Advances in Applied Energy 9 (2023) 100123

Contents lists available at ScienceDirect

Advances in Applied Energy

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

Interpretable machine learning for building energy management: A state-of-the-art review

Zhe Chen[a], Fu Xiao[a,b,*], Fangzhou Guo[a], Jinyue Yan[a]

[a] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[b] Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Building energy efficiency
Building energy flexibility
Interpretable machine learning
Model interpretability
Explainable artificial intelligence

ABSTRACT

Machine learning has been widely adopted for improving building energy efficiency and flexibility in the past decade owing to the ever-increasing availability of massive building operational data. However, it is challenging for end-users to understand and trust machine learning models because of their black-box nature. To this end, the interpretability of machine learning models has attracted increasing attention in recent studies because it helps users understand the decisions made by these models. This article reviews previous studies that adopted interpretable machine learning techniques for building energy management to analyze how model interpretability is improved. First, the studies are categorized according to the application stages of interpretable machine learning techniques: ante-hoc and post-hoc approaches. Then, the studies are analyzed in detail according to specific techniques with critical comparisons. Through the review, we find that the broad application of interpretable machine learning in building energy management faces the following significant challenges: (1) different terminologies are used to describe model interpretability which could cause confusion, (2) performance of interpretable ML in different tasks is difficult to compare, and (3) current prevalent techniques such as SHAP and LIME can only provide limited interpretability. Finally, we discuss the future R&D needs for improving the interpretability of black-box models that could be significant to accelerate the application of machine learning for building energy management.

1. Introduction

The building sector is a major contributor to global energy consumption and carbon emissions. In 2020, it accounted for 36% of global energy consumption and 37% of global CO_2 emissions [1]. Throughout the life cycle of buildings, the operation phase accounts for 80%–90% of total energy consumption [2]. Therefore, building energy management is crucial for global energy-saving and carbon neutrality. Many researchers have quantified the potential of building energy-saving and proposed plans to enhance building energy efficiency. For example, China aims to achieve a 50% reduction in building energy consumption, and one strategy is to adopt efficient equipment and smart building management systems [3]. In Hong Kong, buildings consume 90% of electricity, and therefore the government plans to reduce the electricity consumption of commercial buildings by 30%–40% before 2050 [4]. In the U.S., energy consumption from the building sector can be reduced by efficient heating, control, etc. Researchers estimated that CO_2 emissions from buildings can be reduced by up to 78% by 2050 [5]. In the EU, the building sector would need to reduce its emissions by 60% to reach the EU objective of a 55% reduction by 2030 [6].

Building automation systems (BASs) play an essential role in improving energy efficiency and flexibility during building operations. BASs can implement various smart control strategies in building energy systems, such as heating, ventilating, and air conditioning (HVAC) systems, energy storage systems, and renewable energy systems [7,8]. Traditional control strategies such as rule-based control strategies relying on physics and experience face great challenges in tackling the complicated interactions among building energy systems [9]. Modern buildings are usually equipped with advanced metering infrastructure and numerous sensors; thus, the BAS can collect and store massive energy-related operational data. The prospect of utilizing such big data has opened up due to the advancement in machine learning (ML) algorithms. ML algorithms can discover and learn new knowledge (i.e., data-driven models) from the data and to support energy-efficient/energy-flexible control in the ever-changing energy market [10]. With such data-driven models, building energy systems can be monitored to make decisions autonomously with the support of big data [11].

* Corresponding author at: Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China.
E-mail address: linda.xiao@polyu.edu.hk (F. Xiao).

https://doi.org/10.1016/j.adapen.2023.100123
Received 26 October 2022; Received in revised form 4 January 2023; Accepted 11 January 2023
Available online 13 January 2023

Nomenclature

AI	artificial intelligence
ANN	artificial neural network
BAS	building automation system
CNN	convolutional neural network
DNN	deep neural network
DRL	deep reinforcement learning
DT	decision tree
FDD	fault detection and diagnosis
GAM	general additive model
GRU	gated recurrent unit
HCTSA	highly comparative time-series analysis
HVAC	heating, ventilation, and air conditioning
ICE	individual conditional expectation
IoT	internet of things
kNN	k-nearest neighbors
LIME	local interpretable model-agnostic explanations
LR	linear regression
LSTM	long short-term memory
ML	machine learning
PDP	partial dependence plot
PMV	predicted mean vote
PV	photovoltaics
RNN	recurrent neural network
SHAP	shapley additive explanations
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding

1.1. Machine learning for building energy management

Machine learning has effectively facilitated building energy management in various typical applications in the past decade, including load/power prediction, fault detection and diagnosis (FDD), occupancy-related applications, etc.

Load prediction refers to predicting the cooling/heating/electricity demand in the future hours or days, while power prediction aims to predict the power generation of equipment such as photovoltaic (PV) panels and wind turbines. Accurate load/power prediction is important for improving building energy efficiency and flexibility [12]. Two main applications of load/power prediction models are demand-side management and model predictive control [13]. Demand-side management aims to improve the flexibility of building energy systems by balancing the ever-changing electricity supply and demand caused by the wider uptake of renewable energy systems and dynamic building energy consumption. Model predictive control optimizes building energy systems under constraints (e.g., thermal comfort and setpoint boundaries) to achieve a goal such as minimal cost or energy consumption [14]. Compared with physics-based load/power prediction, ML algorithms require only historical data instead of detailed physical information and thermal balance equations, making them easier to develop and deploy. In the past decades, various ML algorithms have been investigated and achieved satisfactory performance in load/power prediction, mainly including autoregressive methods, tree-based methods, artificial neural networks (ANN), and deep neural networks (DNN) [13]. A previous review on load prediction shows that ANN can deal with real-world problems of considerable complexity [15]. Bahani et al. concluded that ANN and the autoregressive method were the two most accurate algorithms for solar radiation prediction [16]. Although the two methods have similar accuracy, ANN is more flexible as a universal non-linear approximation. Li et al. applied DNN with long short-term memory (LSTM) layers for building electricity demand prediction, and DNN was demonstrated to have better performance than tree-based algorithms [17].

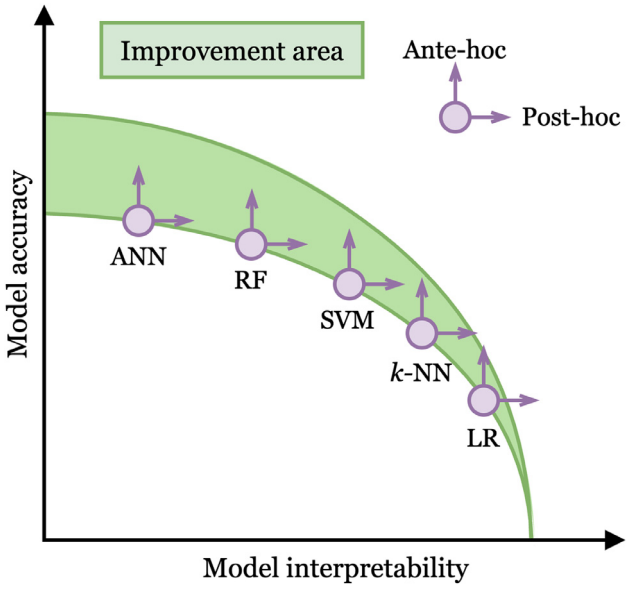

Fig. 1. The trade-off between model interpretability and accuracy [28].

ML has also been used to detect and diagnose faults in building energy systems, i.e., FDD [18]. Early detection of equipment faults is essential for building energy efficiency, especially for energy-intensive equipment such as chillers. Unlike knowledge-driven FDD, data-driven FDD requires less professional knowledge and can distinguish rare and unforeseen energy patterns in real operations [19], which are valuable for FDD.

Occupant thermal comfort is an important index when optimizing building HVAC systems, and accurate thermal comfort models can help improve occupant satisfaction meanwhile reducing energy consumption. Research shows that by adopting an ML-based thermal comfort model instead of predicted means vote (PMV), thermal comfort-related energy consumption and CO_2 levels in buildings can be reduced by up to 58.5% and 24.0%, respectively [20]. Occupancy level prediction and occupancy activity recognition are essential when performing occupancy-based control strategies. According to a literature review, ML-based occupancy level prediction is superior in optimizing the operation of HVAC systems and reduces energy consumption by 23% on average [21]. With the extensive use of Internet of things (IoT) devices, occupancy-related data such as occupancy number and occupancy behaviors can be collected more conveniently, which integrates ML and IoT for building energy efficiency [22,23]. Research shows that the energy consumption of HVAC can be reduced up to 19.8% when occupancy is predicted by ANN using data collected by IoT devices (environmental sensors) [24].

1.2. The need for interpretable machine learning

Model interpretability refers to the degree of how the predictions of an ML model can be understood by human beings [25]. Interpreting predictions can answer the following questions: which features significantly influence the model performance, and which features contribute to the predictions? For example, decision-makers care about the foundation of a fault detection prediction from ML models [26]. Although the applications of ML algorithms have fully demonstrated their values for building energy management, their broad applications are limited by a lack of interpretability [27]. In other words, most ML models are not transparent or explainable.

The trade-off between model accuracy and model interpretability limits the power of machine learning [28,29], as shown in Fig. 1. For

example, ANN usually consists of an input layer, an output layer, and hidden layers. Increasing the number of hidden layers of an ANN model can often improve its accuracy in modeling complicated systems, but the interpretability of the model decreases [30]. The model becomes "deeper" and "darker", making it more difficult for users to understand and interpret the modeling process and results. Meanwhile, the existing ready-to-interpret models, such as linear regression, lack good prediction performance. Researchers made great efforts to improve the prediction performance of black-box ML models, such as ANN and support vector machine (SVM), but generally overlooked their interpretability in building energy management. Fig. 1 also shows the two major approaches to addressing the trade-off between model accuracy and model interpretability: ante-hoc and post-hoc approaches. These approaches are described in detail in Section 2.

The lack of interpretability also challenges the mass deployment of ML models in real-world applications [31]. First, during the training process, the training data are usually incomplete; therefore, the trained ML models need to tackle out-of-distribution data after deployment [32]. During the training process of ML models, physical knowledge and information are usually ignored compared with physics-based modeling. Therefore, decision-makers may find the ML models untrustworthy if the models are not trained on complete operational data and the real-world performance is worse than on the training data. Second, because of the black-box nature, ML models produce output without any explanations. Decision-makers usually need insights into how and why the black-box models produce such predictions so that they can understand, check and apply the models. Generally, there is still significant skepticism in the building industry about the broad application of ML because there is a mismatch between training and deployment environments. Therefore, it is necessary to generate reasonable interpretations that explain the original ML model without oversimplifying essential details or sacrificing prediction performance.

1.3. Scope of the review

Previous review papers on interpretable ML have mainly focused on disease diagnosis [33,34], biomedicine [35,36], and other applications in healthcare [37–39]. In the energy field, Machlev et al. reviewed the applications of interpretable ML in power systems and analyzed the typical interpretable ML techniques [40]. In building energy systems, the interpretability of ML models has become critically important to the mass deployment of AI-empowered smart building energy management after the great efforts in ML model development. To the best of our knowledge, there is no comprehensive review of the applications of interpretable ML for building energy management. The major contributions of this paper are as follows:

1. A comprehensive and critical review of studies adopting interpretable ML and typical interpretable ML techniques in building energy management is presented.
2. The status quo of interpretable ML in building energy management is identified by disclosing how model interpretability is improved from the literature review.
3. Challenges for the wide application of interpretable ML in building energy management are discussed on the basis of this review.
4. Research directions are pointed out by analyzing the limitations of current interpretable ML techniques and the challenges of studies.

The remainder of this paper is constructed as follows. Section 2 presents the background on interpretable ML including the taxonomy of interpretable ML. Section 3 introduces the methodology of how the literature was searched and selected in this study. Section 4 presents the review results according to the taxonomy in Section 2. Section 5 discusses the main findings of the literature review. Finally, the conclusions of this study are given in Section 6.

2. Background on interpretable machine learning

2.1. Interpretable machine learning

A new ML paradigm, known as interpretable machine learning [25], has been adopted by many researchers given the limitations of the traditional ML paradigm. Interpretable ML uses novel approaches and techniques to develop models that are accurate, trustworthy, and easy for users to understand. Fig. 2 illustrates the differences between traditional and interpretable ML paradigms using the example of FDD, a widely researched topic in building energy management (e.g., chiller FDD [41] and air handling unit FDD [42]). In both paradigms, training data are used to develop ML models. When a new input sample is to be examined, the FDD method adopting interpretable ML can not only predict whether the new sample is normal or faulty but also explain the prediction. Therefore, compared to traditional interpretable ML models, interpretable ML models are considered well-founded, trustworthy, and possible to correct mistakes.

2.2. Taxonomy of interpretable ML techniques

According to different criteria, i.e., application stage, interpretability scope, and model dependency, techniques for interpretable ML can be classified into different groups [43], as shown in Fig. 3.

2.2.1. Application stage

First, interpretable ML techniques can be classified according to when the techniques are adopted in building an ML model. Ante-hoc interpretable ML techniques are applied during the model training process, and post-hoc interpretable ML techniques are applied after training. Fig. 4 shows how ante-hoc and post-hoc interpretable techniques are applied at different stages in the model training process.

Ante-hoc interpretable ML models are usually self-explanatory. Therefore, ML models developed using ante-hoc techniques are also called intrinsic or transparent models. For example, linear regression is a simple ante-hoc model for predicting a continuous outcome variable based on one or more predictor variables. Linear regression is self-explanatory because it makes predictions using a linear combination of the input variables, which can be easily understood and explained [43]. Although linear regression has high interpretability according to Fig. 1, it is too simple to address complicated problems in building energy management [44]. In this paper, a variant of linear regression named generalized additive models (GAMs) is reviewed. GAMs have strong flexibility and interpretability in regression and classification tasks [45]. As shown in Fig. 4, post-processing is used to evaluate each input's impact according to the parameters of the intrinsic model. For example, the coefficients of GAMs can be used to evaluate input features' positive or negative effects.

Post-hoc interpretable ML techniques are applied to black-box models after training. They are used to interpret and understand the dependency and significance of specific input features over the output by fitting surrogate models without the need to understand the internal structures. Post-hoc interpretable ML techniques generate interpretation by examining the interrelationship between input features and the predictions.

2.2.2. Interpretability scope

Interpretability scope refers to the scope of model output that needs to be interpreted. As the classification problem shown in Fig. 5, global interpretation explains an ML model based on a full view of the model structures and parameters. In contrast, local interpretation explains each prediction individually.

Global interpretable ML techniques aim to provide a holistic understanding of the ML model by measuring the global effects of the input features on the model prediction. They require only the black-box models and the entire training data. Global interpretation helps decision-

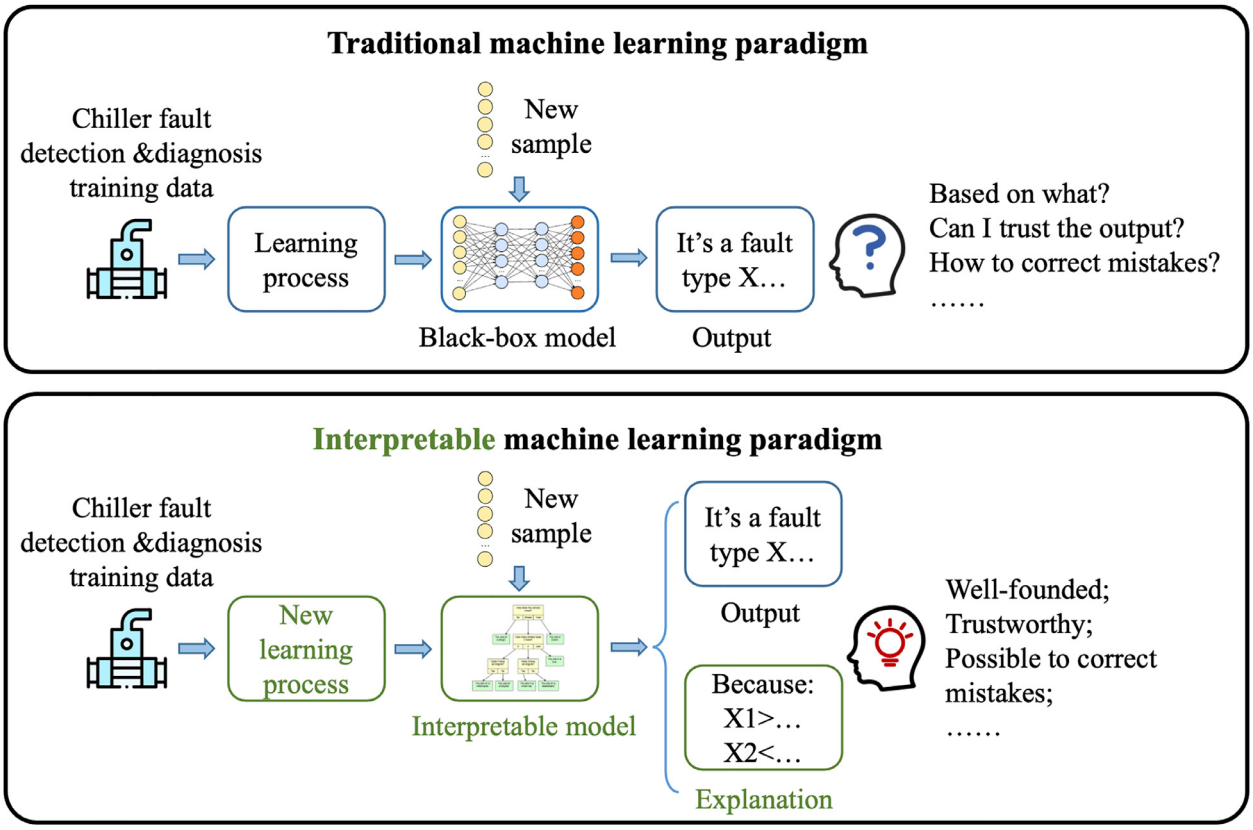

Fig. 2. Comparison of traditional ML paradigm and interpretable ML paradigm in FDD applications.

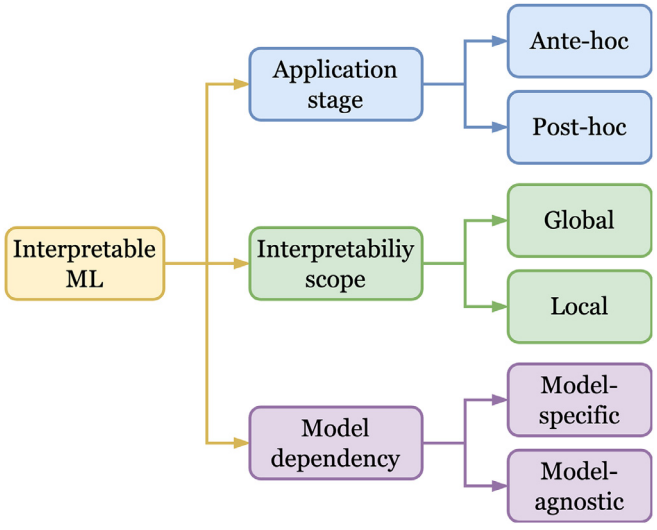

Fig. 3. Taxonomy of interpretable ML [43].

makers gain a macro-level understanding of the ML model, including the most influential input features. In the context of FDD, global interpretability helps explain which features are most significant in predicting equipment faults.

Local interpretable ML techniques provide a transparent understanding of the model prediction for a specific input sample. Instead of global feature importance, local methods focus on the contribution of each feature to a prediction sample and require both the black-box model and the prediction sample. Local interpretation is important for decision-makers to trust the output or correct the wrong output. In the context of FDD, local interpretability helps explain which features contribute the most to the prediction sample, such as high supply air temperature in air handling unit operation.

2.2.3. Model dependency

Model dependency refers to whether the interpretable ML technique can be applied to any ML model or to specific models. Some interpretation techniques treat the ML models as black-box models, and these techniques are applicable to any ML model or are independent of the type of ML model. Therefore, these techniques are model-agnostic, as illustrated in Fig. 6(a). Other techniques can only be applied to interpret certain types of ML models and are thus called model-specific techniques, as shown in Fig. 6(b).

Model-agnostic techniques can be applied to any ML model because they require only the input and output of the ML model without considering its inner structures. Therefore, most post-hoc interpretable ML techniques are model agnostic. For example, LIME is a post-hoc model-agnostic tool that can approximate any ML model locally.

Model-specific techniques can dig into the specific characteristics or architecture of the ML model, providing in-depth interpretability that may not be possible with model-agnostic methods. For example, the attention mechanism is usually employed in neural networks to improve interpretability as a model-specific technique.

2.2.4. Summary

According to the definitions, the above three criteria to classify interpretable ML techniques are not independent. First, ante-hoc techniques are usually model-specific because the interpretation is tied to the ante-hoc models. Post-hoc techniques, on the other hand, are not always model-agnostic. For example, gradients can only be used to interpret neural networks as a post-hoc technique. Additionally, all model-agnostic techniques are post-hoc because they treat ML models as black boxes and generate interpretation by examining the interrelationship

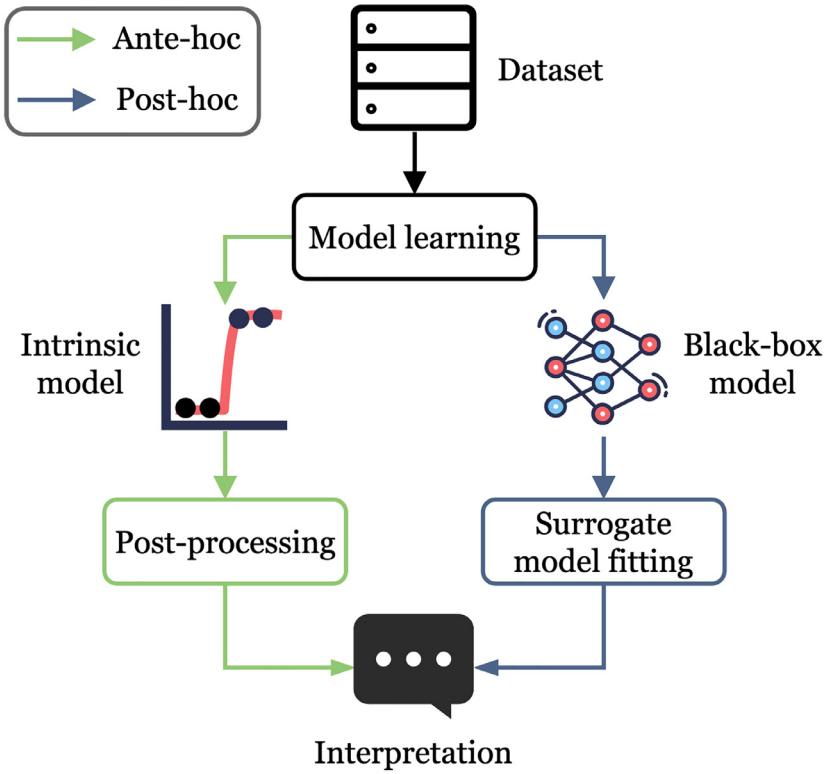

Fig. 4. Ante-hoc and post-hoc interpretability.

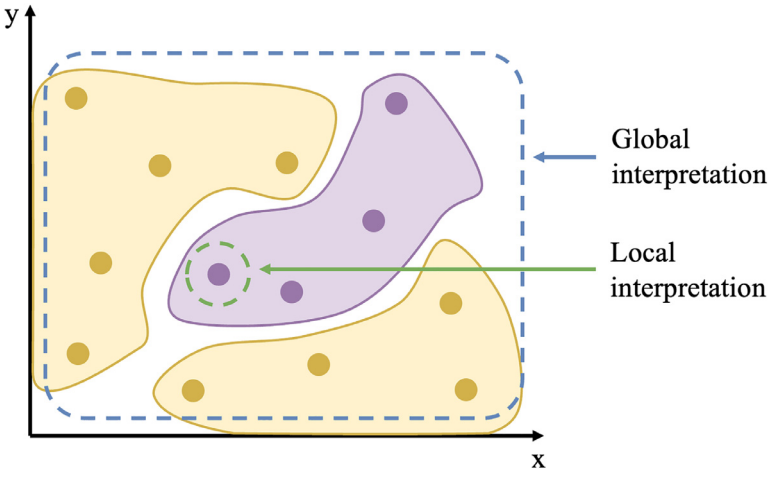

Fig. 5. Global and local interpretability in a classification task [46].

between input features and the predictions. Finally, some techniques can be used to provide both local and global interpretation, e.g., SHAP (refer to Section 4.3.2 for more information).

3. Methodology of literature review

The methodology for collecting relevant literature consists of four steps: keywords construction, keywords search, initial selection, and final selection, as illustrated in Fig. 7.

First, two groups of keywords were constructed. The first group was constructed according to the domain knowledge of building energy management (BEM), including energy-related terms such as *building load/demand* and building energy management-related tasks such as *building control* and *building fault detection and diagnosis*. The second group of keywords is related to interpretable ML. The keywords were determined based on the terminologies frequently used in interpretable ML. *Explainable AI* was also included because ML is a branch of artificial intelligence (AI).

Second, all possible one-to-one combinations of the two groups of keywords were searched in Google Scholar and ScienceDirect. For example, *"building energy* & *interpretable machine learning", "building energy* & *model interpretability"*, and *"building energy* & *explainable AI"*.

Third, the initial selection aims to screen the papers found by searching keywords in Google Scholar and ScienceDirect to ensure the selected papers actually address building energy management and relevant applications. The initial selection was carried out by carefully reviewing the papers' titles, keywords, and abstracts.

The last step is the final selection. The terminologies of interpretable ML in the literature vary and are sometimes not clearly stated in the title, keywords, and abstracts of papers after the initial selection. Therefore, it is necessary to scrutinize these papers' methodology and results/discussion sections to ensure that interpretable ML techniques of interest were adopted.

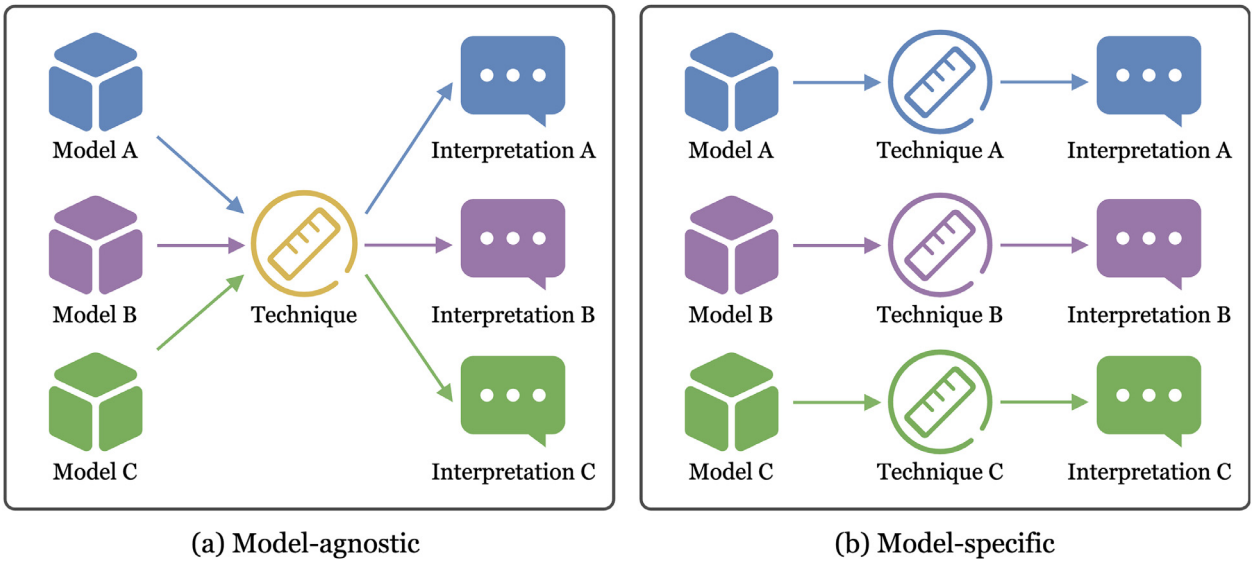

Fig. 6. Model-agnostic and model-specific interpretable ML techniques.

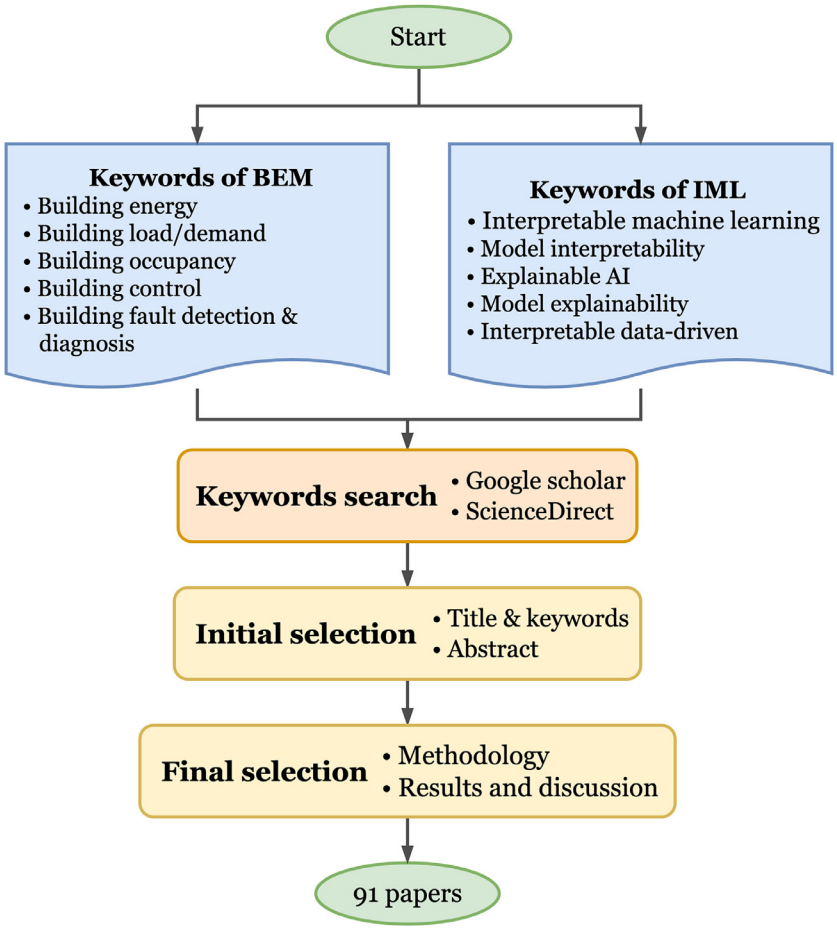

Fig. 7. The workflow of selecting studies on interpretable ML for building energy management.

After the above four steps, 91 papers were selected. The cut-off date of the literature review is June 1st, 2022. These papers were first categorized according to the application stage of interpretable ML techniques (i.e., ante-hoc and post-hoc approaches). After that, the applications in building energy management and specific techniques used to improve model interpretability were analyzed in depth for each approach in Section 4.

4. Ante-hoc and post-hoc interpretable ML techniques for building energy management

4.1. Overview of literature

Typical applications of the 91 papers for building energy management over the years are shown in Fig. 8, including fault detection and di-

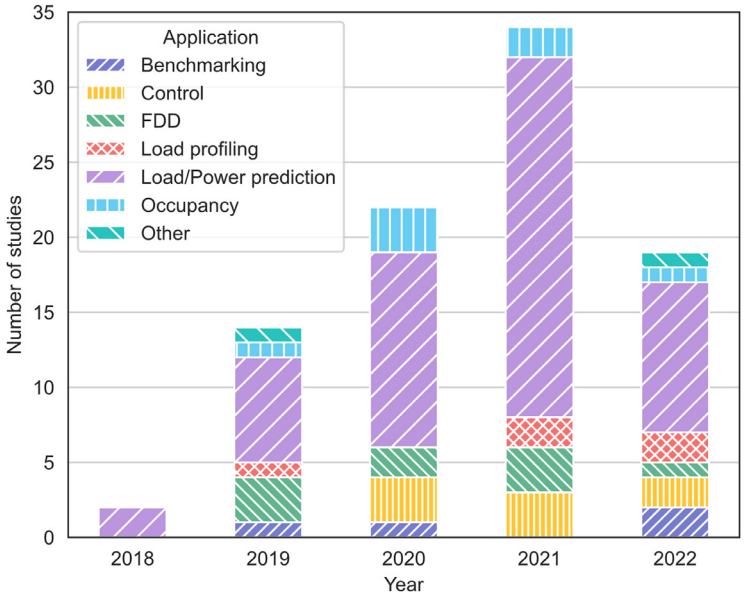

Fig. 8. Distribution of applications in studies by year.

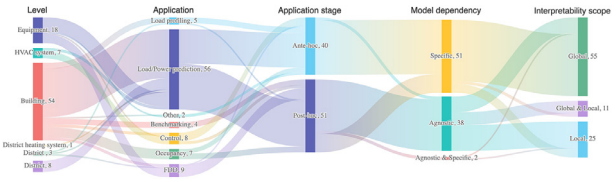

Fig. 9. Sankey diagram depicting the connections of reviewed studies on different levels, applications, stages, model dependency, and interpretability scopes.

agnosis, load/power prediction, control, etc. The overall trend shows an increasing research interest in interpretable ML for building energy management, because increasing applications of ML stimulate the need to understand ML models. Since the cut-off date of this review is June 1st, 2022, the number of publications in 2022 is expected to exceed 2021. Furthermore, in line with the findings from the word cloud, load/power prediction is the most popular application accounting for 61.5% of the studies, followed by FDD and control (on HVAC or other building energy systems).

A Sankey diagram is created to illustrate the classifications of the studies in five dimensions and the connections among them, as shown in Fig. 9. The five dimensions include application stage, model dependency, interpretability scope, application (i.e., the seven typical applications identified above) and level (i.e., equipment, system, building, and district levels). Under each dimension, the reviewed studies are classified into several groups differentiated by labels. A label has two parts, i.e., the group name and its respective value representing the number of studies in the group. For example, under the *Level* dimension, the reviewed studies are divided into six groups. The largest group is "*buildings*" which has 54 papers. The link between two groups in the diagram represents the flow (a further division) from one dimension to another dimension. For example, from the "*buildings*" group in the "*level*" dimension, the majority goes to the "*load/power prediction*" in the "*application*" dimension. In other words, of the 54 papers focusing on building-level applications, most addressed load/power prediction using interpretable ML techniques. This phenomenon is understandable because the accurate prediction of load/power plays a critical role in building energy management [12]. As load/power prediction models are becoming too complex for users to understand, model interpretability has gained increasing popularity [47]. In addition, most studies on control focus on HVAC systems, which is reasonable as HVAC systems are responsible for the largest portion of energy consumption of buildings and have the largest energy-saving potential. Therefore, the interpretability of those methods and models becomes increasingly important as more machine learning methods have been developed to improve HVAC system performance, such as model-based predictive control, model-based optimization, and FDD.

From the connection between the application stage and model dependency, it can be found that nearly all ante-hoc techniques in the literature are model-specific, while most model-agnostic techniques are post-hoc. This is because of the nature of ante-hoc and model-agnostic techniques. Ante-hoc techniques usually improve the model interpretability by adding interpretable characteristics or adopting models with intrinsic interpretability. Therefore, most ante-hoc techniques are model-specific. Post-hoc techniques can be either model-specific or model-agnostic because some post-hoc techniques such as integrated gradients can only be applied to neural networks [48], while other techniques such as SHAP can be adopted to any machine learning models. As for model-agnostic techniques, they do not rely on the model characteristics and are usually applied after model development. The connection between the model dependency and interpretability scope shows that most model-specific studies generate global interpretations while model-agnostic studies can generate global, local, or both global and local interpretations. It is also observed that more studies focus on global interpretation.

4.2. Ante-hoc approach

40 papers out of the 91 papers reviewed in this study adopted the ante-hoc approach to improve model interpretability for building energy management. These papers are further divided into four categories according to the specific ante-hoc techniques adopted: modified neural networks, attention mechanism, clustering and feature extraction, and generalized additive models (GAMs). Fig. 10 shows the number of studies in each category from 2018 to 2022, and Fig. 11 shows the various applications adopting ante-hoc techniques in the reviewed studies. Load/power prediction and control are two main applications adopting ante-hoc techniques.

4.2.1. Modified neural networks

ANN has become popular in building energy management in the past decades. It is well known that ANN is dark for users. A typical approach to improving ANN's interpretability is modifying neural networks' structure, which generates the so-called modified neural networks with en-

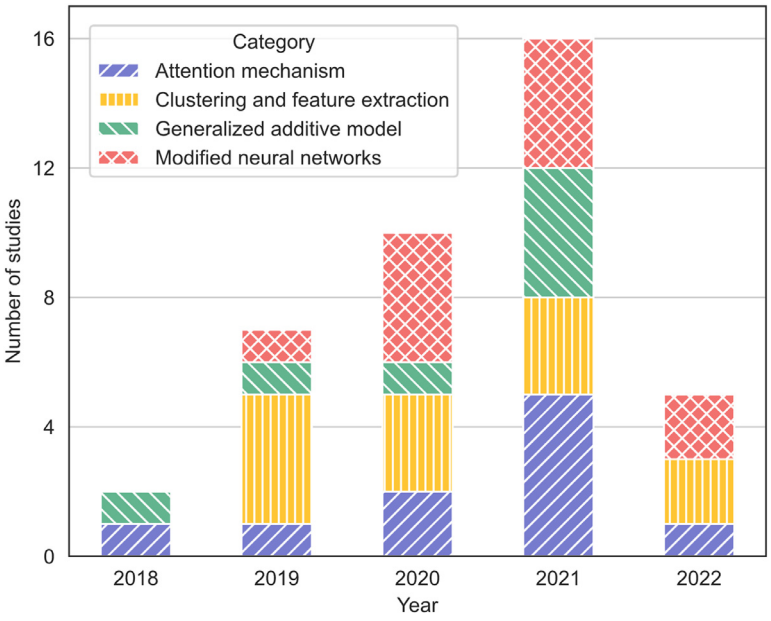

Fig. 10. Distribution of publications by year published and ante-hoc category.

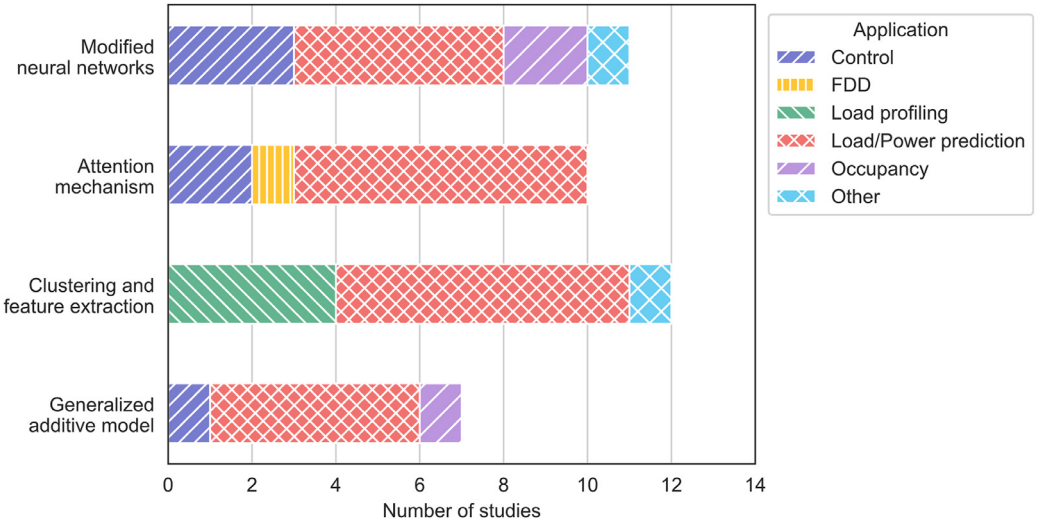

Fig. 11. Breakdown of studies adopting ante-hoc techniques.

hanced interpretability. A summary of studies adopting modified neural networks is listed in Table 1.

There are two ways to modify the structure of neural networks in general. First, elements with physical meanings can be directly added to the models. Shan et al. integrated the gravitational model (GRA) with the gated recurrent units (GRU) model for building energy consumption prediction [49]. In the proposed GRA-GRU model, linear model GRA and non-linear model GRU were ensembled. The weights of the two models were determined using mutual information and weighted entropy. Wang et al. proposed a direct explainable neural network (DXNN) using the ridge function instead of the widely-used sigmoid activation function as the kernel function [51]. Considering the polynomial characteristic of the ridge function, the mathematical relationship between the model input and output can be directly obtained. The DXNN was used for solar irradiance forecasting, and the results showed that the output is a quadratic function of input features. Zhang et al. combined a deep belief network with Takagi-Sugeno-Kang fuzzy classifier to generate interpretable fuzzy rules for indoor occupancy detection [53]. Kim modified traditional convolutional neural network (CNN) and proposed interpretable CNN (I-CNN) for indoor human activity detection by adding temporal convolution and pooling layers into the CNN. The author demonstrated that the proposed I-CNN could rank the importance of sensor signals and improve the performance of I-CNN [50]. Chen and Zhang used domain knowledge to obtain the average trend of district load reflecting the periodic patterns so that the data-driven model could predict irregular local load fluctuations [54]. Similarly, Oreshkin et al. decomposed the load consumption time-series into human-interpretable outputs (i.e., trend and seasonal components) using the Loess method [57]. Li et al. proposed an automatic relevance determination (ARD) network that can incorporate the uncertainty of input and evaluate the importance of input features [56]. The proposed modified neural network can reveal the relevance of input features and model output. The results showed that the hour of day was the most influential feature for hourly electricity prediction.

Second, domain knowledge can be used to guide the design of neural networks or improve the training process of neural networks, making the neural networks and training processes physically explainable. Chen et al. used domain knowledge to guide the design of model input/output and the structure of neural networks for air-conditioner modeling [58]. To make the control signal predicted by the deep Q network trustworthy

Table 1
Summary of studies adopting ante-hoc techniques.

Category	Ref.	Year	Application	Level	ML task	ML algorithm	Model dependency	Scope
Modified neural networks	[49]	2019	Load/Power prediction	Building	Regression	GRA-GRU	Specific	Global
	[50]	2020	Occupancy	Building	Regression	I-CNN	Specific	Global
	[51]	2020	Load/Power prediction	Building	Regression	DXNN	Specific	Global
	[52]	2020	Control	Equipment	Regression	DQN	Specific	Global
	[53]	2020	Occupancy	Building	Classification	DBN-TSK-FC	Specific	Global
	[54]	2021	Load/Power prediction	District	Regression	EnLSTM	Specific	Global
	[55]	2021	Control	Building	Regression	Physics-constrained deep learning	Specific	Global
	[56]	2021	Load/Power prediction	Building	Regression	ARD	Specific	Global
	[57]	2021	Load/Power prediction	District	Regression	N-BEATS	Specific	Global
	[58]	2022	Heat exchanger modeling	Equipment	Regression	NARX-LSTM-MLP	Specific	Global
	[59]	2022	Control	HVAC system	Regression	PCNN	Specific	Global
Attention mechanism	[60]	2018	Load/Power prediction	Building	Regression	Multi-variable LSTM	Specific	Global
	[61]	2019	FDD	Equipment	Classification	Encoder-decoder	Specific	Local
	[62,63]	2020	Control	HVAC system	Regression	ST-Att	Specific	Global
	[64]	2020	Load/Power prediction	Building	Regression	Temporal fusion transformers	Specific	Local
	[65]	2021	Load/Power prediction	District	Regression	Bi-LSTM	Specific	Global
	[47]	2021	Load/Power prediction	Building	Regression	Encoder-decoder	Specific	Global
	[66]	2021	Load/Power prediction	Equipment	Regression	Encoder-decoder	Specific	Global
	[67]	2021	Load/Power prediction	District	Regression	Encoder-decoder	Specific	Global
	[68]	2022	Load/Power prediction	Building	Regression	IM-LSTM	Specific	Global
Clustering and feature extraction	[69]	2019	Building thermal design	Building	Clustering	AAHR	Specific	Global
	[70]	2019	Load/Power prediction	Building	Regression	DT, kNN	Specific	Local
	[71]	2019	Load/Power prediction	Building	Clustering	EXP, CART, CTREE, RF	Specific	Global
	[72]	2019	Load profiling	Building	Classification	hctsa	Specific	Global
	[73]	2020	Load/Power prediction	Building	Regression	kNN	Specific	Global
	[74]	2020	Load/Power prediction	Building	Regression	RF	Specific	Global
	[75]	2020	Load/Power prediction	Equipment	Regression	BDLSTM	Agnostic	Global
	[76]	2021	Load/Power prediction	Building	Regression	LightGBM	Agnostic	Global
	[77]	2021	Load/Power prediction	Building	Regression	kNN	Specific	Global
	[78]	2021	Load profiling	Building	Clustering	k-means	Specific	Global
	[79]	2022	Load profiling	Building	Classification	hctsa	Specific	Global
	[80]	2022	Load profiling	District	Regression	Multi-equation model	Specific	Global
Generalized additive models	[81]	2018	Load/Power prediction	Building	Regression	GAM, LSTM	Specific	Global
	[82]	2019	Occupancy	Building	Regression	GAM	Specific	Global
	[83]	2020	Load/Power prediction	Building	Regression	GAM	Specific	Global
	[84]	2021	Load/Power prediction	District	Regression	PLAM	Specific	Global
	[85]	2021	Load/Power prediction	District	Regression	GAM	Specific	Global
	[86]	2021	Load/Power prediction	Equipment	Regression	GAM	Specific	Global
	[87]	2021	Control	Building	Regression	GAM	Specific	Global

and aligned with domain knowledge, Yu et al. integrated *a priori* knowledge into the searching strategy. They concluded that the knowledge-based search strategy could significantly reduce training time [52]. The modified LSTM proposed in [55] also used thermal dynamics to guide the design of a recurrent neural networks (RNN) model for building thermal modeling, which can learn interpretable dynamic models from measurement data. Di Natale et al. proposed a physically consistent neural network by incorporating domain knowledge into black-box models for building thermal modeling, and the proposed approach was proved to be physically interpretable [59].

4.2.2. Attention mechanism

The attention mechanism was first introduced by Bahdanau et al. to improve the performance of the encoder-decoder model for machine translation [88]. Inspired by the cognitive attention process, the attention mechanism can improve the interpretability of encoder-decoder models by stressing some parts of the input features in making predictions while weakening the rest features based on the context vectors. Because encoder-decoder models deal with time-series data, the attention mechanism can consider the temporal dependency of time-series data [68]. It is an ante-hoc approach because it is embedded into the prediction model [89]. Studies adopting the attention mechanism are listed in Table 1.

Many studies have used the attention mechanism to analyze temporal dependency in time-series data in both regression and classification tasks. According to the individual attention matrix of input samples, Li et al. analyzed the temporal dependency of time-series data and removed redundant features for chiller fault diagnosis [61]. It could provide local interpretation of the importance of sensors on the fault diagnosis resulting from the encoder-decoder network. Attention weight heatmap was used in [66] to explore the features emphasized in the LSTM model for day-ahead daily load prediction. Results showed that one day-ahead load is the most important feature. Similarly, average attention patterns in [65] demonstrated that the impact of historical features on model output exhibited 24-hour periodicity, indicating a strong relationship between energy consumption and the hour of the day. Li et al. adopted the attention mechanism in the ANN model for building cooling load prediction. They found that the most recent energy consumption data had the most significant influence on the next-hour cooling load prediction [47]. In [62,63], spatiotemporal attention values were almost evenly distributed across all input time steps for zone air temperature prediction because air temperature had faster thermal dynamics than the building envelope.

Technically, the attention mechanism generates local interpretations because it treats each input sample individually. Nevertheless, many studies treat the average of attention values as global interpretability. For example, Guo et al. [60] stated that the average attention values of LSTM could represent the importance of input features, which conform to the domain knowledge.

4.2.3. Clustering and feature extraction

Unlike modified neural networks and attention mechanisms embedded into black-box models (neural networks), clustering and feature extraction techniques do not change the structure of the original black-box models. The clustering and feature extraction techniques improve the interpretability of machine learning models by clustering raw data

into several groups with human-interpretable characteristics (e.g., interpretable rules) or extracting interpretable features. A summary of the relevant studies is provided in Table 1.

Bhatia et al. proposed a novel clustering technique named axis-aligned hyper-rectangles [69] for clustering simulated building thermal design data. Compared with other clustering techniques, it could generate hyper-rectangle boundaries that can be described with interpretable rules. It was employed to extract interpretable rules, such as the range of window-to-wall ratio, to assist the design of building envelopes in different climate zones [69]. Laurinec and Lucká proposed an interpretable time-series clustering technique named ClipStream [71] to improve the interpretability of electricity forecasting. The case study showed that extracting interpretable features from a moving window of time-series data improved demand forecast accuracy. For example, the number of data points below average within the moving window could help explain the time-series' overall shape. Some studies have compared the performance of clustering techniques with black-box models. Grimaldo and Novak claimed that the interpretable ML approach did not sacrifice the model's accuracy [77]. Their case study showed that the k-nearest neighbors (kNN) algorithm had similar accuracy for load prediction compared with sophisticated machine learning models such as RF and gradient boosted trees. The kNN algorithm is interpretable because it is model-free and makes predictions according to the nearest neighbors of a sample.

Visualizing the results of clustering techniques can improve interpretability as it shows intuitive differences among different clusters. Grimaldo and Novak used kNN and decision trees (DT) to predict building energy consumption on similar days. They then developed a smart energy dashboard visualizing energy consumption of similar days to help users understand the prediction results [70]. They also presented a radar chart to compare the similarity of weather parameters in the same prediction task [73].

Some studies have extracted interpretable features using clustering techniques during the feature engineering process. Highly comparative time-series analysis (HCTSA) is a toolkit that can generate interpretable time-series features programmed in MATLAB [90]. The features extracted from building energy consumption data were used to explain the classification of primary space usage [72,79]. Hu et al. extracted 21 interpretable features based on domain knowledge from building load data, including 13 global features (e.g., mean value of a daily load pattern) and eight peak-period features (e.g., number of peak periods) [78]. Kasuya et al. got typical energy usage modes as an input feature of next-day load prediction of a test building using the Gaussian mixture model, which is a distribution-based clustering algorithm [74]. Chen et al. generated mode labels as input features using a novel early classification approach to enhance the interpretability and performance of building load prediction [76]. Instead of splitting data into several clusters/modes, Castellini et al. split the problem of predicting the heating load into several subproblems so that each subproblem can be approximated linearly [80]. In this way, the interpretability of the models developed for subproblems was improved.

4.2.4. Generalized additive models (GAMs)

Generalized additive models (GAMs) have gained increasing attention recently owing to their model interpretability. GAMs are a variant of generalized linear models that can model the non-linear additive effects of each feature [43]. The general structure of GAMs is defined as:

$$g(\mathbb{E}(y \mid \mathbf{x})) = \mathbf{w}_0 + f_1(\mathbf{x}_1) + \cdots + f_i(\mathbf{x}_i) \tag{1}$$

where $g(\cdot)$ is the link function that connects the estimated mean $\mathbb{E}(y \mid \mathbf{x})$ to the sum of additive effects, $\mathbf{w}_0$ is the model intercept, and $f_i(\cdot)$ is the additive effect function (e.g., linear, cubic spline) for the feature $\mathbf{x}_i$ to be estimated.

Compared with linear models, GAMs are more flexible and can incorporate irregular and volatile effects to improve flexibility in handling high-resolution data [84]. A summary of studies adopting GAMs is given in Table 1. Bujalski and Madejski used GAMs to predict heat production in a combined heat and power plant system [85]. The results showed that ambient air temperature, solar radiation, and hour of the day had different impacts on the heating load. For example, outdoor air temperature showed a negative linear relation with heating load prediction, while solar radiation showed a negative exponential relationship. In addition, GAMs were also applied to identify operational patterns of HVAC systems [81] and perform sensitivity analysis of input features in thermal comfort modeling [82], thermal energy storage modeling [87], distributed PV power prediction [86], and short-term energy prediction in buildings [83].

4.2.5. Summary

The ante-hoc approach improves model interpretability by modifying existing machine learning models or incorporating interpretable features into the design or training process. Therefore, explanations can be directly obtained from the trained model. Studies have shown that the ante-hoc approach can improve model interpretability in building energy management without sacrificing accuracy. However, ante-hoc interpretable techniques are often specific to a particular model and may not be applicable to other types of models, as mentioned in Section 2.2.4. Modified neural networks and attention mechanism are examples of techniques designed for neural networks. GAM is also a type of model-specific technique because of its self-explanatory nature. Clustering and feature extraction techniques can be either model-specific or model-agnostic, depending on their function during the model development process. For example, clustering techniques can be used to generate interpretable features before model training in a model-agnostic way, as in [75]. In [71], an interpretable clustering technique is used to cluster similar load profiles in a model-specific manner. Overall, the adoption of ante-hoc techniques depends heavily on the ML algorithms and tasks being addressed.

4.3. Post-hoc approach

51 of the 91 papers reviewed in this study adopted the post-hoc approach, as summarized in Table 2. In Fig. 12, the Sankey diagram shows the connections in three dimensions, namely model dependency, post-hoc technique, and interpretability, along with the number of papers in each subdivided category. Because some studies adopted more than one post-hoc technique, the total number of papers in each dimension is greater than the total number of papers adopting the post-hoc approach (i.e., 51). As shown in Fig. 12, the model-agnostic method is more often used than model-specific techniques. Fig. 13 shows the various applications adopting post-hoc techniques in the reviewed studies. The papers are classified according to the post-hoc techniques: SHAP, LIME, visualization and partial dependency plot (PDP), and other techniques. The figure shows that load/power prediction is the main application.

4.3.1. Local interpretable model-agnostic explanations (LIME)

LIME was proposed by Ribeiro et al. in 2016 as a model-agnostic approach to obtain local interpretation for individual predictions [142]. The local interpretation is obtained by training a local surrogate model to approximate the local characteristics of the black-box model in the region around the prediction sample. The interpretable model is obtained by optimizing the following objective $\xi(x)$:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

where f is the black-box model, g is the local surrogate model from searching space G that defines the type of interpretable models such as linear or logistic models, π_x defines locality around data instance x, $\mathcal{L}$ is a loss function that measures the fidelity of the surrogate model g to the black-box model f, and Ω measures the complexity of the surrogate model.

Table 2
Summary of studies adopting post-hoc techniques.

Refs.	Year	Application	Level	ML task	ML algorithm	Post-hoc technique	Model dependency	Scope
[91,92]	2019	Load/Power prediction	Equipment	Regression	ANN	Permutation importance, SHAP	Agnostic	Global
[93]	2019	FDD	Equipment	Classification	GLM, MLP, SVM, RF, XGB	LIME	Agnostic	Local
[94]	2019	Load/Power prediction	Building	Regression	Autoencoder	Visualization	Specific	Global
[95]	2019	Load/Power prediction	Building	Regression	CNN-LSTM	Visualization	Specific	Both
[96]	2019	FDD	Equipment	Classification	SVM, ANN	LIME	Agnostic	Local
[97]	2019	Benchmarking	Building	Regression	XGBoost	SHAP	Agnostic	Local
[98]	2019	Load/Power prediction	HVAC system	Regression	RF, GBDT, XGBoost	Feature importance	Specific	Global
[99]	2020	Benchmarking	Building	Regression	XGBoost	SHAP	Agnostic	Both
[100]	2020	Load/Power prediction	Equipment	Regression	TS-SOM, XGBoost	SHAP	Agnostic	Both
[101]	2020	Load/Power prediction	Building	Regression	LSTM	Visualization	Specific	Global
[102]	2020	Load/Power prediction	Equipment	Regression	RF	LIME, SHAP, ELI5	Agnostic, specific	Local
[103]	2020	Control	HVAC system	Regression	DNN	LIME, PDP, ICE	Agnostic	Both
[104]	2020	Load/Power prediction	Building	Regression	XGBoost	SHAP	Agnostic	Local
[105]	2020	FDD	District	Classification	RF	SHAP	Agnostic	Both
[106]	2020	Load/Power prediction	Building	Regression	CNN	Visualization	Specific	Global
[107]	2020	FDD	Building	Classification	NS-NN	Integrated gradients	Specific	Local
[108]	2020	Load/Power prediction	Building	Regression	LSTM	Feature importance	Agnostic	Global
[109]	2020	Occupancy	Building	Regression	Numerical equations	SHAP, PDP	Agnostic	Both
[110]	2021	Load/Power prediction	Building	Regression	MLP, LSTM, Seq2Seq, kNN, RF	SHAP	Agnostic	Global
[111]	2021	Load/Power prediction	Building	Regression	XGBoost	SHAP	Agnostic	Local
[112]	2021	Load/Power prediction	Equipment	Regression	DNN	SHAP	Agnostic	Local
[113,114]	2021	Load/Power prediction	Building	Regression	Autoencoder	Visualization	Specific	Global
[115]	2021	FDD	Equipment	Classification	CNN	Visualization	Specific	Global
[116]	2021	Load/Power prediction	Equipment	Regression	DNN	SHAP	Agnostic	Local
[117]	2021	Occupancy	Building	Regression	RF, kNN, DNN, LR	SHAP	Agnostic	Both
[118]	2021	Load/Power prediction	District	Regression	ANN, LR	DiCE Diverse Counterfactual Explanations.	Agnostic	Local
[119]	2021	FDD	Building	Classification	XGBoost	SHAP	Agnostic	Both
[120]	2021	Load/Power prediction	Equipment	Regression	XGBoost	ELI5	Specific	Both
[121]	2021	FDD	Equipment	Classification	XGBoost	LIME	Agnostic	Local
[122]	2021	Occupancy	Building	Regression	GBR	LIME, SHAP	Agnostic	Both
[123]	2021	Load/Power prediction	HVAC system	Regression	ANN	Gradient method	Specific	Global
[124]	2021	Load profiling	Building	Classification	RF, CNN, InceptionTime	LIME, SHAP	Agnostic	Local
[125]	2021	Load/Power prediction	District	Regression	LSTM	LIME	Agnostic	Local
[126]	2021	Load/Power prediction	Building	Regression	RF	PDP, Rule extraction	Agnostic	Global
[127]	2021	Load/Power prediction	Building	Regression	LightGBM, RF, Bi-RNN, Bi-LSTM, Bi-GRU	SHAP	Agnostic	Global
[128]	2022	Control	HVAC system	Regression	NSGA-II	Rule extraction	Agnostic	Global
[129]	2022	Benchmarking	Building	Regression	CatBoost	LIME	Agnostic	Local
[130]	2022	FDD	Equipment	Classification	RF, LightGBM	SHAP	Agnostic	Both
[131]	2022	Load/Power prediction	Building	Regression	DNN	LIME	Agnostic	Local
[132]	2022	Load/Power prediction	District	Regression	LSTM	LIME	Agnostic	Local
[133]	2022	Benchmarking	Building	Regression	RF, Adaboost	LIME, feature importance	Agnostic	Both
[134]	2022	Load/Power prediction	Building	Regression	Encoder-decoder	Kullback–Leibler divergence	Specific	Local
[135]	2022	Load/Power prediction	Building	Regression	XGBoost	SHAP	Agnostic	Local
[136]	2022	Load/Power prediction	Building	Regression	Cubist	Feature importance	Specific	Local
[137]	2022	Load/Power prediction	Building	Regression	RAnger	Feature importance, PDP	Specific, agnostic	Global
[138]	2022	Load/Power prediction	Building	Regression	LSTM	PDP	Agnostic	Global
[139]	2022	Load/Power prediction	Building	Regression	Expectile regression	Rule extraction	Agnostic	Global
[140]	2022	Occupancy	Building	Regression	SVM, ANN, RF, GBDT, XGBoost	SHAP	Agnostic	Both
[141]	2022	Load/Power prediction	District	Regression	LSTM, Bi-LSTM, CNN-LSTM, encoder-decoder, etc.	LIME	Agnostic	Local

As LIME can give the contradict or support value of each input feature for a prediction sample, it is valuable to explain the prediction of classification problems. Wastensteiner et al. used LIME to interpret ML-based time-series classification for building energy consumption and analyzed the stability and reliability of the interpretation [124]. Madhikermi et al. trained ANN and SVM for AHU fault diagnosis, and six samples were randomly selected to demonstrate the interpretability of LIME [96]. Srinivasan et al. experimented with interpreting three types of faults of chiller operation (i.e., scaling in condenser fins, sensor errors caused by pulsations in the flow, and false alarm) using LIME. On the one hand, decision-makers can know the foundation of the model output to support fault/normal decisions based on the contradict and support values given by LIME. On the other hand, LIME also provides information for the possible false alarms of the black-box model [121].

Apart from classification, LIME can also be applied to regression tasks. Fan et al. integrated the contradiction and support values into a single metric to evaluate the confidence level of a single prediction of chiller COP efficiency (i.e., low or high efficiency) [93]. Kotevska et al.

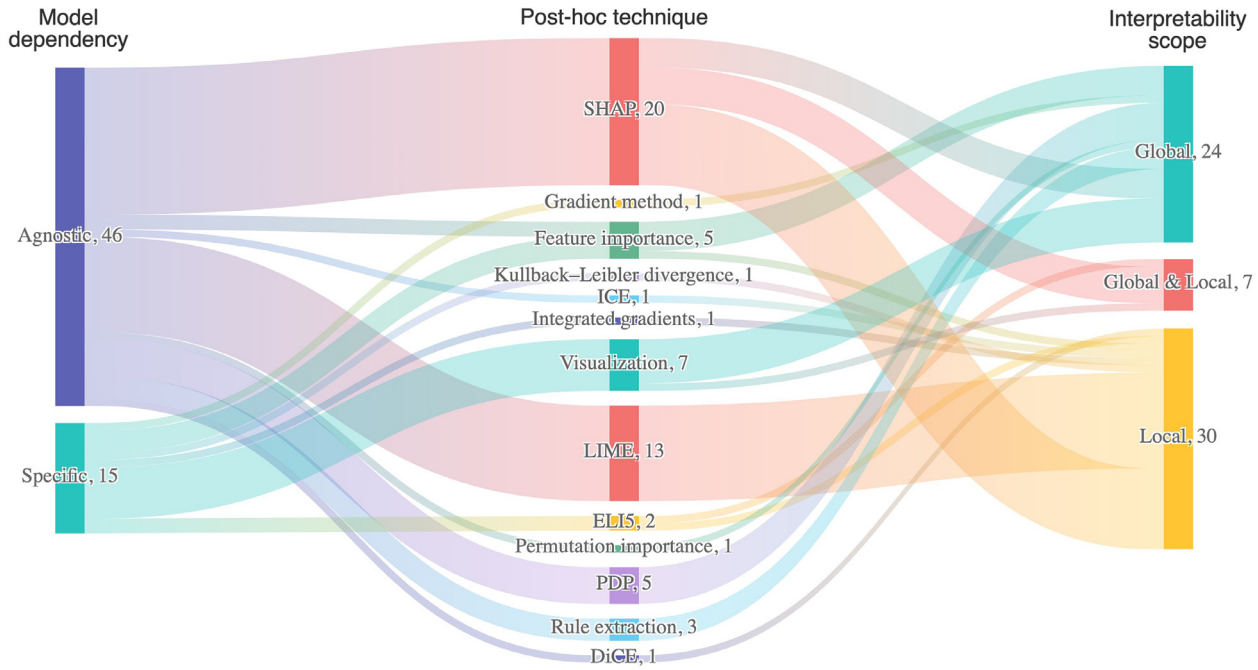

Fig. 12. Sankey diagram depicting the connections of reviewed studies on different model dependencies, post-hoc techniques, and interpretability scopes.

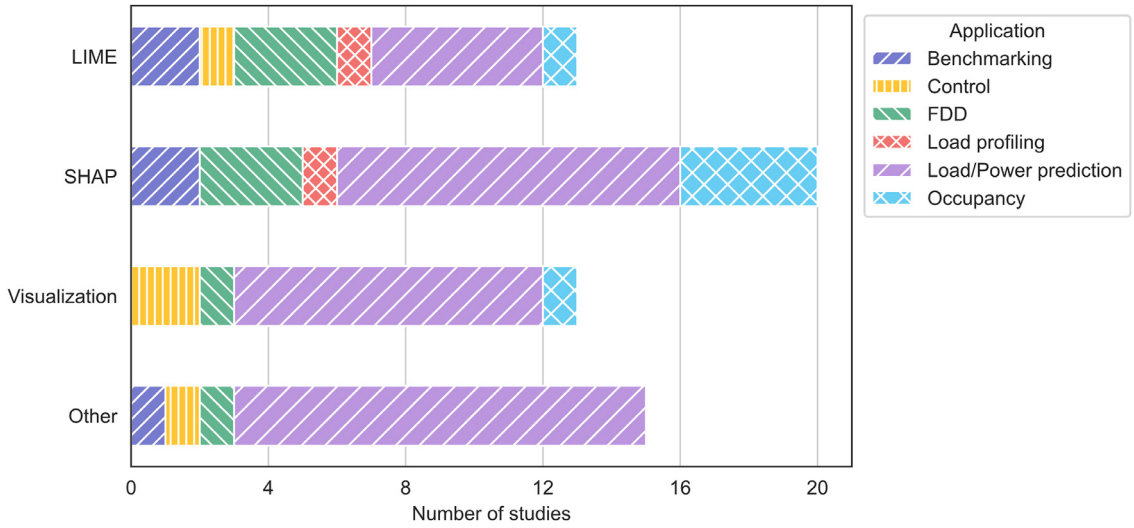

Fig. 13. Breakdown of studies adopting post-hoc techniques.

used LIME to get the local linear approximation of the deep reinforcement learning (DRL) setpoint controller model. The results showed that the impact of zone temperature on setpoint recommendation varies in different ranges [103]. Zdravković et al. employed LIME to generate the local feature importance of prediction samples for heating demand prediction and anomaly detection in district heating systems [125,141]. Likewise, Arjunan et al. adopted LIME to improve the interpretability of the CatBoost model for building energy benchmarking [129]. In the case study, the authors gave an example of a building that consumed less energy than its peer group. According to the local interpretation provided by LIME, it was because the target building had a lower air-conditioned floor area. Jin et al. presented a LIME-based interpretable building energy benchmarking framework that could help evaluators understand the results [133]. For example, a building that consumed more energy than its peers would obtain a low score. Geyer et al. proposed a component-based methodology that predicted the heat flow of envelopes, heating/cooling demand, and final energy consumption by stages. DNN models were used for prediction in each stage, and LIME was employed to interpret the model output [131]. Besides, LIME was also used for other building management-related applications such as distributed PV power prediction [102], electricity demand prediction [124], and indoor CO_2 concentration prediction [122].

Although LIME is a model-agnostic technique suitable for any ML model, the interpretation obtained from LIME depends on ML models. Madhikermi et al. found that the interpretation of ANN and SVM models using LIME differs. For example, in FDD, the temperature of supply air after the heat recovery unit is the most influential feature for ANN, while the temperature of waste air is the most influential feature for SVM [96].

4.3.2. Shapley additive explanations (SHAP)

SHAP is also a model-agnostic tool proposed by Lundberg and Lee in 2017 to interpret individual predictions [143]. SHAP computes Shapley

values of each feature representing marginal contribution using a conditional expectation function. As shown in Fig. 13, SHAP is the most popular post-hoc technique in our reviewed literature. Although SHAP is designed as a local interpretability tool, the aggregation of Shapley values can be regarded as a global interpretation. For example, Carlsson et al. used the average of Shapley values as feature importance of the ANN model and found the most important features for energy consumption [91,92]. Ugwuanyi also used the average Shapley values for the global interpretation of CO_2 prediction [122].

Similar to LIME, SHAP is suitable for explaining the influential features of fault detection. In the study [105], SHAP generated the local and global interpretations of RF for FDD in district heating systems. Local interpretations revealed the influential features for individual prediction, while global interpretation showed the overall impact of each feature in the black-box model. Gao et al. used SHAP to interpret RF and LightGBM models for chiller FDD [130]. Santos et al. adopted XGBoost to detect fraud electricity consumption in the market, and SHAP was used to build interpretations for fraud activities afterward [119]. Additionally, SHAP can be used to interpret time-series classification for building energy consumption [124].

SHAP was also applied to occupancy-related studies such as CO_2 concentration prediction [122]. To interpret the sophisticated numerical equations-based thermal comfort model, Zhang et al. adopted SHAP for the local interpretation of individual PMV output [109]. According to the SHAP values, the authors proposed possible solutions to improve thermal comfort in different weather scenarios. In [117], the authors used the average SHAP values to rank the feature importance for natural ventilation rate prediction. The most influential features were pressure difference, outdoor temperature, wind speed, etc. In addition, the authors also provided the plot of individual SHAP values as local interpretation. To improve the interpretability of black-box models for thermal comfort prediction, Yang et al. adopted SHAP to generate both local and global interpretations [140]. A case study was conducted to interpret three thermal sensation models: hot, neutral, and cold. The results showed that air temperature and relative humidity were the most influential features for all three models.

For building energy benchmarking, SHAP can determine the key features contributing to high or low energy usage intensity of individual buildings. In study [97], SHAP was used to interpret the XGBoost-based residential building energy benchmarking model in New York. According to SHAP values, unit density was the strongest predictor for energy use intensity of residential buildings in New York with the highest positive correlation, followed by property assessed value and number of floors. Arjunan et al. improved the interpretability of a traditional benchmarking method named EnergyStar by combining the XGBoost algorithm and SHAP interpretable ML framework [99].

Load/power prediction is the most popular application for SHAP. Chang et al. adopted SHAP to provide the interpretability analysis to reveal feature importance for PV power generation models (TS-SOM and XGBoost) [100]. Results showed that global horizontal irradiance for center value was the most influential feature, which was consistent with the Pearson correlation analysis. Movahedi and Derriblea investigated the interpretation and interrelationship of three prediction models (electricity, water, and gas consumption) using SHAP [104], and results showed that the type of buildings (i.e., residential buildings or commercial buildings) and water consumption were the most influential feature for electricity prediction. They also found that gas and water consumption were strongly interrelated because gas was used for water heating in target buildings. Bellahsen and Dagdougui used SHAP to rank the feature importance as global interpretation. The three most influential features were historical loads right ahead of the forecasting time, one day, and one week ahead of the forecasting time [110]. Results also showed that the RF model relied heavier on historical features instead of datetime features than other models. According to the SHAP values from the XGBoost model, Chakraborty et al. found that single-family homes were likely to have a more significant increase in building cooling energy consumption under the context of global climate change [111]. Besides, buildings in hot-humid zones would consume more energy for cooling because of global warming. SHAP was adopted to interpret the performance-related indices (i.e., cooling capacity, COP, and wet/dew point efficiency) of a dew point cooler predicted by DNN in [112]. For example, a sample had a higher cooling capacity than the base value because of the relatively high intake air velocity. In [116], SHAP values showed that load and solar generation one hour ahead and the solar irradiance were the top three influential features for hourly ahead distributed PV power prediction. Similarly, Li and Wang summarized that day-ahead energy consumption was the most influential for daily load prediction [135].

4.3.3. Visualization and partial dependency plot

Visualization is a useful technique for users to build a better understanding of black-box models. The t-distributed stochastic neighbor embedding (t-SNE) creates two-dimensional projections for high-dimensional data using a non-linear transformation. Visualizing the embedding or hidden layer of neural networks using t-SNE has been widely adopted because it helps reveal the hidden mechanisms within neural networks. Kim and Cho added the state transition that can be visualized using t-SNE in the autoencoder model to improve the interpretability of electricity demand prediction results [94,114]. In [113], the authors visualized the latent states of autoencoders using t-SNE to explain the possible reasons for high or low energy consumption prediction. Singaravel et al. did similar research by visualizing the embedding layer of the CNN model to improve the understanding of building peak load prediction [106]. It was found that models with good generalization had higher separability than models with poor generalization when plotting using t-SNE.

Heatmap is another commonly used visualization tool that reveals the magnitude of a phenomenon in two dimensions. In [95], Kim and Cho analyzed class activation heatmaps to explore the influential features for load prediction. They found that one of the sub-metering related to an electric water heater and an air conditioner was the most noteworthy feature. Based on a heatmap interpretation tool for DNN named Grad-CAM, Li et al. proposed a modified variant to obtain fault-discriminative information from the one-dimensional CNN for chiller fault diagnosis [115]. To improve the interpretability of LSTM-based electricity load prediction, Kim and Cho proposed a deep learning model that can visualize and analyze the correlation between latent variables and output. The results showed that the two latent variables had different time dependencies, i.e., short-term and long-term dependencies [101].

The partial dependency plot (PDP) is a visualization tool that generates global interpretations for black-box models. PDP measures the effect of a feature by averaging the marginal distribution of other features for the entire dataset. PDP shows the overall effect, whereas the individual conditional expectation (ICE) plot visualizes the impact of a feature for each sample. The limitation of PDP and ICE is that they assume that input features are uncorrelated. In the study [103], PDP indicating the global effect of input showed that indoor temperature was the most influential feature for setpoint recommendation in DRL. Additionally, the ICE plot revealed the feature variation impact of inputs by showing the control upper and lower bound. Overall, most interpretations from the DRL model were consistent with domain knowledge. Zhang et al. adopted PDP to assess the marginal impact of each input feature in the thermal comfort model [109]. They concluded that the marginal impact of each feature was different, and most features had a positive impact on PMV value. In the study [126], PDP was employed for feature importance, and the impact of floor area was much larger than building ID because the PDP curve of floor area had a larger variation. Mouakher et al. found that the dwelling type and the number of bedrooms were influential features for energy consumption prediction according to the PDP of the LSTM load prediction model [138].

4.3.4. Other techniques

Apart from the above-mentioned post-hoc techniques, some studies adopted less-popular techniques. Many studies adopted various techniques to obtain the importance of features. For example, permutation importance calculates the importance of a feature by shuffling the values of the feature. The feature is important if the model prediction shows a significant error when shuffling a feature. Carlsson et al. evaluated the importance of features of an ANN for electricity prediction using permutation importance [91,92]. Similar to PDP, permutation importance is biased when features have a strong correlation. Zhang et al. adopted a dimensionless sensitivity index to quantify the feature importance, and the results showed that time-lag features of cooling load were more influential than other features [108]. Kim and Cho used Kullback–Leibler divergence to measure the relevance of features on prediction using latent states of the encoder-decoder model [135]. Tree-based methods can evaluate the importance of each feature by calculating the contribution of each feature to decrease impurity within the tree model, such as RF [133,137], gradient boosting machine (GBM) [137], XGBoost [98,137], and Cubist [136]. In [136], Cubist regression was used for building load prediction. After the importance of features is ranked, the authors found that outdoor air temperature and holiday index were influential external factors, and one day-ahead and one week-ahead energy load were influential internal factors.

The gradients of neural networks contain the importance of input features to the model output. Sipple et al. proposed an unsupervised anomaly detection method to detect the failures of power meter devices in 145 office buildings and adopted integrated gradients approach to interpret anomalies by analyzing the most influential input dimensions for normal and abnormal samples [107]. Similar work was done by Wang et al. by employing gradients of the ANN models, which quantified the marginal impact of the feature on the prediction based on the backpropagation rule [123]. Besides, gradients of the ANN were used to select important features in their study.

Counterfactual explanation generates local interpretation of a sample by creating nearby samples with the smallest changes in features that change the model output. Sakkas et al. first selected features via statistical analysis and then used the Diverse Counterfactual Explanation (DiCE) framework to perform counterfactual analysis for interpreting energy demand forecasting [118].

ELI5, short for *explain like I'm five*, is a Python package aiming to interpret popular black-box machine learning models such as XGBoost, LightGBM, CatBoost, Keras, and Scikit-learn. Sarp et al. used ELI5 to interpret the XGBoost model to facilitate the deployment of the ML-based renewable energy prediction model [120]. Time index and irradiance were the most influential feature for the XGBoost model to predict solar power generation overall. Besides the global interpretation, the authors also investigated the local interpretation of two samples using ELI5. Similarly, Kuzlu et al. used ELI5 to interpret renewable energy prediction models [102].

Some studies attempted to approximate black-box models for building energy management using simple and interpretable surrogate models. Zhang et al. trained a rule-set surrogate model to replace the RF model for building energy prediction [126]. The study [115] used a surrogate model to replace the multi-objective optimization algorithm for HVAC setpoint control. Moreover, results showed that simple DT-like rule sets could achieve about 90% of the detailed model predictive controller performance and save substantial computational costs. In the study [139], a novel rule extraction algorithm was used to interpret the global features of the load prediction model.

4.3.5. Summary

Post-hoc interpretability can evaluate and compare different machine learning models, allowing for selecting the most effective model for a given building energy management problem. According to Fig. 12, LIME and SHAP are the two most popular post-hoc techniques for explaining the predictions or decisions of ML models. Both techniques are model-agnostic and have their pros and cons. Although LIME is computationally efficient, the interpretation obtained from LIME depends on ML models, meaning that LIME has poor instability compared with SHAP [143]. SHAP, owing to the concept of cooperative game theory, has better stability and fairness. Another advantage of SHAP is its flexibility in generating both global and local interpretations. However, there are also some problems with SHAP. First, the computational cost is high if the number of features increases. Second, SHAP does not provide a surrogate model like LIME, so it cannot be used to evaluate how an increase or decrease in a particular feature will change the output. Lastly, the current visualization of SHAP does not have good readability compared with traditional visualization tools such as line diagrams [124].

5. Discussions

5.1. Benefits of interpretable ML

Using interpretable machine learning in building energy efficiency and building energy flexibility offers several potential benefits. First, by using interpretable ML algorithms, building managers can gain a better understanding of the factors that affect energy consumption prediction [47]. This can help them develop more accurate and efficient energy management strategies, reducing energy waste and lowering operating costs. Second, interpretable ML techniques can be used to identify the factors that affect building energy consumption and production and to develop strategies for increasing the flexibility of building energy systems [99,133]. Third, interpretable ML algorithms can provide explanations for their predictions, which can help building managers understand the reasoning behind their energy management decisions [121]. This can promote transparency and accountability in decision-making and enable building managers to make more informed and effective decisions.

5.2. Limitations and challenges of interpretable ML

Although interpretable ML has gained increasing attention in building energy management in recent years, its broad application is faced with several challenges based on this literature review.

The first challenge is related to the various terminologies adopted to describe the interpretability of ML models. In the literature, the most commonly used terminology is interpretable machine learning (e.g., [93]) and explainable artificial intelligence (e.g., [105]), and some studies have adopted less popular terminology such as model interpretability (e.g., [79]) explainable machine learning (e.g., [72]). Some studies adopted terminology related to specific applications, such as interpretable building energy benchmarking [133]. Although interpretability and explainability are often used interchangeably, there is a subtle difference between interpretability and explainability. Interpretability represents the extent to which the model input and output can be observed in a causal way. For example, linear regression determines the coefficient for each input feature, representing the causal impact of each input feature from the viewpoint of the regression model [144]. As shown in Fig. 1, models with high accuracy usually have sophisticated structures and low interpretability. In other words, a model with high interpretability can be understood by users intuitively, but this does not indicate the degree to which the model approximates reality. Explainability, on the other hand, is about the explanation for the prediction and the reason why users should trust the model. For example, the proposed building energy benchmarking approach can explain why a building achieves a particular score using SHAP [99]. Therefore, researchers may find it confusing to collect related studies and conduct exhaustive reviews, which is the intention of this review.

The difficulty in describing interpretability also leads to incomparability between studies dealing with the same tasks. The traditional ML paradigm can use model accuracy or error metrics to compare the

performance of the two methods. However, for interpretable ML, the interpretability of the two methods is usually difficult to measure and compare, especially for ante-hoc techniques. Besides, model interpretations should be consistent and stable [25]. Consistency means the explanations for two models trained on the same task and similar predictions should be similar. Stability requires explanations for similar samples (with similar features) are similar. Up to now, only a few studies have analyzed the consistency and stability of post-hoc techniques. In [117], it was observed that feature rankings for RF, kNN, DNN, and linear regression (LR) using SHAP were slightly different because of the different structures of the models. Sakkas et al. used counterfactual explanations to perform counterfactual analysis for local interpretability analysis. They found that the counterfactuals yielded from the linear regression and ANN models were quite similar despite the huge structure differences between the two models [118]. Wastensteiner performed the most in-depth experiment to analyze the consistency and stability of different post-hoc techniques. He compared the consistency and stability of LIME and SHAP in forecasting building energy consumption, and the results showed that SHAP outperformed LIME in both metrics. Gao et al. compared the post-hoc interpretation from RF and LightGBM using SHAP [130]. They adopted SHAP to interpret two different diagnosis models (RF and LightGBM). It was found that the influential features of the model output were similar, although the two models had different structures.

Another challenge is related to the limitations of current interpretable ML techniques. LIME and SHAP are two popular techniques suitable for all ML models. Ugwuanyi did a non-expert survey to test the interpretability of LIME and SHAP for CO_2 prediction. All testers responded that SHAP is better than LIME in terms of the readability of the output [122]. Kuzlu et al. also compared three interpretable ML techniques (LIME, SHAP, and ELI5) for interpreting renewable energy prediction [102]. They summarized that each technique had its advantages and disadvantages. Generally, SHAP provides a more detailed interpretation but is much slower than LIME and ELI5. Besides, LIME is criticized for the instability of the explanations [93,143]. On the other hand, SHAP is computationally intensive, especially for large and complex machine learning models, making it impractical to use in some large and complex applications. As discussed in Section 4.3.4, some studies adopted feature importance from tree-based models as global interpretation. However, Chang et al. stated that using feature importance ranking methods in tree-based models such as RF and XGBoost is not reasonable because it violates the consistency principle of feature importance [100]. To this end, they adopted SHAP to reveal the most influential features of PV power generation. PDP and ICE are intuitive for global and local interpretation of black-box models, but they assume that features are independently distributed. The common disadvantage of the above-mentioned post-hoc techniques is the neglect of the correlation among features. Although powerful tools such as LIME and SHAP can provide each feature's positive/negative impact, they cannot offer in-depth explanations for black-box models considering the possible correlation among features.

5.3. Research directions

Based on the literature, ante-hoc and post-hoc techniques have shown their power to improve interpretability. In future studies, we believe interpretable ML will be employed extensively in various building energy management-related applications because of the compelling need for model interpretability with the rapid development and broad deployment of ML. Several research directions are identified for maximizing the values of interpretable ML in building energy management.

The first research direction is to dig into the interpretability of ML-based classification tasks for building energy management such as FDD. In the literature, most studies, with 75 out of 91 studies, focus on regression tasks such as load/power predictions that usually provide global interpretations to explain the most influential features for predictions. Only 13 of 91 papers deal with classification-related applications including FDD and load profiling. As demonstrated in other fields such as healthcare and biomedicine, interpretability is highly valuable to classification tasks as it can provide more straightforward explanations for ML-based decisions. For example, global interpretations can reveal the most important sensor measurements or variables for a fault in FDD applications, which is very useful for diagnosing the fault in facility management. System operators usually have a stronger need for local interpretation of classification tasks such as FDD. For example, operators need explanations for a fault alarm, but they usually do not need to know the specific reason for cooling load prediction. Furthermore, extracting interpretable rules from black-box models for classification tasks is valuable to overcome the difficulties of expert rule-based FDD in real applications due to system diversities and uncertainties.

Another practical need for interpretable ML in building energy control systems is the ability to understand the control strategies recommended by black-box models. Most ML-based control strategies/policies simply adopt the black-box prediction models to construct optimal objective functions based on physical knowledge, such as minimizing energy consumption while maintaining thermal comfort. Strictly speaking, the control strategies/policies are not learned by ML. As a result, the interpretability of those ML-based controls is a problem of interpretability of the prediction models. DRL is believed to be a promising method for efficient and flexible control in buildings [145,146] that learns control policies from the interactions between the reinforcement models/agents and the controlled environment. However, one challenge in deploying DRL models is the lack of transparency in the learning process. The control strategies recommended by DRL models may not be easily understood by building operators or control engineers. This makes it difficult to implement and trust the model's recommendations, especially in safety-critical cases such as demand control ventilation [147,148]. To improve the interpretability of DRL-based control, the ante-hoc approach can be adopted by integrating domain knowledge into the DRL models during the model structure design or training process (such as in [52]). As for the post-hoc approach, tools such as SHAP can provide clear and transparent explanations of the reasons behind the model's recommendations [149]. In addition, surrogate models (refer to Section 2.2.1) can be developed to obtain the rules and decision boundaries of the DRL model. These measures to improve interpretability build trust in the model and make it easier for building operators and control engineers to understand and implement the control strategies.

The interpretation of ML models can be customized for different stakeholders and applications in future studies. The necessity and aims of adopting interpretable ML are usually not explained clearly in the literature. In practice, the level of interpretability required depends on the specific application and the needs of the situation. For example, control engineers may need a higher level of interpretability to understand the decisions and actions being taken by the model. At the same time, ML researchers may be more interested in the underlying algorithms and data used to train the model. Building operators and owners may be more interested in the practical implications of the model, such as how it will affect the building's energy consumption and costs. For example, when poor energy performance such as low COP is detected in chillers using data-driven methods (such as in [93]), the operators need the details such as specific fault location and cause to support the detection. More importantly, possible solutions to improve the efficiency of chillers. Regulators and policymakers, in contrast, emphasize model transparency more. In some applications, such as building energy benchmarking, more attention should be paid to fairness and stability in interpretability evaluation. Therefore, model explanations will have better applications if the aims and end-users are identified clearly. To conclude, the specific approach to defining the needs for interpretability will depend on the specific context and requirements of the situation.

Lastly, more benchmark datasets of good quality should be developed in typical applications to enhance the comparability between model explanations and boost the application of interpretable ML for

building energy management. Interpretable ML is not as mature as general ML that can be evaluated by accuracy (RMSE, MAE, CVRMSE, etc.), computation efficiency, generalization capability, etc. In the existing literature, the performance of interpretable ML has seldom been evaluated, as interpretability is difficult to quantify. *Building Data Genome Project 2* is the largest open dataset for building energy prediction containing energy meters from 1636 buildings [150]. However, there are many missing values and outliers; thus, some studies only used part of the dataset. Li et al. developed a synthetic building operation dataset for building energy benchmarking, thermal and energy load prediction, model predictive control, etc. [151]. The entire dataset was created using EnergyPlus simulation, so its application is limited. In addition, there are some open datasets for FDD [152,153] and energy benchmarking [154]. However, these datasets are suitable for developing data-driven models but not tailored for evaluating and comparing model explanations. Dedicated datasets for evaluating the performance of interpretable ML should design not only training/test data but also typical regression/classification tasks according to practical engineering needs. For example, the dataset for chiller FDD should include necessary features such as temperature and water flowrate in the evaporate/condenser. In addition, typical fault detection/diagnosis tasks with various data availability of labeled data should be designed to evaluate the generalization ability of models [41]. Based on the dedicated datasets, evaluation metrics that quantify the quality of model explanations should also be developed accordingly, such as stability and consistency.

6. Conclusions

This article provides a comprehensive review of previous literature that utilizes interpretable ML techniques for building energy management, which is a novel and promising research topic to facilitate the adoption and deployment of ML models in buildings. First, the applications of machine learning for building energy management in previous studies are presented. Although ML algorithms have demonstrated flexibility in dealing with various tasks, criticism against the broad application of ML exists because of the black-box nature of ML models. The issues of the traditional ML paradigm call for the interpretability of ML, i.e., interpretable ML. Second, the basics of interpretable ML are introduced according to the taxonomy from the perspective of computer science. Then, the keywords and searching methods are presented.

After the initial and final selection, 91 papers are analyzed in this review. The papers are divided into two categories: ante-hoc and post-hoc. In each category, the papers are analyzed according to the type of interpretable techniques. The main findings include: (1) load/power prediction is the most popular application, followed by FDD in both ante-hoc and post-hoc categories, (2) the interpretability of neural networks can be improved by adding physical characteristics, integrating domain knowledge, or adopting attention mechanisms, (3) most reviewed studies focus on global interpretability models, i.e., evaluating the overall impact of each feature on the model predictions, and (4) SHAP and LIME are two most frequently used post-hoc techniques owing to its applicability to all ML models.

Although interpretable ML improves the trust in the application of ML for building energy management, the research is still in its infancy and faces significant challenges including (1) the difficulty in describing interpretability, (2) the difficulty in comparing model explanations between studies dealing with the same tasks, and (3) current techniques such as SHAP and LIME cannot provide extensive explanations. To fully leverage the power of interpretable ML and promote its application, we propose a few possible opportunities: (1) in-depth research can be carried out for classification tasks such as FDD, (2) interpretable DRL models can be developed for building control, (3) customized interpretable models can be established for different end-users and applications, and (4) open datasets with typical tasks tailored for interpretable building energy management are needed.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

Acknowledgement

The authors gratefully acknowledge the support of this research by the National Key Research and Development Program of China (2021YFE0107400), and the Research Grant Council of the Hong Kong SAR (C5018–20GF).

References

[1] IEA 2021 Global status report for buildings and construction. U N Environ Programme; 2021.
[2] Ramesh T, Prakash R, Shukla KK. Life cycle energy analysis of buildings: an overview. Energy Build 2010;42:1592–600. doi:10.1016/j.enbuild.2010.05.007.
[3] Zhou N, Khanna N, Feng W. Policy roadmap to 50% energy reduction in Chinese buildings by 2050. Procs ACEEE Summer Study Energy Effic Build 2016 9–1.
[4] Environmental Bureau HKSAR. Hong Kong's Climate Action Plan 2050. 2021.
[5] Langevin J, Harris CB, Reyna JL. Assessing the Potential to Reduce U.S. Building CO2 Emissions 80% by 2050. Joule 2019;3:2403–24. doi:10.1016/j.joule.2019.07.013.
[6] European Environment Agency. Greenhouse gas emissions from energy use in buildings in Europe n.d. https://www.eea.europa.eu/data-and-maps/indicators/greenhouse-gas-emissions-from-energy/assessment (accessed October 20, 2022).
[7] Hurtado LA, Mocanu E, Nguyen PH, Gibescu M, Kamphuis RIG. Enabling cooperative behavior for building demand response based on extended joint action learning. IEEE Trans Ind Inform 2018;14:127–36. doi:10.1109/TII.2017.2753408.
[8] Tang H, Wang S, Li H. Flexibility categorization, sources, capabilities and technologies for energy-flexible and grid-responsive buildings: state-of-the-art and future perspective. Energy 2021;219:119598. doi:10.1016/j.energy.2020.119598.
[9] Chen Y, Chen Z, Yuan X, Su L, Li K. Optimal control strategies for demand response in buildings under penetration of renewable energy. Buildings 2022;12:371. doi:10.3390/buildings12030371.
[10] Capozzoli A, Cerquitelli T, Piscitelli MS. Chapter 11 - enhancing energy efficiency in buildings through innovative data analytics technologiesa. In: Pervasive computer. Boston: Academic Press; 2016. p. 353–89. doi:10.1016/B978-0-12-803663-1.00011-5.
[11] IBM. Building artificial intelligence into buildings. https://www.ibm.com/thought-leadership/institute-business-value/report/buildingintelligence} (accessed October 20, 2022).
[12] Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y, et al. A review of machine learning in building load prediction. Appl Energy 2021;285:116452. doi:10.1016/j.apenergy.2021.116452.
[13] Chen Y, Guo M, Chen Z, Chen Z, Ji Y. Physical energy and data-driven models in building energy prediction: a review. Energy Rep 2022;8:2656–71. doi:10.1016/j.egyr.2022.01.162.
[14] Chen Z, Chen Y, He R, Liu J, Gao M, Zhang L. Multi-objective residential load scheduling approach for demand response in smart grid. Sustain Cities Soc 2022;76:103530. doi:10.1016/j.scs.2021.103530.
[15] Zhao H, Magoulès F, Liu X, Zhao T, Chang C-T, Fu CJ, et al. A review on renewable energy and electricity requirement forecasting models for smart grid and buildings. Sustain Cities Soc 2020;55:102052. doi:10.1016/j.scs.2020.102052.
[16] Bahani K, Ali-Ou-Salah H, Moujabbir M, Oukarfi B, Ramdani M. A novel interpretable model for solar radiation prediction based on adaptive fuzzy clustering and linguistic hedges. New York, NY, USA: Association for Computing Machinery; 2020. p. 1–6. proc. 13th int. conf. intell. syst. theor. appl. doi:10.1145/3419604.3419807.
[17] Li G, Li F, Xu C, Fang X. A spatial-temporal layer-wise relevance propagation method for improving interpretability and prediction accuracy of LSTM building energy prediction. Energy Build 2022;271:112317. doi:10.1016/j.enbuild.2022.112317.
[18] Mariano-Hernández D, Hernández-Callejo L, Zorita-Lamadrid A, Duque-Pérez O, Santos García F. A review of strategies for building energy management system: model predictive control, demand side management, optimization, and fault detect & diagnosis. J Build Eng 2021;33:101692. doi:10.1016/j.jobe.2020.101692.
[19] Gaur M, Makonin S, Bajić IV, Majumdar A. Performance evaluation of techniques for identifying abnormal energy consumption in buildings. IEEE Access 2019;7:62721–33. doi:10.1109/ACCESS.2019.2915641.
[20] Qavidel Fard Z, Zomorodian ZS, Korsavi SS. Application of machine learning in thermal comfort studies: a review of methods, performance and challenges. Energy Build 2022;256:111771. doi:10.1016/j.enbuild.2021.111771.
[21] Dai X, Liu J, Zhang X. A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings. Energy Build 2020;223:110159. doi:10.1016/j.enbuild.2020.110159.

[22] Amayri M, Ali S, Bouguila N, Ploix S. Machine learning for activity recognition in smart buildings: a survey. In: Energy smart homes algorithms technology application. Cham: Springer International Publishing; 2021. p. 199–228. doi:10.1007/978-3-030-76477-7_6.

[23] Khaoula E, Amine B, Mostafa B. Machine learning and the internet of things for smart buildings: a state of the art survey. In: Proceedings of the 2nd international conference on innovative research in applied science, engineering and technology. IRASET; 2022. p. 1–10. doi:10.1109/IRASET52964.2022.9738256.

[24] Javed A, Larijani H, Wixted A. Improving energy consumption of a commercial building with IoT and machine learning. IT Prof 2018;20:30–8. doi:10.1109/MITP.2018.053891335.

[25] Molnar C. Interpretable machine learning. Morisville, North Carolina: Lulucom; 2019.

[26] Vollert S, Atzmueller M, Theissler A. Interpretable Machine Learning: a brief survey from the predictive maintenance perspective. In: Proceedings of the 26th IEEE international conference on emerging technologies and factory automation ETFA; 2021. p. 01–8. doi:10.1109/ETFA45728.2021.9613467.

[27] Naser MZ. An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: navigating causality, forced goodness, and the false perception of inference. Autom Constr 2021;129:103821. doi:10.1016/j.autcon.2021.103821.

[28] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. doi:10.1016/j.inffus.2019.12.012.

[29] Thampi A. Interpretable AI. Manning 2022.

[30] Fei J, Chen Y, Liu L, Fang Y. Fuzzy multiple hidden layer recurrent neural control of nonlinear system using terminal sliding-mode controller. IEEE Trans Cybern 2022;52:9519–34. doi:10.1109/TCYB.2021.3052234.

[31] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res 2021;70:245–317. doi:10.1613/jair.1.12228.

[32] Yan K, Zhong C, Ji Z, Huang J. Semi-supervised learning for early detection and diagnosis of various air handling unit faults. Energy Build 2018;181:75–83. doi:10.1016/j.enbuild.2018.10.016.

[33] Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics 2022;12:237. doi:10.3390/diagnostics12020237.

[34] Mahapatra D. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. ArXiv210406087 Cs 2021.

[35] Lötsch J, Kringel D, Ultsch A. Explainable artificial intelligence (XAI) in biomedicine: making ai decisions trustworthy for physicians and patients. BioMedInformatics 2022;2:1–17. doi:10.3390/biomedinformatics2010001.

[36] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell 2020;2:573–84. doi:10.1038/s42256-020-00236-4.

[37] Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Min Knowl Discov 2020;10:e1379. doi:10.1002/widm.1379.

[38] Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. Explainable AI (xAI) for Anatomic Pathology. Adv Anat Pathol 2020;27:241–50. doi:10.1097/PAP.0000000000000264.

[39] Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci 2021;11:5088. doi:10.3390/app11115088.

[40] Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable artificial intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities. Energy AI 2022;9:100169. doi:10.1016/j.egyai.2022.100169.

[41] Li B, Cheng F, Zhang X, Cui C, Cai W. A novel semi-supervised data-driven method for chiller fault diagnosis with unlabeled data. Appl Energy 2021;285:116459. doi:10.1016/j.apenergy.2021.116459.

[42] Fan C, Liu X, Xue P, Wang J. Statistical characterization of semi-supervised neural networks for fault detection and diagnosis of air handling units. Energy Build 2021;234:110733. doi:10.1016/j.enbuild.2021.110733.

[43] Kamath U, Liu J. Explainable artificial intelligence: an introduction to interpretable machine learning. Cham: Springer International Publishing; 2021. doi:101007/978-3-030-83356-5.

[44] Sha H, Xu P, Lin M, Peng C, Dou Q. Development of a multi-granularity energy forecasting toolkit for demand response baseline calculation. Appl Energy 2021;289:116652. doi:10.1016/j.apenergy.2021.116652.

[45] Zhuang H, Wang X, Bendersky M, Grushetsky A, Wu Y, Mitrichev P, et al. Interpretable ranking with generalized additive models. In: Procededings of the 14th ACM international conference on web search and data mining. Association for Computing Machinery; 2021. p. 499–507. doi:10.1145/3437963.3441796.

[46] Interpretability n.d. https://ww2.mathworks.cn/en/discovery/interpretability.html (accessed October 20, 2022).

[47] Li A, Xiao F, Zhang C, Fan C. Attention-based interpretable neural network for building cooling load prediction. Appl Energy 2021;299:117238. doi:10.1016/j.apenergy.2021.117238.

[48] Qi Z, Khorram S, Li F. Visualizing deep networks by optimizing with integrated gradients. CVPR Workshop 2019;2.

[49] Toubeau J-F, Bottieau J, Wang Y, Vallee F, Shan S, Cao B, et al. Forecasting the short-term electricity consumption of building using a novel ensemble model. IEEE Access 2019;7:88093–106. doi:10.1109/ACCESS.2019.2925740.

[50] Kim E. Interpretable and accurate convolutional neural networks for human activity recognition. IEEE Trans Ind Inform 2020;16:7190–8. doi:10.1109/TII.2020.2972628.

[51] Wang H, Cai R, Zhou B, Aziz S, Qin B, Voropai N, et al. Solar irradiance forecasting based on direct explainable neural network. Energy Convers Manag 2020;226:113487. doi:10.1016/j.enconman.2020.113487.

[52] Yu Z, Yang X, Gao F, Huang J, Tu R, Cui J. A Knowledge-based reinforcement learning control approach using deep Q network for cooling tower in HVAC systems. In: Proceedings of the Chinese automation congress CAC; 2020. p. 1721–6. doi:10.1109/CAC51589.2020.9327385.

[53] Zhang X, Chung F-L, Wang S. An interpretable fuzzy DBN-based classifier for indoor user movement prediction in ambient assisted living applications. IEEE Trans Ind Inform 2020;16:42–53. doi:10.1109/TII.2019.2912625.

[54] Chen Y, Zhang D. Theory-guided deep-learning for electrical load forecasting (TgDLF) via ensemble long short-term memory. Adv Appl Energy 2021;1:100004. doi:10.1016/j.adapen.2020.100004.

[55] Drgoňa J, Tuor AR, Chandan V, Vrabie DL. Physics-constrained deep learning of multi-zone building thermal dynamics. Energy Build 2021;243:110992. doi:10.1016/j.enbuild.2021.110992.

[56] Li L., Yan J., Yang X., Jin Y. learning interpretable deep state space model for probabilistic time series forecasting. ArXiv210200397 Cs Stat 2021.

[57] Oreshkin BN, Dudek G, Pełka P, Turkina E. N-BEATS neural network for mid-term electricity load forecasting. Appl Energy 2021;293:116918. doi:10.1016/j.apenergy.2021.116918.

[58] Chen Z, Xiao F, Shi J, Li A. Dynamic model development for vehicle air conditioners based on physics-guided deep learning. Int J Refrig 2022;134:126–38. doi:10.1016/j.ijrefrig.2021.11.021.

[59] Di Natale L., Svetozarevic B., Heer P., Jones C.N. Physically consistent neural networks for building thermal modeling: theory and analysis. ArXiv211203212 Cs Eess 2022.

[60] Guo T., Lin T., Lu Y. An interpretable LSTM neural network for autoregressive exogenous model. ArXiv180405251 Cs Stat 2018.

[61] Li D, Li D, Li C, Li L, Gao L. A novel data-temporal attention network based strategy for fault diagnosis of chiller sensors. Energy Build 2019;198:377–94. doi:10.1016/j.enbuild.2019.06.034.

[62] Gangopadhyay T, Tan SY, Jiang Z, Sarkar S. Interpretable deep attention model for multivariate time series prediction in building energy systems. In: Lecture notes computer science. Cham: Springer International Publishing; 2020. p. 93–101. doi:10.1007/978-3-030-61725-7_13.

[63] Gangopadhyay T, Tan SY, Jiang Z, Meng R, Sarkar S. Spatiotemporal attention for multivariate time series prediction and interpretation. In: Proceedings of the ICASSP IEEE the international conference on acoustics, speech, & signal processing. ICASSP; 2021. p. 3560–4. doi:10.1109/ICASSP39728.2021.9413914.

[64] Lim B., Arik S.O., Loeff N., Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. ArXiv191209363 Cs Stat 2020.

[65] Azam MF, Younis MS. Multi-horizon electricity load and price forecasting using an interpretable multi-head self-attention and EEMD-based framework. IEEE Access 2021;9:85918–32. doi:10.1109/ACCESS.2021.3086039.

[66] Gao Y, Ruan Y. Interpretable deep learning model for building energy consumption prediction based on attention mechanism. Energy Build 2021;252:111379. doi:10.1016/j.enbuild.2021.111379.

[67] Toubeau J-F, Bottieau J, Wang Y, Vallee F. Interpretable probabilistic forecasting of imbalances in renewable-dominated electricity systems. IEEE Trans Sustain Energy 2022;13:1267–77. doi:10.1109/TSTE.2021.3092137.

[68] Li C, Dong Z, Ding L, Petersen H, Qiu Z, Chen G, et al. Interpretable memristive LSTM network design for probabilistic residential load forecasting. IEEE Trans Circuits Syst Regul Pap 2022;69:2297–310. doi:10.1109/TCSI.2022.3155443.

[69] Bhatia A, Garg V, Haves P, Pudi V. Explainable clustering using hyper-rectangles for building energy simulation data. IOP Conf Ser Earth Environ Sci 2019;238:012068. doi:10.1088/1755-1315/238/1/012068.

[70] Grimaldo AI, Novak J. User-centered visual analytics approach for interactive and explainable energy demand analysis in prosumer scenarios. In: Computer vision systems. Cham: Springer International Publishing; 2019. p. 700–10. vol. 11754. doi:10.1007/978-3-030-34995-0_64.

[71] Laurinec P, Lucká M. Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting. Data Min Knowl Discov 2019;33:413–45. doi:10.1007/s10618-018-0598-2.

[72] Miller C. What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification. Energy Build 2019;199:523–36. doi:10.1016/j.enbuild.2019.07.019.

[73] Grimaldo AI, Novak J. Combining machine learning with visual analytics for explainable forecasting of energy demand in prosumer scenarios. Procedia Comput Sci 2020;175:525–32. doi:10.1016/j.procs.2020.07.074.

[74] Kasuya T, Takeshi T, Esaki H. Building activity profiling: explainable and predictive modeling for building automation. In: Proceedings of the international conference on artificial intelligence in information and communication. ICAIIC; 2020. p. 242–7. doi:10.1109/ICAIIC48513.2020.9065268.

[75] Sun M, Zhang T, Wang Y, Strbac G, Kang C. Using bayesian deep learning to capture uncertainty for residential net load forecasting. IEEE Trans Power Syst 2020;35:188–201. doi:10.1109/TPWRS.2019.2924294.

[76] Chen Z, Chen Y, Xiao T, Wang H, Hou P. A novel short-term load forecasting framework based on time-series clustering and early classification algorithm. Energy Build 2021;251:111375. doi:10.1016/j.enbuild.2021.111375.

[77] Grimaldo A, Novak J. Explainable needn't Be (Much) less accurate: evaluating an explainable ai dashboard for energy forecasting. In: Artificial intelligence applications and innovations. AIAI 2021 IFIP WG 12.5 international workshops. Cham: Springer International Publishing; 2021. p. 340–51. doi:10.1007/978-3-030-79157-5_28.

[78] Hu M, Ge D, Telford R, Stephen B, Wallom DCH. Classification and characterization of intra-day load curves of PV and non-PV households using interpretable feature extraction and feature-based clustering. Sustain Cities Soc 2021;75:103380. doi:10.1016/j.scs.2021.103380.

[79] Xiao T, Xu P, Ding R, Chen Z. An interpretable method for identifying mislabeled commercial building based on temporal feature extraction and ensemble classifier. Sustain Cities Soc 2022;78:103635. doi:10.1016/j.scs.2021.103635.

[80] Castellini A, Bianchi F, Farinelli A. Generation and interpretation of parsimonious predictive models for load forecasting in smart heating networks. Appl Intell 2022;52:9621–37. doi:10.1007/s10489-021-02949-4.

[81] Pathak N, Ba A, Ploennigs J, Roy N. Forecasting gas usage for big buildings using generalized additive models and deep learning. In: Proceedings of the IEEE international conference on smart computing. SMARTCOMP; 2018. p. 203–10. doi:10.1109/SMARTCOMP.2018.00092.

[82] Charalampopoulos I. A comparative sensitivity analysis of human thermal comfort indices with generalized additive models. Theor Appl Climatol 2019;137:1605–22. doi:10.1007/s00704-019-02900-1.

[83] Khamma TR, Zhang Y, Guerrier S, Boubekri M. Generalized additive models: an efficient method for short-term energy prediction in office buildings. Energy 2020;213:118834. doi:10.1016/j.energy.2020.118834.

[84] Amato U, Antoniadis A, De Feis I, Goude Y, Lagache A. Forecasting high resolution electricity demand data with additive models including smooth and jagged components. Int J Forecast 2021;37:171–85. doi:10.1016/j.ijforecast.2020.04.001.

[85] Bujalski M, Madejski P. Forecasting of heat production in combined heat and power plants using generalized additive models. Energies 2021;14:2331. doi:10.3390/en14082331.

[86] Sundararajan A, Ollis B. Regression and generalized additive model to enhance the performance of photovoltaic power ensemble predictors. IEEE Access 2021;9:111899–914. doi:10.1109/ACCESS.2021.3103126.

[87] Voss M, Heinekamp JF, Krutzsch S, Sick F, Albayrak S, Strunz K. Generalized additive modeling of building inertia thermal energy storage for integration into smart grid control. IEEE Access 2021;9:71699–711. doi:10.1109/ACCESS.2021.3078802.

[88] Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate 2022. 10.48550/arXiv.1409.0473.

[89] Rojat T., Puget R., Filliat D., Del Ser J., Gelin R., Díaz-Rodríguez N. Explainable artificial intelligence (XAI) on TimeSeries Data: a survey. ArXiv210400950 Cs 2021.

[90] Fulcher BD, Jones NS. hctsa : a computational framework for automated time-series phenotyping using massive feature extraction. Cell Syst 2017;5:527–31 e3. doi:10.1016/j.cels.2017.10.001.

[91] Carlsson L, Samuelsson P, Jönsson P. Using interpretable machine learning to predict the electrical energy consumption of an electric arc furnace. Stahl Eisen 2019;139:24–9 1881.

[92] Carlsson LS, Samuelsson PB, Jönsson PG. Interpretable machine learning—tools to interpret the predictions of a machine learning model predicting the electrical energy consumption of an electric arc furnace. Steel Res Int 2020;91:2000053. doi:10.1002/srin.202000053.

[93] Fan C, Xiao F, Yan C, Liu C, Li Z, Wang J. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Appl Energy 2019;235:1551–60. doi:10.1016/j.apenergy.2018.11.081.

[94] Kim J-Y, Cho S-B. Electric energy consumption prediction by deep learning with state explainable autoencoder. Energies 2019;12:739. doi:10.3390/en12040739.

[95] Kim T-Y, Cho S-B. Predicting residential energy consumption using CNN-LSTM neural networks. Energy 2019;182:72–81. doi:10.1016/j.energy.2019.05.230.

[96] Madhikermi M, Malhi AK, Främling K. Explainable artificial intelligence based heat recycler fault detection in air handling unit. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Cham: Springer International Publishing; 2019. p. 110–25. doi:10.1007/978-3-030-30391-4_7.

[97] Papadopoulos S, Kontokosta CE. Grading buildings on energy performance using city benchmarking data. Appl Energy 2019;233–234:244–53. doi:10.1016/j.apenergy.2018.10.053.

[98] Wang R, Lu S, Li Q. Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. Sustain Cities Soc 2019;49:101623. doi:10.1016/j.scs.2019.101623.

[99] Arjunan P, Poolla K, Miller C. EnergyStar++: towards more accurate and explanatory building energy benchmarking. Appl Energy 2020;276:115413. doi:10.1016/j.apenergy.2020.115413.

[100] Chang X, Li W, Ma J, Yang T, Zomaya AY. Interpretable machine learning in sustainable edge computing: a case study of short-term photovoltaic power output prediction. In: Proceedings of the ICASSP IEEE international conference on acoustics, speech, and signal processing. ICASSP; 2020. p. 8981–5. doi:10.1109/ICASSP40776.2020.9054088.

[101] Kim J-Y, Cho S-B. Electric energy demand forecasting with explainable time-series modeling. In: Proceedings of the international conference on data mining workshop ICDMW; 2020. p. 711–16. doi:10.1109/ICDMW51313.2020.00101.

[102] Kuzlu M, Cali U, Sharma V, Guler O. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access 2020;8:187814–23. doi:10.1109/ACCESS.2020.3031477.

[103] Kotevska O, Munk J, Kurte K, Du Y, Amasyali K, Smith RW, et al. Methodology for interpretable reinforcement learning model for HVAC energy control. In: Proceedings of the IEEE international conference on big data. IEEE; 2020. p. 1555–64. doi:10.1109/BigData50022.2020.9377735.

[104] Movahedi A., Derrible S. Interrelated Patterns of electricity, gas, and water consumption in large-scale buildings 2020. 10.31224/osf.io/ahn3e.

[105] Park S, Moon J, Hwang E. Explainable anomaly detection for district heating based on shapley additive explanations. In: Proceedings of the international conference on data mining workshop ICDMW; 2020. p. 762–5. doi:10.1109/ICDMW51313.2020.00111.

[106] Singaravel S, Suykens J, Janssen H, Geyer P. Explainable deep convolutional learning for intuitive model development by non–machine learning domain experts. Des Sci 2020;6. doi:10.1017/dsj.2020.22.

[107] Interpretable SJ. Multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In: Proceedings of the 37th the international conference on machine learning. PMLR; 2020. p. 9016–9025.

[108] Zhang C, Li J, Zhao Y, Li T, Chen Q, Zhang X. A hybrid deep learning-based method for short-term building energy load prediction combined with an interpretation process. Energy Build 2020;225:110301. doi:10.1016/j.enbuild.2020.110301.

[109] Zhang W, Wen Y, Tseng KJ, Jin G. Demystifying thermal comfort in smart buildings: an interpretable machine learning approach. IEEE Internet Things J 2021;8:8021–31. doi:10.1109/JIOT.2020.3042783.

[110] Bellahsen A, Dagdougui H. Aggregated short-term load forecasting for heterogeneous buildings using machine learning with peak estimation. Energy Build 2021;237:110742. doi:10.1016/j.enbuild.2021.110742.

[111] Chakraborty D, Alam A, Chaudhuri S, Başağaoğlu H, Sulbaran T, Langar S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. Appl Energy 2021;291:116807. doi:10.1016/j.apenergy.2021.116807.

[112] Golizadeh Akhlaghi Y, Aslansefat K, Zhao X, Sadati S, Badiei A, Xiao X, et al. Hourly performance forecast of a dew point cooler using explainable Artificial Intelligence and evolutionary optimisations by 2050. Appl Energy 2021;281:116062. doi:10.1016/j.apenergy.2020.116062.

[113] Kim JY, Cho S-B. Interpretable deep learning with hybrid autoencoders to predict electric energy consumption. In: "Advances in intelligent systems and computing. Cham: Springer International Publishing; 2021. p. 133–43. doi:10.1007/978-3-030-57802-2_13.

[114] Kim J-Y, Cho S-B. Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space. Expert Syst Appl 2021;186:115842. doi:10.1016/j.eswa.2021.115842.

[115] Li G, Yao Q, Fan C, Zhou C, Wu G, Zhou Z, et al. An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems. Build Environ 2021;203:108057. doi:10.1016/j.buildenv.2021.108057.

[116] Lu Y, Murzakhanov I, Chatzivasileiadis S. Neural network interpretability for forecasting of aggregated renewable generation. In: Proceedings of the IEEE international conference on communications, control, and computing technologies for smart grids; 2021. p. 282–8. doi:10.1109/SmartGridComm51999.2021.9631993.

[117] Park H, Park DY. Comparative analysis on predictability of natural ventilation rate based on machine learning algorithms. Build Environ 2021;195:107744. doi:10.1016/j.buildenv.2021.107744.

[118] Sakkas N, Yfanti S, Daskalakis C, Barbu E, Domnich M. Interpretable forecasting of energy demand in the residential sector. Energies 2021;14:6568. doi:10.3390/en14206568.

[119] Santos RN, Yamouni S, Albiero B, Vicente R, Silva J, Souza T, et al. Gradient boosting and Shapley additive explanations for fraud detection in electricity distribution grids. Int Trans Electr Energy Syst 2021;31:e13046. doi:10.1002/2050-7038.13046.

[120] Sarp S, Kuzlu M, Cali U, Elma O, Guler O. An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool. In: Proceedings of the IEEE power Amp energy innovative smart grid technologies conference ISGT; 2021. p. 1–5. doi:10.1109/ISGT49243.2021.9372263.

[121] Srinivasan S, Arjunan P, Jin B, Sangiovanni-Vincentelli AL, Sultan Z, Poolla K. Explainable AI for chiller fault-detection systems: gaining human trust. Computer (Long Beach Calif) 2021;54:60–8. doi:10.1109/MC.2021.3071551.

[122] Ugwuanyi C. Using interpretable machine learning for indoor CO_2 level prediction and occupancy estimation. University of Strathclyde; 2021. PhD Thesis.

[123] Wang M, Wang Z, Geng Y, Lin B. Interpreting the neural network model for HVAC system energy data mining. Build Environ 2022;209:108449. doi:10.1016/j.buildenv.2021.108449.

[124] Wastensteiner J. Explainable AI for tailored electricity consumption feedback – an experimental evaluation of visualizations 2021:19.

[125] Zdravković M, Ćirić I, Ignjatović M. Towards explainable AI-assisted operations in district heating systems. IFAC-Pap 2021;54:390–5. doi:10.1016/j.ifacol.2021.08.044.

[126] Zhang W, Liu F, Wen Y, Nee B. Toward explainable and interpretable building energy modelling. In: Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation. Association for Computing Machinery; 2021. p. 255–8. doi:10.1145/3486611.3491127.

[127] Zhang Y, Ma R, Liu J, Liu X, Petrosian O, Krinkin K. Comparison and explanation of forecasting algorithms for. Energy Time Ser Math 2021;9:2794. doi:10.3390/math9212794.

[128] Yu MG, Pavlak GS. Extracting interpretable building control rules from multi-objective model predictive control data sets. Energy 2022;240:122691. doi:10.1016/j.energy.2021.122691.

[129] Arjunan P, Poolla K, Miller C. BEEM: data-driven building energy benchmarking for Singapore. Energy Build 2022;260:111869. doi:10.1016/j.enbuild.2022.111869.

[130] Gao Y, Han H, Lu H, Jiang S, Zhang Y, Luo M. Knowledge mining for chiller faults based on explanation of data-driven diagnosis. Appl Therm Eng 2022;205:118032. doi:10.1016/j.applthermaleng.2021.118032.

[131] Geyer P., Singh M.M., Chen X. Explainable AI for engineering design: a unified approach of systems engineering and component-based deep learning. ArXiv210813836 Cs Eess 2022.

[132] Grzeszczyk TA, Grzeszczyk MK. Justifying short-term load forecasts obtained with the use of neural models. Energies 2022;15:1852. doi:10.3390/en15051852.
[133] Jin X, Xiao F, Zhang C, Li A. GEIN: an interpretable benchmarking framework towards all building types based on machine learning. Energy Build 2022;260:111909. doi:10.1016/j.enbuild.2022.111909.
[134] Kim J-Y, Cho S-B. Predicting residential energy consumption by explainable deep learning with long-term and short-term latent variables. Cybern Syst 2022;0:1–16. doi:10.1080/01969722.2022.2030003.
[135] Li M, Wang Y. Power load forecasting and interpretable models based on GS_XGBoost and SHAP. J Phys Conf Ser 2022;2195:012028. doi:10.1088/1742-6596/2195/1/012028.
[136] Moon J, Park S, Rho S, Hwang E. Interpretable short-term electrical load forecasting scheme using cubist. Comput Intell Neurosci 2022;2022:1–20. doi:10.1155/2022/6892995.
[137] Moon J, Park S, Rho S, Hwang E. Robust building energy consumption forecasting using an online learning approach with R ranger. J Build Eng 2022;47:103851. doi:10.1016/j.jobe.2021.103851.
[138] Mouakher A, Inoubli W, Ounoughi C, Ko A. Expect: eXplainable prediction model for energy consumption. Mathematics 2022;10:248. doi:10.3390/math10020248.
[139] Rajapaksha D., Bergmeir C. LIMREF: local interpretable model agnostic rule-based explanations for forecasting, with an application to electricity smart meter data. ArXiv Prepr ArXiv220207766 2022.
[140] Yang Y, Yuan Y, Han Z, Liu G. Interpretability analysis for thermal sensation machine learning models: an exploration based on the SHAP approach. Indoor Air 2022;32:e12984. doi:10.1111/ina.12984.
[141] Zdravković M, Ćirić I, Ignjatović M. Explainable heat demand forecasting for the novel control strategies of district heating systems. Annu Rev Control 2022. doi:10.1016/j.arcontrol.2022.03.009.
[142] Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. ArXiv160204938 Cs Stat 2016.
[143] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30 Curran Associates, Inc.
[144] Mishra P. Practical explainable ai using python: artificial intelligence model explanations using python-based libraries, extensions, and frameworks. Berkeley, CA: Apress; 2022. doi:101007/978-1-4842-7158-2.
[145] Wang Z, Hong T. Reinforcement learning for building controls: the opportunities and challenges. Appl Energy 2020;269:115036. doi:10.1016/j.apenergy.2020.115036.
[146] Yu L, Qin S, Zhang M, Shen C, Jiang T, Guan X. A review of deep reinforcement learning for smart building energy management. IEEE Internet Things J 2021;8:12046–63. doi:10.1109/JIOT.2021.3078462.
[147] An Y, Xia T, You R, Lai D, Liu J, Chen C. A reinforcement learning approach for control of window behavior to reduce indoor PM2.5 concentrations in naturally ventilated buildings. Build Environ 2021;200:107978. doi:10.1016/j.buildenv.2021.107978.
[148] Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2020;20:124. doi:10.1186/s12911-020-1120-5.
[149] Xu D., Fekri F. Interpretable model-based hierarchical reinforcement learning using inductive logic programming 2021. 10.48550/arXiv.2106.11417.
[150] Miller C, Kathirgamanathan A, Picchetti B, Arjunan P, Park JY, Nagy Z, et al. The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition. Sci Data 2020;7:368. doi:10.1038/s41597-020-00712-x.
[151] Li H, Wang Z, Hong T. A synthetic building operation dataset. Sci Data 2021;8:213. doi:10.1038/s41597-021-00989-6.
[152] Granderson J, Lin G, Harding A, Im P, Chen Y. Building fault detection data to aid diagnostic algorithm creation and performance testing. Sci Data 2020;7:65. doi:10.1038/s41597-020-0398-6.
[153] Sun J, Im P, Bae Y, Munk J, Kuruganti T, Fricke B. Dataset of low global warming potential refrigerant refrigeration system for fault detection and diagnostics. Sci Data 2021;8:144. doi:10.1038/s41597-021-00927-6.
[154] Commercial Buildings Energy Consumption Survey. U.S. energy information administration.; 2012.