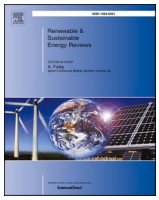
Renewable and Sustainable Energy Reviews 193 (2024) 114283

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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



A protocol for developing and evaluating neural network-based surrogate models and its application to building energy prediction

D. Hou [a,b,*], R. Evins [a,b]

[a] Energy in Cities Group, Department of Civil Engineering, University of Victoria, V8P 5C2, Canada
[b] Institute for Integrated Energy Systems, University of Victoria, V8P 5C2, Canada

ARTICLE INFO

Keywords:
Protocol
Surrogate model
Meta-model
Neural networks
Building energy
Synthetic data

ABSTRACT

Because of their low computational costs, surrogate models (SMs), also known as meta-models, have attracted attention as simplified approximations of detailed simulations. Besides conventional statistical approaches, machine-learning techniques, such as neural networks (NNs), have been used to develop surrogate models. However, surrogate models based on NNs are currently not developed in a consistent manner. The development process of the models is not adequately described in most studies. There may be some doubt regarding the abilities of such models due to a lack of documented validation. In order to address these issues, this paper presents a protocol for the systematic development of NN-based surrogate models and how the procedure should be reported and justified. The protocol covers the model development procedure sample generation, data processing, SM training and validation, how to report the implementation, and how to justify the modeling choices.

The protocol is used to critically review the quality of NN-based SMs in the prediction of building energy consumption. Sixty-eight papers are reviewed, and details of the developed surrogate models are summarized. The reported developing procedures were evaluated using the criteria proposed in the protocol. The results show that the selection of the number of neurons is the best-implemented step with a justification, followed by the determination of model architecture, mostly justified in a discussion way. While greater focus should be given to sample dataset generation, especially input variables selection, considering independence check and clear report of model validation on training and test data. Also, data preprocessing is strongly recommended.

1. Introduction

Computational simulations have proven to be extremely useful when studying complex physical phenomena. The process of conducting a high-fidelity simulation, however, is typically labor-intensive, time-consuming, and computationally expensive. In particular, the required computational time and cost are prohibitive when batch simulation is required, especially in large quantities. A surrogate model (SM) is an approximation of an original model that is more expensive to evaluate; the SM maps input data to outputs in a data-driven manner. Generally, SMs are used when the simulations are computationally expensive or when the relationships between input variables and output variables are not well understood.

In the field of building energy research, SMs have been used for conceptual design, system control and operation, building retrofit analysis, sensitivity analysis, uncertainty analysis, and design optimization [1]. Bracht et al. [2] developed an SM for predicting the annual thermal load of each room considered in a building information model. Building construction information extracted from the building information model and default values for missing information are fed into the SM. The developed SM and the corresponding integration tool are exploited to facilitate the exchange of data, especially in the early stages of design, to assist designers in selecting the most appropriate parameters. Zhu et al. [3] proposed a hybrid SM method for predicting the energy consumption of complex building forms by decomposing them. With a particular input of solar radiation, the SM can assist designers in estimating the differences in energy demand caused by inter-building shading and reflections in an urban environment, as well as the effects of self-shading. Luo et al. [3] embedded an SM into a Bayesian framework for estimating the flexibility of building operations under demand response considering the indoor thermal environment. In Ref. [4], the authors proposed a novel surrogate retrofit model with easily accessible inputs for predicting near-optimal retrofit solutions, which is non-experts used-friendly. In contrast to the conventional retrofit process, the proposed surrogate model exhibits an average accuracy of 0.9

* Corresponding author. Energy in Cities Group, Department of Civil Engineering, University of Victoria, V8P 5C2, Canada.
E-mail address: danlinhou@uvic.ca (D. Hou).

https://doi.org/10.1016/j.rser.2024.114283
Received 27 May 2023; Received in revised form 2 January 2024; Accepted 3 January 2024
Available online 9 January 2024

Nomenclature

Abbreviation

ACO	Ant Colony Optimization
Adam	Adaptive Moment estimation
BBO	Biogeography Based Optimization
BFGS	Broyden Fletcher Goldfarb Shanno
BR	Bayesian Regularization
CD	Cross Validation
D	Discrete Value
ELM	Extreme Learning Machine
FF	Feedforward
GD	Gradient Descent
GA	Genetic Algorithm
GRNN	General Regression Neural Network
GS	Grid Search
LHS	Latin Hypercube Sampling
LM	Levenberg-Marquardt
LSTM	Long Short-Term Memory
L&W	Level Samples and Weight Samples
MAPE	Mean Absolute Percentage Error
MLP	Multi-Layer Perceptron
N	Normalization
NN	Neural Network
PSO	Particle Swarm Optimization
R2	Coefficient of Determination
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
S	Standardization
SCG	Scaled Conjugate Gradient
SM	Surrogate Model
T&E	Trial and Error

Symbol

tanh	Hyperbolic tangent

for both continuous variables and categorical variables. Zhu et al. [5] identified key variables affecting cooling and heating loads with high reliability, taking into account the uncertainty of model inputs by using different SMs. Bre et al. [6] dynamically coupled an SM trained based on EnergyPlus samples with the multi-objective Non-dominated Sorting Genetic Algorithm-II to optimize building performance design considering energy efficiency and thermal comfort. The results of a real-world housing case demonstrate that the proposed method can significantly reduce the number of building energy simulations required to identify the Pareto front of a multi-objective building performance optimization problem while maintaining good accuracy.

Both grey-box and black-box models can be used as SMs. A grey-box SM can be developed by simplifying complex physical relationships between input variables and output variables under certain assumptions, while black-box-based SMs rely entirely on statistical techniques. The behavior of a particular system can be described using a function derived from a suitable database, which eliminates the need for a physical explanation completely. In addition to creating SMs with a single type of model (i.e., pure grey-box or pure black-box), researchers may also integrate them together to develop a holistic model. For example, Singhavel et al. developed a surrogate for predicting building heating and cooling loads called a component-based machine-learning model [7]. In their study, a single building thermal zone is divided into multiple physical entities, such as walls, windows, floors, etc. A black-box model is developed for each physical entity to simulate the energy fluxes through it. Then, all entities are assembled based on domain knowledge to build a grey-box model for a building with single/multiple thermal zones to make a final prediction.

With the development of artificial intelligence and the expansion of computers' computing capabilities, machine-learning approaches are gaining more attention and have often been applied to the development of SMs. In comparison to SMs based on conventional regression techniques, machine-learning-based SMs perform better in capturing complex processes. Neural networks (NNs) are a popular machine-learning method well-suited to regression problems which is inspired by biological neural networks. An NN is composed of a collection of connected units or nodes called neurons. Similar to synapses in a biological brain, each connection can transmit a signal from the receiving neuron to neighboring neurons. In each connection, the output is calculated by some nonlinear function of the sum of its inputs. During a typical learning process, weights are adjusted in order to increase or decrease the strength of the signal. It is possible for neurons to have a threshold beyond which signals are only sent if the aggregate signal exceeds it. Neurons are usually organized into layers. Different layers may perform different transformations on their inputs. From the first layer (the input layer), signals are likely to travel multiple layers before reaching the last layer (the output layer). More details of NNs can be found in Ref. [8].

We collected references about the SMs developed by NN for building energy-related topics and listed them in Table 1. Some researchers developed NN-based SMs to represent the original building energy/performance models to simplify fundamental building physics [7,9,10]. Some NN-based SMs are developed for special purposes by considering particular model features. Lopes and Lamberts [11] developed an SM for predicting the cooling energy consumption of office buildings. By including a new model feature in their NN models, proposed as Cooling Enthalpy Hours, which is more reasonable for hot and humid regions rather than Cooling Degree Days and Cooling Degree Hours, their model can be applied to different climates with good accuracy. In the study conducted by Jia et al. [12], besides general model inputs for building-level prediction, parameters of orientation, exterior wall area, window area, roof area, and apartment height above ground are included in the model features to downscaling the prediction of monthly cooling load from general building-level to apartment-level. Some researchers developed models for optimization issues. Chegari et al. [13] developed two NNs with identical model features to target annual energy thermal demand and annual weighted average of discomfort degree-hours separately. Then, these models integrate with commonly used metaheuristic algorithms to identify the optimal building envelope design considering energy efficiency and thermal comfort. Some NNs are applied to energy consumption prediction on a larger scale. In Ref. [12], the authors developed an SM to estimate the monthly cooling load at the building level with information on outdoor weather conditions, building construction information and envelope characteristics, and HVAC system operations. Then, the cooling load at a district level can be estimated by summating the cooling load of each residential building located in the district calculated by the developed SM. Nagpal et al. [14] employed the developed NN-based SMs with an optimization algorithm to find the highest-ranking combination of unknown parameters of a target building during an auto-calibration procedure to vastly reduce the computational expense. The proposed methodology makes it possible for campus operators to reasonably estimate energy impact, which is 500 times faster than traditional approaches.

There existed some previous review studies focusing on building energy forecasting with machine-learning approaches [15,16]; some of them even specified NN techniques [17–19]. Some authors captured the prosperous trend of developing SMs for building energy studies and reviewed the state-of-the-art [1]. However, none of them emphasized the particular scope of NN-based SMs in building energy-related studies.

Table 1
Details of data generation and processing for the development of neural network-based surrogate models.

Ref.	Simulation Tool					Sample Generation							Dataset Processing					
	EnergyPlus	Ecotect	DesignBuilder	TRNSYS	Other	Range	Distribution	Sample Method	Sample Size	Significance			Independence	Preprocessing				Training/Testing Splitting	Scaling
										Sensitivity	Correlation	Ad-Hoc		Encoding	Randomization	Removing	Other		
[11]	√					√	√	LHS	250,000	√			×	√		√		10-fold- CD	×
[26]	√					×	×	×	268		√		×	√	√		√	70-15-15	N
[27]		√				A dataset is used			768	√			×		√			55-15-30	N
[28]			√			√	D	×	463			√	×			√		70-0-30	N
[30]			√			Time series			58,237			√	×	×				×	×
[31]					√	Hourly data			×			√	×		√			90-0-10	N
[32]	√					√	√	LHS	500	√			×	×				90-0-10	×
[33]				√		Time series			742			√	×	√				66-0-33	×
[9]	√					√	D	×	714			√	×	√				80-0-20	×
[34]			√			√	×	×	240			√	×	×				70-0-30	×
[35]				√		Time series			8760			√	×	×				Two datasets	N
[36]				√		Time series			8760			√	×	×				Two datasets	N
[37]		√				A dataset is used			768		√		×		√			5-fold- CD	N
[38]				√		×	×	×	8000			√	×	×				67-15-15	×
[39]	√					√	×	Box-Behnken	12,960			√	×	×				70-15-15	×
[40]					√	A dataset is used			40,000	√			×		√			70-15-15	×
[41]					√	√	D	×	77,000		√		×		√			50-25-25 10-fold CD	N
[42]					√	√	×	×	×			√	×	×				×	×
[43]					√	√	×	×	620,000			√	×	×				70-0-30 Extra data for validation	×
[44]		√				A dataset is used			768	×			×	×				70-0-30	×
[45]					√	√	√	×	148			√	×	×				80-0-20	S
[10]	√					√	D	×	180			√	×	×				70-0-30	×
[46]	√					√	D	×	5625			√	×	×				Two datasets	×
[47]	√					√	×	×	×			√	×	×				×	×
[48]	√					×	×	×	6435	Stepwise regression			×			√		10-fold CD	S
[49]	√					√	×	×	200		√		×	×				×	N
[50]		√				×	D	×	586			√	×	×				10-fold CD	N
[51]	√					A dataset is used			12,000	×			×		√			Two datasets	N
[52]	√					Time series			8760	×			×	√				Two datasets	×
[12]	√					√	√	LHS	11,700	√			√		√			80-0-20 Extra dataset for validation	N
[53]					√	√	×	×	2611			√	×	×				50-25-25	N
[54]	√					Time series			3761		√		×	×				60-0-40	×
[55]			√			√	×	×	54		√		×	×				80-0-20	×
[16]		√				A dataset is used			768	×			×	×				90-0-10	×
[56]	√					×	×	×	5240	×			×			√		70-15-15	×
[57]					√	×	×	×	1920	×			×	×				87.5-0-22.5	N
[58]					√	×	×	×	×		√		×	×				×	×

(*continued on next page*)

Table 1 (*continued*)

Ref.	Simulation Tool					Sample Generation							Dataset Processing					
	EnergyPlus	Ecotect	DesignBuilder	TRNSYS	Other	Range	Distribution	Sample Method	Sample Size	Significance			Independence	Preprocessing				Training/Testing Splitting	Scaling
										Sensitivity	Correlation	Ad-Hoc		Encoding	Randomization	Removing	Other		
[59]					√	Time series			168			√	×	×				60-20-20	×
[60]	√					√	×	×	1008			√	×	×				×	N
[61]					√	√	D	×	2608			√	×	×				70-15-15	×
[62]					√	×	×	×	10,080		√		×	×				70-15-15	×
[63]				√		A dataset is used			2184	×			×	×				85-0-15	N
[64]					√	√	×	×	328			√	×	×				56-24-20	√
[65]			√			√	×	LHS	450	×			×	×				90-0-10	×
[66]					√	√	×	LHS	9000	×			×	×				80-0-20 10-fold CD	N
[67]	√					Time series			2016	×			×	×				52-0-48	×
[68]	√					Time series			744			√	×	×				×	×
[69]	√					√	×	×	1602	√			×	×				70-15-15	×
[70]	√					Time series			2,041,344			√	×	×				Two datasets	×
[71]				√		Time series			8760*3			√	×				√	33-33-33	×
[72]					√	√	D	×	442			√	×		√			75-0-25 10-fold CD	√
[73]				√		×	D	×	1560		√		×		√			85-0-15	×
[74]					√	×	D	×	243			√	×		√			90-0-10 10-fold CD	√
[75]	√					√	D	×	172,980			√	×	×				87-0-13	×
			√			√	D	×	243			√	×	×				Two datasets	×
[76]	√					Time series			3288	√			×	√				70-0-30	×
[77]					√	√	×	×	~5000			√	×		√			80-0-20	N
[78]					√	√	D	×	91			√	×		√			70-0-30	N
[79]		√				A dataset is used			768	×			×	×				70-0-30	×
[80]	√					A dataset is used			5000	×			×	×				80-0-10	×
		√				A dataset is used			768	√			×	×				75-0-25	×
[81]	√					√	√	LHS	300			√	×	×				80-0-20	N
[24]	√					√	×	LHS	52,560			√	×				√	64-16-20	S
[82]					√	A dataset is used			10,368	×			×	Performed				70-20-10	N
[83]				√		√	D	×	35			√	×		√			70-15-15	×
[84]	√					√	×	LHS	6400			√	×	×				Two datasets	N
[85]	√					√	×	LHS	15,000			√	×	×				10-fold CD	N
[86]	√					Time series			35,040		√		×	×				75-0-25 10-fold CD	×
[14]	√					×	×	LHS	270			√	×	×				×	×
[13]				√		√	×	L&W	35			√		×				70-15-15	×

In fact, though authors reported how they developed the models in the publications, there is no consistent procedure available that stipulates how to determine NN details and how to report them in a consistent manner. Also, the underlying dataset used for training is vitally important in determining the performance of the SM. The specification of the training dataset should also be stipulated by the procedure. Finally, for surrogate models, it is important to critically examine and evaluate the performance of the model in comparison to the original detailed model. Many papers lack adequate details of this process, making it impossible to draw any definite conclusions about how well the surrogate models performed. For studies that adopted a rigorous development process and gave adequate details of the process, there is a much greater degree of confidence in the results presented.

Therefore, in this paper, we propose a protocol for developing NN-based SMs and reporting the details of the model and testing process. Compared to previously developed protocols for general NN models [20,21], we extended the steps by adding the procedure of dataset generation. We also improved several sub-steps to make the NN models perform better. We then demonstrate this protocol by applying it to 68 papers that provide a comprehensive review of the literature relating to surrogate modeling of buildings. The paper is organized as follows. Section 2 describes the protocol step by step, including the generation of sample datasets, the determination of the NN architecture and hyperparameters, and the evaluation of the surrogate model. The criteria for evaluating performance and assessing the confidence level of the results are also outlined in this section. Section 3 gives a summary of the technical papers collected, including the simulation tools, the sample generation, and detailed SM information. Moreover, the quality of each SM is assessed using the proposed protocol, and the quality of all studies is summarized. Lastly, we provide conclusions and directions for further extending this work.

2. Protocol for developing and reporting surrogate models

SMs have been applied in a wide variety of applications [18,22], but this paper provides the first detailed protocol for developing such models in the area of building energy use. Forrester et al. [23] provide a very broad overview of how to develop an SM. Generally, there are three stages of the modeling processes: (1) preparing the data and choosing a modeling approach, (2) parameter estimation and training, and (3) model testing. Refs. [20,21] provide a protocol for developing general NN models, covering the model development process, method, and level of detail for each step and its justification. On that basis, we proposed a protocol that can be used in developing SMs for any field of modeling; in the following section, we take a building energy model as an example to illustrate the protocol step by step.

The protocol is composed of three parts, shown in Fig. 1, including a systematic surrogate model development procedure, criteria for assessing the reporting of model details, and criteria for assessing the justification of modeling choices. The first part, shown on the left in Fig. 1, is used as a guideline for how to develop an NN-based SM. Each step is identified with several sub-steps to outline the general developing procedures. The second and third parts of the protocol, shown on the right in Fig. 1, focus on ensuring sufficient detail in reporting the implementation of the model and the associated justifications. The categories for reporting and justification are shown in Fig. 2; the specific criteria for each category for every sub-step are given in Appendix A in Table A.1 and A.2. In order to make our contributions clear, we have listed the adaptations and extensions we made to the previously existing protocol:

a. Four steps were proposed in the systematic development procedure by considering some steps in the previous protocol as sub-steps. As a result of this revision, the purpose of each step will be more clearly defined.
b. We expanded the steps of Input Selection and Data Collection in Ref. [20] to Sample Generation. This step is an important element of surrogate model development since the scope of the dataset can directly limit the SM capability. Specifically, we added sub-steps such as input samples generation, output variables selection, and corresponding outputs generation.
c. We extend the step of Data Splitting in Ref. [20] to Data Processing by supplementing data pre-processing and data scaling, which can impact model performance.
d. In the third step of NN-Based Surrogate Model Training, we define it with four sub-steps, namely architecture selection (same in Ref. [20]), hyperparameters determination, NN model creation, and hyperparameter optimization.
e. In the last step of the development procedure, we removed the structural validity included in Ref. [20] and kept replicative validity and productive validity, considering that the purpose of SM validation is to evaluate its accuracy and effectiveness to represent the original detailed simulation model.
f. We expanded on Ref. [20] to include a more detailed assessment of the reporting and justification of each sub-step.
g. For the assessing of justification of modeling choice, we define three categories instead of four categories in Ref. [20]. Specifically, we removed the category of "Not Necessary" since all NN models have several model hyperparameters to be tuned.

2.1. Systematic surrogate model development procedure

As one of the main purposes of the study, we propose a systematic development procedure to make it clear how to develop an NN-based SM. Here, a building energy model is taken as an example of the original detailed simulation model. For other types of detailed simulation models, the development procedure should be the same. Similarly, for SMs based on machine learning concepts other than NNs, the procedure should be applicable with the appropriate changes to the NN block. The main differences between a general NN and a surrogate NN are the data source and model application. For a general NN, "real" data typically from measurements are used to train and test the model, while for an NN-based SM, the "synthetic" data used for training and testing are generated from the detailed simulator. The determination of inputs and outputs should consider the number of simulations and the application of the surrogate. Four main steps are included in the procedure, which are sample generation, data processing, NN-based SM training, and SM validation. Below is a detailed explanation of each step.

2.1.1. Sample generation

As a starting point, we assume that an appropriate detailed simulation model is available and that it can be parameterized and executed automatically. Once the reference simulation model is ready, developers have to determine the inputs and outputs of the SM according to its intended application. Target output variable(s) should be determined first since they likely impact the input variable choice. Generally, an SM with multiple output variables is more complicated than one with a single output. Training for one output can yield better accuracy [24], though many multi-output models exist. Overall, the selection of output variables should carefully consider the purpose of the SM; what metrics will be required by the end user?

Once the output variables are determined, the input variables can be determined according to their impact on the outputs. Considering too many input variables will increase the computational time and cost from more parametric simulations and more complex SM development. Too few input variables will result in limited capabilities of the developed SM, especially if significant variables are excluded. However, considering irrelevant or insignificant inputs may impact the SM accuracy while increasing the training time and making the network structure more complex [25]. Including redundant input variables can lead to similar results. Therefore, in this step, a significance check and independence check should be conducted to filter the appropriate input

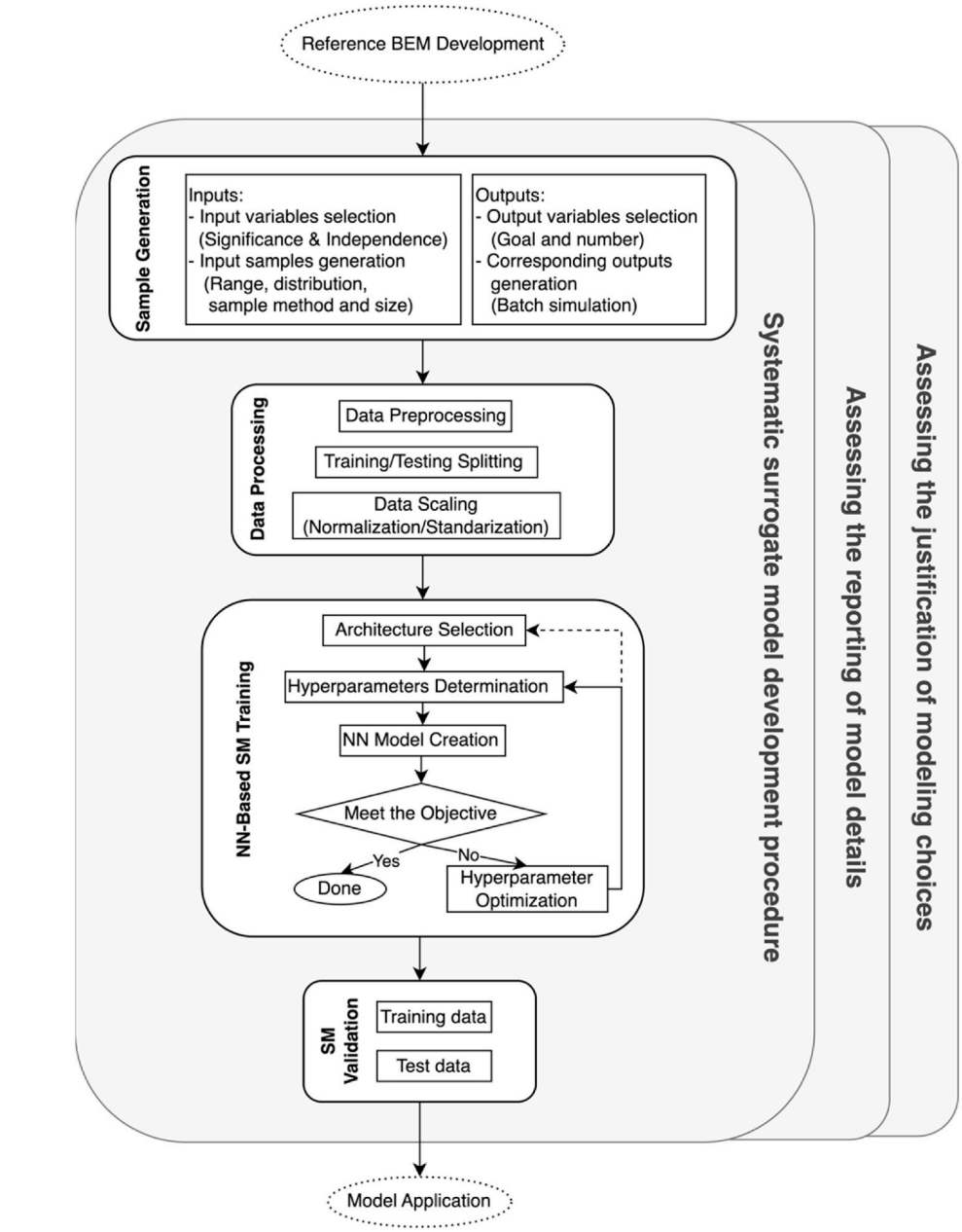


Fig. 1. The proposed protocol for developing and evaluating surrogate models using Neural-Network.

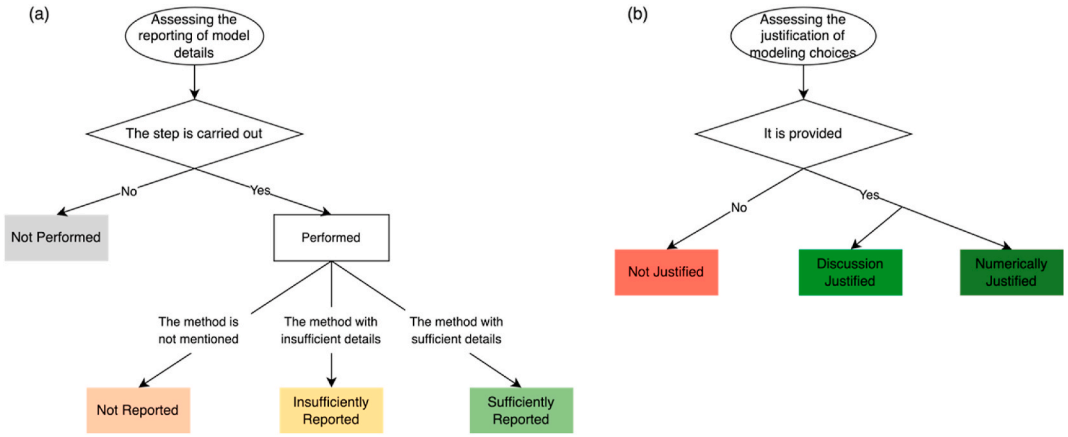


Fig. 2. Categories of the level of assessing the reporting of model details.

variables. Again, the selection of input variables should consider the purpose of the surrogate model; it is easy to specify models that are too simple to be meaningful or too complex to be useable (or trainable with a finite sample set).

Once the input variables and output variables are determined, it is important to pay careful attention to all other values in the model that will remain fixed. These set the "context" for the surrogate model, as the outputs are only valid while these remain unchanged.

Finally, developers must generate input samples and corresponding output samples. The sampling process should consider the ranges and distributions of input variables, sampling method, and sample size. Ranges and distributions can impact the SM feasibility and utility. Again, the balance must be found between being too narrow/coarse-grained and too wide/fine-grained. The sampling method is also important since it can impact the sample size required and the representativeness of the sample set. The number of samples required should be selected according to the number of selected input variables and sample method. If the sample set is too small, it will not cover the input domain adequately. However, a large sample set means long simulation times and potentially long NN model fitting times. The sample set will impact the NN model accuracy. There must be a balance between the sample set size, SM accuracy, robustness, simulation, and computing time. Unfortunately, there are no easy rules of thumb for determining sample size, as it greatly depends on the complexity of the underlying problem space. It may be necessary to fit a preliminary model and assess its accuracy, then return to generate more samples if required. Once the input sample set is determined, a parametric simulation should be conducted by using the input samples as parameter values in the detailed simulation model to generate corresponding output samples.

2.1.2. Data processing

Once the parametric simulation is completed, the input samples and corresponding outputs should be extracted from the simulation results to form the sample dataset. There are three sub-steps required to prepare this dataset for model fitting: data preprocessing, dataset splitting, and data scaling.

Firstly, the integrity of the dataset should be checked. This includes the exclusion of erroneous data; data points with severe errors or which are not representative of the problem space should be removed. Next, the data may need to be encoded. Some input variables are categorical rather than taking continuous or integer values, such as "chilled water pump configuration" or the day of the week. NN modeling best practice is to encode these using "one hot encoding," in which each categorical choice becomes a binary input variable that takes a value of one if that category is selected and zero otherwise. For example, for days of the week, there will be seven input variables.

Sometimes, missing data points have to be generated during the preprocessing sub-step, for example, for observed local weather if the weather features are included in the model inputs [26]. It is noted that, for the data preprocessing, randomization should be applied if the SM is not using time-series data. Sadeghi et al. found that randomization can improve the performance of the SM indicated by RMSE by 44 % and 50 % for heating and cooling load prediction, respectively [27].

The next sub-step is to split the data for training and testing of the SM. In general, there are two methods. The first method is that the sample dataset generated in the last step is split into the training subset and testing subset according to a certain percentage. The training dataset can be further split into sub-datasets for training and validation according to a certain percentage. Validation is used to measure the generalization of the model and to discontinue training when generalization ceases to improve. Testing is used to assess the performance of the model during and following training. The main disadvantage of this approach is that the representativeness of the original sample, which is ideal for training, is no longer guaranteed in the testing dataset. The second approach is using the sample dataset for training while building an extra dataset for testing.

Finally, data should be scaled using normalization or standardization to remove the impact of magnitude and units, which can confuse the model fitting process. Normalization is often used when the range of the data varies widely, and the absolute values are not as important as their relative positions. Typically, the rescaling range of a variable is between 0 and 1. Sometimes, the range between −1 and 1 is employed. It is also useful when dealing with multiple variables that have different units of measurement. Standardization, on the other hand, transforms the values of a variable to have a mean of 0 and a standard deviation of 1. This scaling technique is useful when the data has a Gaussian distribution and when the absolute values of the data are important.

2.1.3. NN-based surrogate model training

Once the steps of sample generation and the data process have been completed, the prepared training data can be used for SM development. Model training typically begins with finding an appropriate machine-learning algorithm. Here, we take NN as an example. Typically, this process involves the determination of model architecture and hyperparameters by using hyperparameter optimization (HPO) to achieve the desired level of model performance. Sometimes, model architecture is classified as a model hyperparameter and is involved in optimization. In this procedure, a "k-fold" cross-validation method is often used. The sample dataset is divided into k folds, approximately the same size. One of the folds is used for validation, while the other folds are for training. Repeat the previous k times by using a different subset for validation. The disadvantage of the cross-validation method is that the SM must be trained k times. So, the parameter k should be considered carefully to avoid a large increment of the computational cost.

2.1.3.1. Architecture selection. The NN model architecture determines the overall structure and information flow of the model. NN models can generally be classified into two types: feedforward neural networks and recurrent neural networks. If the information is passed only in one direction from the input layer to the output layer and terminated in the output layer, it is a feedforward NN, the most common NN architecture. A typical feedforward NN is a multi-layer perceptron (MLP), which consists of an input layer, several hidden layers, and an output layer. If the information is not only passed forward from the input layer to the output layer but also backward from the output layer to the input/hidden layer through a feedback loop, it is a recurrent NN. The information feedback from the output layers can be used to update the weights or even structure of the model, which enables the model to capture the complexities of highly dynamic systems. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Feedforward NNs are most suitable for surrogate modeling, particularly if time-resolved outputs are not required.

2.1.3.2. Determination of model hyperparameters. The model hyperparameters include the number of hidden layers, the number of neurons in each layer, how they process signals, etc. Generally, there is one input layer and one output layer, and the number of neurons in these layers matches the number of input and output variables. The number of hidden layers can determine the NN model complexity and is often the first target for hyperparameter tuning [28]. Increasing the number of hidden layers can enhance the generalizability of the NN model. The number of neurons in a hidden layer is an essential value as it impacts the capability of dealing with complex data. However, too many neurons in hidden layers may result in overfitting/overtraining problems [28]. In this case, the generalization may be impaired because the model fits some "noise" in the dataset. Concurrently, another problem that also affects NN performance is underfitting, which occurs in shallow NNs with too few neurons in hidden layers. Underfitting can result in large errors in the NN [28]. There is no clear criterion for determining the number of hidden layers and the number of neurons in hidden layers. The

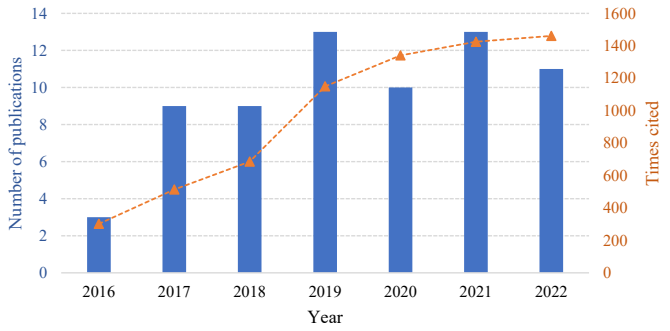


Fig. 3. Distribution of reviewed papers by year of publication.

transfer/activation function translates the input signals to output signals. Four types of transfer functions are commonly used: ReLU (Rectified Linear Unit), sigmoid, tanh (hyperbolic tangent), and softmax functions.

2.1.3.3. NN model creation. After determining the architecture and hyperparameters, the NN model is created to map the relationship between inputs and outputs under the model's configuration. The architecture defines the general framework of the model. The number of layers and the number of neurons in each layer regulate the complexity and capability of the model. The transfer function stipulates the signal transferred from inputs to outputs. The batch size defines the number of samples that will be propagated through the network. The learning rate controls how much to change the model in response to the estimated error each time the model weights are updated. Finally, the outputs of the model have to be compared with the results from the detailed computer model to assess its performance.

2.1.3.4. Hyperparameter optimization. After creating the NN model, the outputs of the model are compared to the target of the detailed computer model under a certain numerical index. The determination of the index should consider the application of the SM. For example, absolute values should be used when the SM is within the context of a net-zero energy/emission target. Relative errors can be employed to estimate building retrofit performance. If the value of the numerical index meets the required threshold, the training process is completed. The trained model is ready for validation. If the value of the numerical index cannot meet the requirement, the process of hyperparameter optimization should be performed. Hyperparameter optimization or hyperparameter tuning refers to the process of finding a set of model hyperparameters (e.g., learning rate, network depth, activation function, etc.) to enable the model to better capture the relationship between the inputs and outputs. HPO can have a great impact on model performance [29]; any study that reports performance values without performing this step is at risk of underestimating the potential accuracy if the model is tuned properly. Many methods can be used to obtain the optimal model hyper-parameters to improve performance, including global optimization methods, such as genetic algorithms; stepwise methods, such as pruning and constructive methods; Bayesian optimization methods; and ad-hoc methods, such as trial-and-error or selecting a structure based on previous experience. HPO must be conducted during model training since, in order to evaluate a set of parameters, the performance of the trained model must be known. This makes it a computationally expensive process, as many training cycles are required. It should be noted that the processing time of the optimization is an important selection criterion, which has a linear behavior with the sample size, i.e., the larger the sample size, the longer the optimization time. In Ref. [11], a synthesis index is used to consider the statistic performance index and computing time simultaneously.

2.1.4. Surrogate model validation

This step aims to evaluate the capability of the SM to capture the mapping relationship between inputs and outputs in the detailed computer models. Only SMs whose outputs are close to the outputs of detailed computer models within an acceptable error range with the same inputs can employ future analysis to replace the detailed computer model for the intended purpose. Therefore, the SMs should be validated through the accuracy assessment both on training data and test data. We call the former replicative validity and the latter alternative validity. If the trained SM can generate precise outputs under the input set involved in the training procedure, it is considered explicatively validated. If the trained SM's outputs are aligned to the dataset that is excluded from the training procedure, always called "unseen data," we can conclude that the SM is predictively validated, and its generalization ability is acceptable. To measure if an SM is validated on training data or test data, generally, standard statistical figures, including means and variances, analysis of variance, goodness-of-fit, regression and correlation analysis, and confidence interval construction, can be estimated as is done for traditional statistical models. The selection of the statistical index should consider the application of SMs; either absolute errors or relative errors are sensitive to the application goal. To estimate retrofit performance, generally, relative errors are employed. While for meeting the net-zero goal in buildings, absolute values should be selected.

2.2. Assessing the reporting of model details

Research can only progress in a given field if the knowledge developed in each contribution can be understood and passed on to fellow researchers in related fields. For this reason, not only must a rigorous model development process be followed, but the modeling procedures must be documented and reported in order to ensure that they can be replicated and built upon in the future. A common problem in surrogate modeling research is that insufficient model details are reported to allow reproducibility.

In this paper, we use four levels of reporting following Ref. [20], which are further developed according to the proposed development procedure in section 2.1. The different levels of reporting are explained in Fig. 2(a); these are applied separately to every sub-step of the model development procedure above, except the sub-step of NN model creation in NN-based SM training, which procedure is realized in the code and no mention in the reporting. If the step is not reported, we assume it is not applicable or not conducted and assign Level 0 (Not performed). Level 1 (Not reported) represents the case where a particular modeling step is carried out, but the details of the method are not mentioned. Level 2 (Insufficiently reported) is assigned if the method used is mentioned, but insufficient details have been provided, so the step cannot be replicated by other researchers. Level 3 (Sufficiently reported) is used where detailed information on the implementation is provided so that it can be repeated for similar studies by other researchers. Details of how exactly we defined the level of reporting for each step are shown in Appendix A in Table A1.

2.3. Assessing the justification of modeling choices

The credibility and confidence in research results relating to surrogate modeling depend on the level of justification provided for the use of a particular method at a particular step in the model development process. The greater the level of justification, the lower the uncertainty surrounding the selection of a particular method. This justification may be presented in two ways: discussion-based or numerical. A discussion-based justification is explanatory; the research reported is based upon findings from previous studies or argued based on known principles. Numerical justification undertakes direct comparative studies to show which choices perform better. The latter can drive methodological advances in the field by providing numerical evidence for future studies.

We use three levels of justification revised from Ref. [20] to match

our surrogate model development process. We assess the justification provided for the use of a particular method in each model development step, as shown in Fig. 2(b). Level 1 (Not justified) indicates that no justification is provided; Level 2 (Discussion justified) means that a discussion-based justification is provided based on findings from previous studies or logical arguments; Level 3 (Numerically justified) indicates that numerical justification is provided by comparing the method selected with alternatives.[1] Table A2 in Appendix A gives details of how the levels of justification were defined for every step.

3. Quality assessment of reviewed papers based on the proposed protocol

Finally, 68 papers were selected, including 57 journal papers and 11 conference papers. Fig. 3 shows the number of publications and citations per year. There is an increasing trend in the number of papers published since 2017, compared with only three papers in 2016. The method of collecting the reviewed papers is explained in Appendix B.

Details of each sub-step of the developed optimal NN-based SM for certain target output(s) are summarized and categorized according to the proposed protocol. Particularly, Table 1 summarizes that how the authors generated the samples, how the data were processed, and which simulation tools they used. Table 2 presents details about the models' training and validation, including model's architecture and hyperparameters, and HPO methods. We mark it using" $\surd$"and " $\times$ " to show if the information is presented and if the step is performed but the method is not mentioned. We specify the method used in each sub-step to give an overview of typical architectures and hyperparameters used in building energy NN-based SMs.

The studies cover residential and/or non-residential buildings. For time series prediction, the minimum time scale is 10 min. Most cases focus on monthly/annual prediction. EnergyPlus is the most popular simulation tool that is used to generate sample datasets for SM (41 %), likely due to the ease of parameterization and automated execution. LHS is the most popular sampling method. The dataset size varies hugely from 35 to over 2 million. For the significance check, Ad-Hoc is widely used, which means the authors select model input variables according to domain knowledge or model development experience. Unfortunately, a crucial sub-step, independence check, is seldom conducted. For dataset splitting for training and testing, a percentage of 70-15-15 (considering 15 % of data for validation in training) or 70-30 are widely selected. Normalization is used to scale the data more frequently than standardization. FF-MLP is the most prominent model architecture. In the reported optimal number of model layers, 3 to 6 is preferred. Layers of 3 or 4 are more general. The transfer function of sigmoid is adopted over half of the cases with a clear statement of transfer function information (51 %), followed by tanh function. Levenberg-Marquardt is popularly applied with other optimization methods to tune hyperparameters, such as PSO and BR. For the model validation, almost all cases selected numerical accuracy performance without considering the cost of computer resources or computational time.

Fig. 4 summarizes the overall assessment results of each sub-step for all reviewed cases. In addition, the quality of reporting and justification of each model shown in Tables 1 and 2 is shown in Appendix C, which is assessed according to levels of reporting and justification shown in the proposed protocol (Sections 2.2 and 2.3). In the following, we present the quality assessment result of each sub-step.

[1] It should be noted that Level 3 does not apply to the sub-steps of range and distribution in sample generation step, as in contrast to the other steps, where there is no numerical way to justify the range and distribution selection. Also, Level 3 does not apply to model validation step, where model predictive ability can be used to assess the relative performance of different methods (e.g., calibration methods, methods for selecting the number of hidden nodes etc.), there is no quantitative measurement for validation performance.

3.1. Sample generation

The range, distribution, sampling method, and sample set size used are critical aspects of the SM development procedure. However, this is rarely given enough attention in reporting. Over half of the reviewed papers do not report the ranges used for sampling, while only 10 % of the papers state the ranges with solid references. 80 % of papers do not present the selected range with evidence or references. For the studies that do not employ existing datasets, 58 % of the papers do not mention the distribution used for sampling, though distribution has an important role in data sampling and the determination of proper scaling techniques.

For the sampling method, 44 % of papers are Level 0 (Not performed), meaning either an existing dataset from a batch simulation is used or discrete values are selected arbitrarily and used as samples. Only 31 % of authors report the method used for sampling, with 23 % describing the method in detail or via references. 86 % of papers do not show the justification for the selected sampling method.

For the determination of sample set size, although the information is disclosed in more than 90 % of papers, only 6 % of the authors justified the chosen value via a numerical test. For most papers, we have no idea why the value was selected.

For the input variable selection, a significance check and independence check should be conducted to filter the best input variables to use as model features. Features with limited importance on the target outputs will increase the model complexity and computational cost of the SM with limited benefit or even a detrimental impact on model performance. Redundant features have a similar impact. Therefore, significance check and independence check are highly recommended. However, our results show that 90 % of studies did not provide enough details about significance checks. Most researchers select the important input variables according to their knowledge or experience, while only 12 % of researchers justify their choices numerically. Unfortunately, few studies conduct the independence check between the input variables. Only Ref. [12] considered it during the procedure of linear regression.

3.2. Data processing

For the three sub-steps in sample dataset processing, only 9 % of authors clearly show the data preprocessing procedure. In reality, this sub-step cannot be skipped since the simulation results should be examined to see if there are any failed running or unreasonable cases. For certain input variables, encoding is needed to convert the information to a useable value. In the studies in which authors generated their input-output datasets, over 50 % have no mention of data preprocessing. Only 9 % reported the procedure with enough detail. Only one out of 68 studies numerically justified the procedure used [10]. It is indicated that randomization and moving average can improve RMSE of heating load by 44.37 % and 16.34 %, respectively. The former can improve 50.07 % of cooling load prediction. In contrast, the latter has no improvement in cooling load prediction [10]. The sub-step of training/testing splitting is about the dataset preparation for model training, validation, and testing, which is an important step for model validation from a different perspective. If the developed model only works with the training dataset and its capability is not tested in unseen data, the generalization performance of the model cannot be guaranteed. Most researchers conduct the procedure using cross-validation. 68 % of studies built the different datasets for training and testing with a clear description. However, the results show that researchers select a conventional percentage for data separation instead of a numerical determination. The data obtained from the step of sample generation should be scaled to remove the impact of magnitude and units. Normalization is used in 78 % of studies that conduct the scaling step.

Table 2
Details of training and validation of neural network-based surrogate models.

Ref.	NN-based Surrogate Model Training								Surrogate Model Validation	
	Architecture		Hyperparameters					Hyperparameter Optimization	Training Data	Test Data
	FF/FF-MLP	Other	Number of Layers	Number of Neurons	Transfer Function					
					Sigmoid	Tanh	ReLU			
[11]	×		14-29-1		√			BFGS	Num Time	Num
[26]	√		14-25-4		×			LM	Num Plot	Num Plot
[27]	√		8-10-8-8-2			√		GS	Num	Num
[28]	√		10-5-6-4-2			√		LM-BR/GS	Num	Num
[30]	√		7-20-20-3		√			BR	Num	
[31]	√		10-5-2-1		√			BP	Num	Plot
[32]	√		29-x-1		√			LM T&E	Num	Num
[33]	√		9-x-1 12-x-1			√		LM	Num	
[9]	√		3-6-4		√			BR	Num	Num
[34]	√		1-3-1		√			LM/T&E	Num	Num
[35]	√		22-14-4		√			LM	Num Time	Num Time
		√	22-8-4		√			BP/Adam	Num Time	Num Time
[36]	√		6-20/19/9/11/10-1				√	LM	Num	Num
			6-12/13/1415/18/19-1				√	BR	Num	Num
[37]	√		4-4-2			√		BR/T&E	×	Num
[38]	√		3-6-1		×			×	Num	Num
[39]	×		11-25-24-15-1 11-13-19-13-1		√			LM GA	Num	Num
[40]	√		8-50-13		√			SCG	Num	
[41]	√		8-14-1 8-10/14/12-1 8-15-1		√			BFGS	Num	
[42]	√		4-18-20-4		×			LM	Num	×
[43]	√		12-13-1			√		LM-BR	Num	
[44]	√		8-5-2		×			BBO	Num	Num
[45]	√		4-7-5-1		√			LM	Num	Num
[10]	√		6-3/6/10-1		√			ELM	Num	Num
[46]	√		6-24-12-2				√	GD	Num	Num
[47]	√		7-13-13-13-13-1		×			×	Num	×
[48]	√		14-12-12-12-1			√		LM	Num	Num
[49]	√		19-8-1 16-9-1			√		PSO& ACO	×	Num
[50]	√		8-3-2		√			BP	Num	
[51]	√		16-10-16-1			√		LM/T&E	Num	Num
[52]	√		×		×			BP	Num	Num
[12]	√		14-13-1				√	RMSprop	Num	Num
[53]	√		6-13-1		Nonlinear			LM	Num	Num
[54]		√	8-12-1		×			BP	Num	Num
[55]	√		×		×			BP	Num	
[16]		√	×		√			×	Num	Num
[56]	√		6-14-15-1				√	LM	Num	Num
[57]	√		×			√		×	Num	
[58]		√	×		×			LM	Num	Num
[59]	√		8-10-1-1		√			LM	Num	Num
[60]	√		7-23-1		×			×	Num	
[61]	√		6-60-1-1-1		×			BR	Num	Num
[62]	√		3-20-1		√			BP	Num	
[63]	√		12-8-1 12-30-1 12-400-120-1 12-500-250-1			√		BP	Num	Num
[64]	√		9-10-1		√			LM/Graywolf	Num	Num
[65]	√		19-8-1		×			×	Num	
[66]	√		39-50-50-50-1 39-100-100-100-1 39-150-150-150-1				√	GD	Num	Num
[67]	√		6-x-2		√			×	Num	

(continued on next page)

Table 2 (*continued*)

Ref.	NN-based Surrogate Model Training								Surrogate Model Validation	
	Architecture		Hyperparameters					Hyperparameter Optimization	Training Data	Test Data
	FF/FF-MLP	Other	Number of Layers	Number of Neurons	Transfer Function					
					Sigmoid	Tanh	ReLU			
[68]		√	Scatter is 0.16		NA			Radial	Num, Time	
[69]	√		6-10-1		×			LM	Num	
[70]	√		12-100-100-100-1		×			Iteration	Num	Num
[71]		√	6-4-1 6-10-1 6-2-1 6-8-1		×			LM	Num	
[72]	×		×		×			×	×	Num
[73]	√		29-50-25-7			√		BP	Num	Num
[74]	√		14-6-1		×			×	×	Num
[75]	√		6-20-3			√		LM/T&E	Num	Num
	√		8-20-2			√		LM	Num	Num
[76]	×		8-x-1		√			LM	Num	Num
[77]	√		4-8-1			√		LM	Num	Num
		√	Number of cluster 9		NA			NA	Num	Num
[78]	×		6-5-3-4-2-1			√		GA	Num	Num
[79]	√		8-X-1		√			GD	Num	
[80]	√		×		×			×	Num	
	√		×		×			×	Num	
[81]	√		Two hidden layers		×			×	×	Num
[24]		√	×				√	×	Num	Num
[82]	√		8-x-x-x-x-2 X between 15 and 18				√	Adam	×	Num
[83]	√		11-4-1		√			LM/PSO	Num	
[84]	√		12-x-x-4		√			LM-BR	Num	Num
[85]	√		15-6-1 15-9-1		√			LM/ACO	Num	
[86]	√		39-39-1 69-15-1		√			BP	Num	
[14]	×		32-x-1		×			×	Num	Num
[13]	√		11-4-1		√			LM/PSO	Num	Num

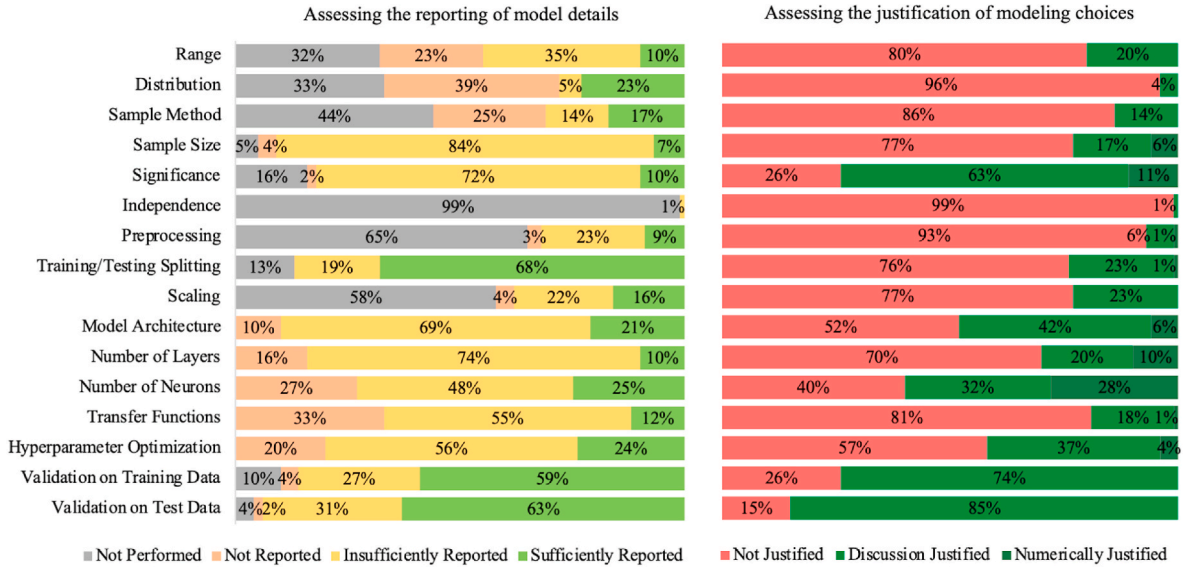


Fig. 4. Quality assessment of the reviewed papers using the proposed protocol for each sub-step.

3.3. NN-based surrogate model training

The choice of architecture is the foundation of an NN-based SM since it determines the nature of the NN models that are used to approximate the sampling space. From our review, it is concluded that FF-MLP is the most popular one employed in building energy prediction. However, in some studies, the authors do not report the architecture type. Many state that an NN model is employed with an illustration figure of an FF architecture. Only four models justify the selection of the architecture used in a numerical way. Once an architecture is selected, the number of layers and number of neurons in each layer should be determined subsequently since they are the two most important hyperparameters. These

values should be reported and discussed in detail. However, around half (44 %) of researchers selected a single hidden layer by default and without any exploration of multiple layers. To determine the number of neurons, generally, there are two approaches. Several researchers proposed equations to calculate the number of neurons in the hidden layers according to the number of model features (input variables), number of outputs, and/or data size. Another method is to find the number through optimization, such as global methods, grid search (GS), or trial and error (T&E). Regarding transfer functions, 88 % of studies do not report it with sufficient detail. Most selected functions (81 %) are without any justification, discussion, or supporting references; only one study justified the choice of function in a numerical way [51]. The authors proposed 12 combinations of different transfer functions in the hidden layer and output layer to investigate the effect. The MAPE and R2 calculated on test data range from 5.58 % to 23.45 % and 0.73 to 0.98, respectively. The transfer function of SigmoidAxon applied to both the hidden layer and output layer outperforms. 24 % of studies clearly described the selection of the learning algorithm or the optimization method used to improve model performance, with 4 % of studies justifying it numerically. Reynolds et al. [30] conclude that the most sensitive hyperparameters are the number of layers and the number of hidden neurons; little difference is found between Bayesian Regularization and Levenberg-Marquardt learning algorithms or Log-sigmoid and Tan-sigmoid transfer functions. Wang et al. [31] also confirm the lack of difference between the transfer functions of sigmoid and tanh. In addition, they conclude that nonlinear models perform much better in prediction than linear models since the R2 drops dramatically when there is no activation used. In Fonseca and Pereira's research [9], they stated that, although different activation functions impact the accuracy improvement slightly, the time taken for hyperparameter optimization using Bayesian Regularization and Levenberg-Marquardt is much lower (less than 10 min) compared with the gradient descent algorithm (1.0–2.5 h). Balancing performance and speed, Bayesian Regularization is determined to be the best choice. From the review, we found that in studies about SM-based performance optimization or control optimization for building system or building design, the procedure of hyperparameters are always absent, even key ones such as the number of hidden layer and number of neurons. It is noted that the concept of automated machine learning (AutoML) has emerged, which refers to automating the end-to-end process of applying machine learning to real-world problems. The purpose of AutoML is to make machine learning more accessible to non-experts, as well as to reduce the amount of time and resources required for the development and deployment of machine learning models. A recent survey of the current state of the art of AutoML has been published by He et al. [87].

3.4. Surrogate model validation

In some cases, the type of validation is not clearly pointed out since the authors do not mention which type of dataset is used in the accuracy calculation. Instead of summarizing the assessment results for all studies, we reported the assessment results for which cases the validation type is clear (49/67). It is found that 90 % of studies calculate the accuracy index both on training data and test data. It is noted that no Level 3 of Justification in the validation of training data and test data is shown in Fig. 4 since numerical justification is not applicable for this step. However, no authors sufficiently discussed the rationality of the selected performance index.

4. Conclusions and future work

In this paper, we propose a protocol for surrogate modeling development, especially using neural networks. The protocol is composed of three segments, including the systematic development procedure, the assessing of reporting of model details, and the assessing of justification of the implementation of each sub-step. The protocol can be applied to surrogate modeling in any research field. The authors take an example of building energy prediction to show the application. In total, 68 papers are collected, and the quality of the developed optimal SM is assessed according to the protocol. Details of the developed optimal NN-based SMs, such as simulation tools for parametric simulation, sample generation, type of model architecture, values of model hyperparameters, the HPO method, etc., are summarized.

It is shown that, in building energy prediction, the simulation tool of EnergyPlus is popularly employed in the parametric simulation to generate input-output samples. Information on the range and distribution of model parameters for propagation is not reported sufficiently. Authors pretend not to provide solid references for their selection. However, this information is essential since it determines the feasible region of the developed SMs. Significance check is performed well either in a discussion way or a numerical way. A few papers report the implementation of the independence check, which cannot be skipped since it is critical to model feature selection. LHS is applied in most cases. Data preprocessing is strongly recommended since it can guarantee data integrity and assist in the model's performance improvement. The splitting percentage of 70-15-15 or 70-30 is accepted through the common practice of training data and splitting data generation. Normalization is the most common scaling technique. For the NN model development, feedforward/feedforward multiple perception is widely used in research, even for time-series prediction. The report quality of hyperparameters and the justification for their determination can be higher by reporting more details and justification procedures to convince readers of findings and conclusions. In the SM validation part, the accuracy of the developed SM should be reported both on training data and test data to show the model's performance to seen and unseen data. Notably, the performance of SMs is critical whether in retrofit assessment or in SM-based optimization controls. Therefore, reporting sufficient details of the SM development and justifying the implementation is important. Only conclusions and findings obtained from replicable procedures are meaningful and trustable for readers.

This work gives insight to readers for a test at the beginning of their surrogate NNs development and how to report the development procedure and conduct the justification. In the future, the authors' group would like to focus on how to develop high-performance SM with affordable computing resources and computing time. Particularly, we will put the main effort into suggestions of NN model type under particular application purpose, the relationship between sample size, model features, and model complexity.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ralph Evins, Danlin Hou reports financial support was provided by Natural Sciences and Engineering Research Council of Canada.

Data availability

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

Acknowledgements

This work was supported by the ReBuild Initiative NSERC Alliance grant.

Appendix A

Table A.1
Proposed assessment categories for assessing the reporting of NN-based surrogate model details.

Steps		Category 0 (Not Performed)	Category 1 (Not reported)	Category 2 (Insufficiently Reported)	Category 3 (Sufficiently Reported)
Sample generation	Range	NA. A dataset is used, or time series data is used.	Not mentioned.	Information is shown in the paper, but insufficient details on how to determine it.	Information is shown with sufficient details on how to determine it.
	Distribution	NA. A dataset is used, or time series data is used.	Not mentioned.	Information is shown in the paper, but insufficient details on how to determine it.	Information is shown with sufficient details on how to determine it.
	Sample Method	NA. A dataset is used, or time series data is used.	Not mentioned.	The method is mentioned with insufficient details. So, it cannot repeat.	The method is mentioned with sufficient details to repeat it.
	Sample Size	NA. A dataset is used, or time series data is used.	Not mentioned.	The size is mentioned. But insufficient details about how to determine the size.	The size is mentioned. Sufficient details about how to determine the size.
	Significance check	Not mentioned.	Performed. But the method is not mentioned.	The method is mentioned; however, the procedure cannot be repeated due to a lack of information.	The method is described in detail (or relevant references are provided), so that it can be repeated for the same/similar studies.
	Independence check	Not mentioned.	Performed. But the method is not mentioned.	The method is mentioned; however, the procedure cannot be repeated due to a lack of information.	The method is described in detail (or relevant references are provided), so that it can be repeated for the same/similar studies.
Data processing	Preprocessing	Not mentioned.	Performed. But the method is not mentioned.	The method is mentioned; however, the procedure cannot be repeated due to a lack of information.	The method is described in detail (or relevant references are provided), so that it can be repeated for the same/similar studies.
	Training/Testing splitting	Not mentioned.	Performed but the method is not mentioned.	The method is mentioned; however, the procedure cannot be repeated due to a lack of information.	The method is described in detail (or relevant references are provided), so that it can be repeated for the same/similar studies.
	Data scaling	Not mentioned.	Performed but the scaling method is not mentioned.	Data scaling method mentioned but not enough details to repeat.	The data scaling method used is described in detail (or relevant references are provided), so that it can be recreated.
NN-based SM training	Architecture selection	NA.	The model architecture is not mentioned.	The model architecture used is mentioned; however, not enough information is provided so that the architecture can be recreated.	The model architecture used is described in detail (or relevant references are provided), so that it can be recreated.
	Determination of model hyperparameters (Number of layers, Number of neurons, Transfer/activation function, etc.)	NA.	The hyperparameter is not reported or just mentioned without any details about how to determine it.	The method of how to determine the hyperparameter is mentioned; however, the procedure cannot be repeated due to a lack of details about how to determine it.	The method of how to determine the hyperparamter is described in detail (or relevant references are provided), so that it can be recreated.
	Hyperparameter optimization	Not mentioned.	Performed but the method is not mentioned.	The method used for HPO is mentioned; however, the procedure cannot be repeated due to a lack of information.	The HPO method is described in detail (or relevant references are provided), so that it can be recreated.
Model validation	Training data & Test data	Not mentioned.	Performed but the method is not mentioned.	The model validation method(s) used is mentioned, however, not enough information (e.g., details/ references) is provided so that the procedure can be repeated.	The model validation method(s) is described in detail (or relevant references are provided), so that it can be repeated for the same/similar studies.

Table A.2
Proposed assessment categories for the justification of modeling choices in the NN-based surrogate model development process.

Steps		Category 1 (Not Justified)	Category 2 (Discussion Justified)	Category 3 (Numerically Justified)
Sample generation	Range	No justification.	The selection of a particular range is justified in a discussion way, such as with clear evidence or reference.	NA.
	Distribution	No justification.	The selection of a particular distribution is justified in a discussion way, such as with clear evidence or reference.	NA.
	Sample Method	No justification.	The use of a particular sample method is justified in a discussion way.	The use of a particular sample method is justified by comparing it to alternative sample method.
	Sample Size	No justification.	The determination of the sample size is justified in a discussion way.	The determination of the sample size is justified by comparing it to other sample size.
	Significance	No justification.	The use of a particular significance check method is justified in a discussion way.	The use of a particular significance check method is justified by comparing it to alternative significance check method.
	Independence	No justification.	The use of a particular independence check method is justified in a discussion way.	The use of a particular independence check method is justified by comparing it to alternative independence check method.
Data processing	Preprocessing	No justification.	The Justified in a discussion way.	Justified by comparing it to alternative model architectures.
	Training/Testing splitting	No justification.	The generation method is justified in a discussion way.	The generation method is justified by comparing it to alternative method.
	Data scaling	No justification.	The use of a particular scaling method is justified in a discussion way.	The selection of a particular scaling method is justified by comparing it to alternative scaling method.
NN-based SM training	Model Architecture	No justification.	The use of a particular architecture is justified in a discussion way.	The selection of a particular architecture is justified by comparing it to alternative model architectures.
	Determination of model hyperparameters (Number of layers, Number of neurons, Transfer/activation function, etc.)	No justification.	The optimal hyperparameter is justified in a discussion way.	The optimal hyperparameter is justified by comparing it to other numbers.
	Hyperparameter optimization	No justification.	The use of a particular optimization algorithm is justified in a discussion way.	The use of a particular optimization algorithm is justified by comparing it to alternative method.
Surrogate model validation	Training data & Test data	No justification.	The use of a particular model validation method is justified via discussion.	NA.

Appendix B

Methodology for review paper collection

To collect papers related to the topic, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology was applied [88]; the PRISMA flowchart is shown in Figure B1.

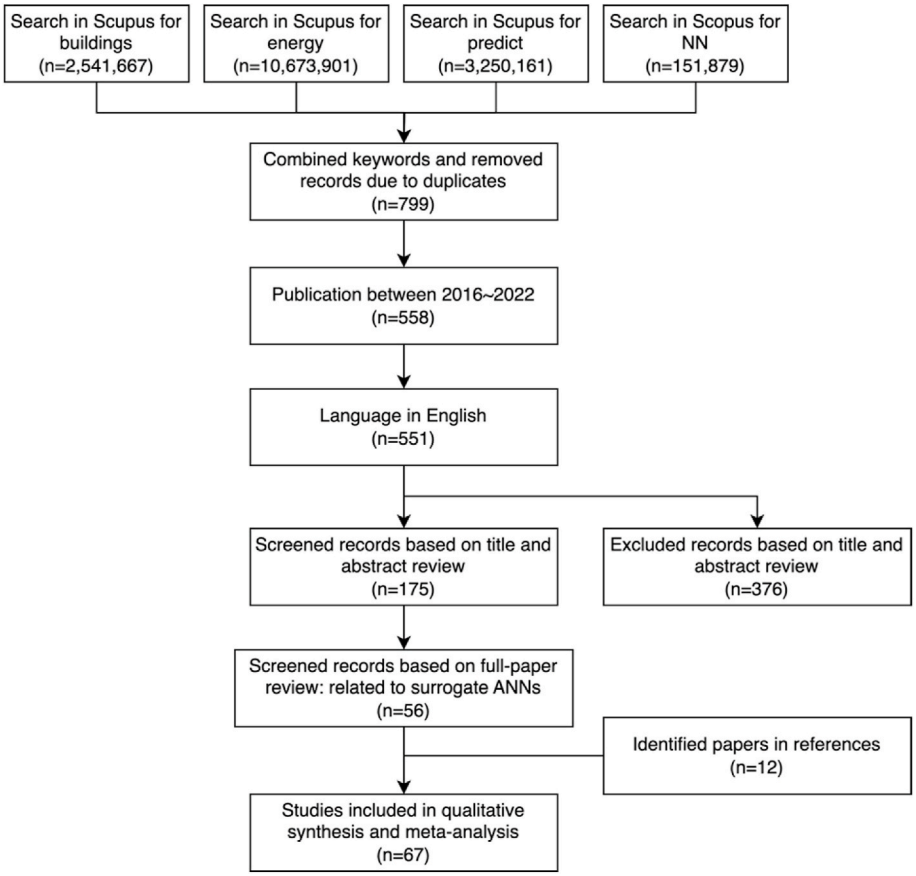


Fig. B.1. The PRISMA flowchart of this review.

Scopus is the main search engine of the methodology, where the full list of searching keywords is the full combination of each sub-keywords. Here, three kinds of sub-keywords are identified. The first sub-keyword narrows the paper to focus on building energy systems. The full list of the first sub-keywords is: "building", "HVAC", "in buildings", "dwelling", "household", "cooling", and "heating". The second sub-keyword defines the energy type to be predicted. The full list of the second sub-keywords is: "energy", "load", "electricity", "consumption", "demand", "gas", and "steam". The third sub-keyword defines the action of prediction; the full list of the third sub-keywords is: "forecast", "predict", and "estimate". The example search keywords are "building electricity predict", "cooling demand forecast", "electricity estimates in buildings", and so forth. The total number of search keywords in this paper is $7 * 7 * 3 = 147$ keywords. Then, "NN" was added to the search. After the combination of all four sub-keywords and removed duplicates, 799 papers were identified. When we range the publication year between 2016 and 2022, the number of related papers decreases to 558, including 551 papers in English. One hundred seventy-five papers were selected after screening records based on title and abstract. Since we cannot define a list of words to scope the studies focus on surrogate NN development, these kinds of sub-keywords were not included. Instead, a manual check was conducted to identify papers related to surrogate NNs, which brought the list down to 56 papers. By double-checking the references of the 56 selected papers, an extra 12 papers were added to the paper pool.

Appendix C

Table C.1
Summary of the assessment results of reporting and justification of the developed neural network-based surrogate models.

Note: "?" means it is not clear for which type of validation.

References

[1] Westermann P, Evins R. Surrogate modeling for sustainable building design – a review. Energy Build 2019;198:170–86. https://doi.org/10.1016/j.enbuild.2019.05.057.

[2] Bracht MK, Melo AP, Lamberts R. A meta-model for building information modeling-building energy modeling integration in early design stage. Autom ConStruct 2021;121:103422. https://doi.org/10.1016/j.autcon.2020.103422.

[3] Luo N, Langevin J, Chandra-Putra H, Lee SH. Quantifying the effect of multiple load flexibility strategies on commercial building electricity demand and services via surrogate modeling. Appl Energy 2022;309:118372. https://doi.org/10.1016/j.apenergy.2021.118372.

[4] Thrampoulidis E, Mavromatidis G, Lucchi A, Orehounig K. A machine learning-based surrogate model to approximate optimal building retrofit solutions. Appl Energy 2021;281:116024. https://doi.org/10.1016/j.apenergy.2020.116024.

[5] Zhu L, Zhang J, Gao Y, Tian W, Yan Z, Ye X, et al. Uncertainty and sensitivity analysis of cooling and heating loads for building energy planning. J Build Eng 2022;45:103440. https://doi.org/10.1016/j.jobe.2021.103440.

[6] Bre F, Roman N, Fachinotti VD. An efficient metamodel-based method to carry out multi-objective building performance optimizations. Energy Build 2020;206:109576. https://doi.org/10.1016/j.enbuild.2019.109576.

[7] Singaravel Sundaravelpandian, Geyer Philipp, Suykens Johan. Deep-learning neural-network architectures and methods: using component-based models in building-design energy prediction. Adv Eng Inf 2018;38:81–90. https://doi.org/10.1016/j.aei.2018.06.004.

[8] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 2000;43:3–31. https://doi.org/10.1016/S0167-7012(00)00201-3.

[9] da Fonseca Raphaela Walger, Ruttkay Pereira Fernando Oscar. Metamodeling of the energy consumption of buildings with Daylight harvesting – application of artificial neural networks sensitive to orientation. J Daylighting 2021;8:255–69. https://doi.org/10.15627/jd.2021.20.

[10] Naji S, Keivani A, Shamshirband S, Alengaram UJ, Jumaat MZ, Mansor Z, et al. Estimating building energy consumption using extreme learning machine method. Energy 2016;97:506–16. https://doi.org/10.1016/j.energy.2015.11.037.

[11] Lopes M, Lamberts R. Development of a meta-model to predict cooling energy consumption of HVAC systems in office buildings in different climates. Sustainability 2018;10:4718. https://doi.org/10.3390/su10124718.

[12] Jia B, Hou D, Kamal A, Hassan IG, Wang L. Developing machine-learning meta-models for high-rise residential district cooling in hot and humid climate. J Build Perform Simulat 2022;15:553–73. https://doi.org/10.1080/19401493.2021.2001573.

[13] Chegari B, Tabaa M, Simeu E, Moutaouakkil F, Medromi H. Multi-objective optimization of building energy performance and indoor thermal comfort by combining artificial neural networks and metaheuristic algorithms. Energy Build 2021;239:110839. https://doi.org/10.1016/j.enbuild.2021.110839.

[14] Nagpal S, Mueller C, Aijazi A, Reinhart CF. A methodology for auto-calibrating urban building energy models using surrogate modeling techniques. J Build Perform Simulat 2019;12:1–16. https://doi.org/10.1080/19401493.2018.1457722.

[15] 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. https://doi.org/10.1016/j.apenergy.2021.116452.

[16] Moradzadeh A, Mohammadi-Ivatloo B, Abapour M, Anvari-Moghaddam A, Roy SS. Heating and cooling loads forecasting for residential buildings based on hybrid machine learning applications: a comprehensive review and comparative analysis. IEEE Access 2022;10:2196–215. https://doi.org/10.1109/ACCESS.2021.3136091.

[17] Mohandes SR, Zhang X, Mahdiyar A. A comprehensive review on the application of artificial neural networks in building energy analysis. Neurocomputing 2019;340:55–75. https://doi.org/10.1016/j.neucom.2019.02.040.

[18] Lu C, Li S, Lu Z. Building energy prediction using artificial neural networks: a literature survey. Energy Build 2022;262:111718. https://doi.org/10.1016/j.enbuild.2021.111718.

[19] Runge J, Zmeureanu R. Forecasting energy use in buildings using artificial neural networks: a review. Energies 2019;12:3254. https://doi.org/10.3390/en12173254.

[20] Wu W, Dandy GC, Maier HR. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ Model Software 2014;54:108–27. https://doi.org/10.1016/j.envsoft.2013.12.016.

[21] Maier HR, Jain A, Dandy GC, Sudheer KP. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ Model Software 2010;25:891–909. https://doi.org/10.1016/j.envsoft.2010.02.003.

[22] Á Bárkányi, Chován T, Németh S, Abonyi J. Modelling for digital twins—potential role of surrogate models. Processes 2021;9:476. https://doi.org/10.3390/pr9030476.

[23] Forrester Alexander IJ, Sobester Andras, Keane Andy J. Engineering design via surrogate modelling. first ed. John Wiley & Sons, Ltd; 2008. https://doi.org/10.1002/9780470770801.

[24] Jowett-Lockwood, Liam, Evins, Ralph. Using a convolutional neural network to determine the thermal characteristics of a building. Proceedings of eSim 2022: 12th conference of IBPSA-Canada, n.d.

[25] Osman ZH, Awad ML, Mahmoud TK. Neural network based approach for short-term load forecasting. In: 2009 IEEE/PES power systems conference and exposition; 2009. p. 1–8. https://doi.org/10.1109/PSCE.2009.4840035.

[26] Uba F, Apevienyeku HK, Nsiah FD, Akorli A, Adjignon S. Energy analysis of commercial buildings using artificial neural network. Model Simulat Eng 2021;2021:1–10. https://doi.org/10.1155/2021/8897443.

[27] Sadeghi A, Younes Sinaki R, Young WA, Weckman GR. An intelligent model to predict energy performances of residential buildings based on deep neural networks. Energies 2020;13:571. https://doi.org/10.3390/en13030571.

[28] Sharif SA, Hammad A. Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA. J Build Eng 2019;25:100790. https://doi.org/10.1016/j.jobe.2019.100790.

[29] Feurer Matthias. Hyperparameter optimization. Automated machine learning: methods, systems, challenges. Springer International Publishing; 2019. p. 3–33.

[30] Reynolds J, Hippolyte J-L, Rezgui Y. A smart heating set point scheduler using an artificial neural network and genetic algorithm. In: 2017 international conference on engineering, technology and innovation (ICE/ITMC). Funchal: IEEE; 2017. p. 704–10. https://doi.org/10.1109/ICE.2017.8279954.

[31] Wang R, Wang F, Nan Z, Xiao M, Ding A. Precise control for heating supply to households based on heating load prediction. In: Wang Z, Zhu Y, Wang F, Wang P, Shen C, Liu J, editors. Proceedings of the 11th international symposium on heating, ventilation and air conditioning (ISHVAC 2019). Singapore: Springer Singapore; 2020. p. 855–63. https://doi.org/10.1007/978-981-13-9524-6_89.

[32] Ascione F, Bianco N, De Masi RF, De Stasio C, Mauro GM, Vanoli GP. Chapter 11 - artificial neural networks for predicting the energy behavior of a building category: a powerful tool for cost-optimal analysis. In: Pacheco-Torgal F, Granqvist C-G, Jelle BP, Vanoli GP, Bianco N, Kurnitski J, editors. Cost-effective energy efficient building retrofitting. Woodhead Publishing; 2017. p. 305–40. https://doi.org/10.1016/B978-0-08-101128-7.00011-3.

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

[34] Saber A. Effects of window-to-wall ratio on: application of numerical and ANN approaches. J Soft Comput Civ Eng 2021;5. https://doi.org/10.22115/scce.2021.281977.1299.

[35] Luo XJ, Oyedele LO, Ajayi AO, Akinade OO. Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads. Sustain Cities Soc 2020;61:102283. https://doi.org/10.1016/j.scs.2020.102283.

[36] Luo XJ. A novel clustering-enhanced adaptive artificial neural network model for predicting day-ahead building cooling demand. J Build Eng 2020;32:101504. https://doi.org/10.1016/j.jobe.2020.101504.

[37] Dan TX, Phuc PNK. Application of machine learning in forecasting energy usage of building design. In: 2018 4th international conference on green technology and sustainable development (GTSD), Ho chi minh city. Vietnam: IEEE; 2018. p. 53–9. https://doi.org/10.1109/GTSD.2018.8595595.

[38] Sanaye S, Sarrafi A. Cleaner production of combined cooling, heating, power and water for isolated buildings with an innovative hybrid (solar, wind and LPG fuel) system. J Clean Prod 2021;279:123222. https://doi.org/10.1016/j.jclepro.2020.123222.

[39] Kim W, Jeon Y, Kim Y. Simulation-based optimization of an integrated daylighting and HVAC system using the design of experiments method. Appl Energy 2016;162:666–74. https://doi.org/10.1016/j.apenergy.2015.10.153.

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

[41] Pino-Mejías R, Pérez-Fargallo A, Rubio-Bellido C, Pulido-Arcas JA. Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions. Energy 2017;118:24–36. https://doi.org/10.1016/j.energy.2016.12.022.

[42] Ahmed MS, Mohamed A, Shareef H, Homod RZ, Ali JA. Artificial neural network based controller for home energy management considering demand response events. In: 2016 international conference on advances in electrical, electronic and systems engineering (ICAEES). Putrajaya, Malaysia: IEEE; 2016. p. 506–9. https://doi.org/10.1109/ICAEES.2016.7888097.

[43] Mui KW, Wong LT, Satheesan MK, Balachandran A. A hybrid simulation model to predict the cooling energy consumption for residential housing in Hong Kong. Energies 2021;14:4850. https://doi.org/10.3390/en14164850.

[44] Moayedi H, Mosavi A. Double-target based neural networks in predicting energy consumption in residential buildings. Energies 2021;14:1331. https://doi.org/10.3390/en14051331.

[45] Akkurt GG. Performance indices of soft computing models to predict the heat load of buildings in terms of architectural indicators. J Therm Eng 2017;3:1358–74. https://doi.org/10.18186/journal-of-thermal-engineering.330179.

[46] Jihad AS, Tahiri M. Forecasting the heating and cooling load of residential buildings by using a learning algorithm "gradient descent", Morocco. Case Stud Therm Eng 2018;12:85–93. https://doi.org/10.1016/j.csite.2018.03.006.

[47] Kim TY, Lee JM, Hong SH, Choi JM, Lee KH. Artificial neural network based optimized control of condenser water temperature set-point, vol. 5; 2021.

[48] Amasyali K, El-Gohary N. Deep learning for building energy consumption prediction. In: 6th CSCE-CRC international construction specialty conference 2017 - held as part of the Canadian society for civil engineering annual conference and general meeting 2017; 2017. p. 466–74.

[49] Himmetoğlu S, Delice Y, Aydoğan EK. PSACONN mining algorithm for multi-factor thermal energy-efficient public building design. J Build Eng 2021;34:102020. https://doi.org/10.1016/j.jobe.2020.102020.

[50] Jitkongchuen D, Pacharawongsakda E. Prediction heating and cooling loads of building using evolutionary grey wolf algorithms. In: 2019 joint international

conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT-NCON). Nan, Thailand: IEEE; 2019. p. 93–7. https://doi.org/10.1109/ECTI-NCON.2019.8692232.
[51] Elbeltagi E, Wefki H. Predicting energy consumption for residential buildings using ANN through parametric modeling. Energy Rep 2021;7:2534–45. https://doi.org/10.1016/j.egyr.2021.04.053.
[52] Ayoub N, Musharavati F, Pokharel S, Gabbar HA. ANN model for energy demand and supply forecasting in a hybrid energy supply system. In: 2018 IEEE international conference on smart energy grid engineering (SEGE). Oshawa, ON: IEEE; 2018. p. 25–30. https://doi.org/10.1109/SEGE.2018.8499514.
[53] Magalhães SMC, Leal VMS, Horta IM. Modelling the relationship between heating energy use and indoor temperatures in residential buildings through Artificial Neural Networks considering occupant behavior. Energy Build 2017;151:332–43. https://doi.org/10.1016/j.enbuild.2017.06.076.
[54] Kim J-H, Seong N-C, Choi W. Modeling and optimizing a chiller system using a machine learning algorithm. Energies 2019;12:2860. https://doi.org/10.3390/en12152860.
[55] Liu Y, Chen H, Zhang L, Feng Z. Enhancing building energy efficiency using a random forest model: a hybrid prediction approach. Energy Rep 2021;7:5003–12. https://doi.org/10.1016/j.egyr.2021.07.135.
[56] Lee S, Jung S, Lee J. Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. Energies 2019;12:608. https://doi.org/10.3390/en12040608.
[57] Zhenhua X, Xiangyang Z. Research on the ultra-short-time load prediction method of air source heat pump considering the input of neural network. In: 2018 China international conference on electricity distribution (CICED). Tianjin, China: IEEE; 2018. p. 260–3. https://doi.org/10.1109/CICED.2018.8592161.
[58] Shiralievich TS, Ivanovic SA, Mamanazarovna SS, Olimovich SO, Yunusov P. Learning algorithm of artificial neural network factor forecasting power consumption of users. Bulletin EEI 2022;11:602–12. https://doi.org/10.11591/eei.v11i2.3172.
[59] Khan AA, Minai AF, Devi L, Alam Q, Pachauri RK. Energy demand modelling and ANN based forecasting using MATLAB/simulink. In: 2021 international conference on control, automation, power and signal processing (CAPS). Jabalpur, India: IEEE; 2021. p. 1–6. https://doi.org/10.1109/CAPS52117.2021.9730746.
[60] Lee C, Jung DE, Lee D, Kim KH, Do SL. Prediction performance analysis of artificial neural network model by input variable combination for residential heating loads. Energies 2021;14:756. https://doi.org/10.3390/en14030756.
[61] Jaber AA, Saleh A, Mohammed Ali HF. Prediction of hourly cooling energy consumption of educational buildings using artificial neural network. Int J Adv Sci Eng Inf Technol 2019;9:159. https://doi.org/10.18517/ijaseit.9.1.7351.
[62] Ihsane I, Miegeville L, Ait-Ahmed N, Guerin P. New evaluation metrics for electrical demand forecasting: application to the residential sector. In: 2018 AEIT international annual conference. Bari, Italy: IEEE; 2018. p. 1–6. https://doi.org/10.23919/AEIT.2018.8577363.
[63] Ciulla G, D'Amico A, Lo Brano V, Traverso M. Application of optimized artificial intelligence algorithm to evaluate the heating energy demand of non-residential buildings at European level. Energy 2019;176:380–91. https://doi.org/10.1016/j.energy.2019.03.168.
[64] Keshtkarbanaeemoghadam A, Dehghanbanadaki A, Kaboli MH. Estimation and optimization of heating energy demand of a mountain shelter by soft computing techniques. Sustain Cities Soc 2018;41:728–48. https://doi.org/10.1016/j.scs.2018.06.008.
[65] Lin Y, Zhou S, Yang W, Li C-Q. Design optimization considering variable thermal mass, insulation, absorptance of solar radiation, and glazing ratio using a prediction model and genetic algorithm. Sustainability 2018;10:336. https://doi.org/10.3390/su10020336.
[66] Ekici B, Turkcan OFSF, Turrin M, Sariyildiz IS, Tasgetiren MF. Optimising high-rise buildings for self-sufficiency in energy consumption and food production using artificial intelligence: case of europoint complex in rotterdam. Energies 2022;15:660. https://doi.org/10.3390/en15020660.
[67] Kapetanakis D-S, Christantoni D, Mangina E, Finn DP. Evaluation of machine learning algorithms for demand response potential forecasting. 2017. 10.
[68] Wang Z, Zhu Y, Wang F, Wang P, Shen C, Liu J, et al. Comparative study of building energy use prediction based on three artificial neural network algorithms. In: Proceedings of the 11th international symposium on heating, ventilation and air conditioning (ISHVAC 2019). Singapore: Springer Singapore; 2020. https://doi.org/10.1007/978-981-13-9528-4_38. pp. 371–379.
[69] Ilbeigi M, Ghomeishi M, Dehghanbanadaki A. Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm. Sustain Cities Soc 2020;61:102325. https://doi.org/10.1016/j.scs.2020.102325.
[70] Hajj-Hassan M, Awada M, Khoury H, Srour I. A behavioral-based machine learning approach for predicting building energy consumption. Construction research congress. Tempe, Arizona: American Society of Civil Engineers; 2020. p. 1029–37. https://doi.org/10.1061/9780784482865.109. 2020.
[71] Santos-Herrero JM, Lopez-Guede JM, Flores Abascal I, Zulueta E. Energy and thermal modelling of an office building to develop an artificial neural networks model. Sci Rep 2022;12:8935. https://doi.org/10.1038/s41598-022-12924-9.
[72] Li X, Yao R. Modelling heating and cooling energy demand for building stock using a hybrid approach. Energy Build 2021;235:110740. https://doi.org/10.1016/j.enbuild.2021.110740.
[73] D'Amico A, Ciulla G, Traverso M, Lo Brano V, Palumbo E. Artificial Neural Networks to assess energy and environmental performance of buildings: an Italian case study. J Clean Prod 2019;239:117993. https://doi.org/10.1016/j.jclepro.2019.117993.
[74] Ngo N-T. Early predicting cooling loads for energy-efficient design in office buildings by machine learning. Energy Build 2019;182:264–73. https://doi.org/10.1016/j.enbuild.2018.10.004.
[75] Li Z, Dai J, Chen H, Lin B. An ANN-based fast building energy consumption prediction method for complex architectural form at the early design stage. Build Simulat 2019;12:665–81. https://doi.org/10.1007/s12273-019-0538-0.
[76] Martellotta F, Ayr U, Stefanizzi P, Sacchetti A, Riganti G. On the use of artificial neural networks to model household energy consumptions. Energy Proc 2017;126:250–7. https://doi.org/10.1016/j.egypro.2017.08.149.
[77] Gao W, Moayedi H, Shahsavar A. The feasibility of genetic programming and ANFIS in prediction energetic performance of a building integrated photovoltaic thermal (BIPVT) system. Sol Energy 2019;183:293–305. https://doi.org/10.1016/j.solener.2019.03.016.
[78] Azari R, Garshasbi S, Amini P, Rashed-Ali H, Mohammadi Y. Multi-objective optimization of building envelope design for life cycle environmental performance. Energy Build 2016;126:524–34. https://doi.org/10.1016/j.enbuild.2016.05.054.
[79] Roy SS, Samui P, Nagtode I, Jain H, Shivaramakrishnan V, Mohammadi-ivatloo B. Forecasting heating and cooling loads of buildings: a comparative performance analysis. J Ambient Intell Hum Comput 2020;11:1253–64. https://doi.org/10.1007/s12652-019-01317-y.
[80] Seyedzadeh S, Pour Rahimian F, Rastogi P, Glesk I. Tuning machine learning models for prediction of building energy loads. Sustain Cities Soc 2019;47:101484. https://doi.org/10.1016/j.scs.2019.101484.
[81] Cant, Kevin, Evins, Ralph. Improved calibration of building models using approximate Bayesian calibration and neural networks n.d. https://www.tandfonline.com/doi/epdf/10.1080/19401493.2022.2137236?needAccess=true&role=button (accessed January 15, 2023).
[82] Zhang H, Feng H, Hewage K, Arashpour M. Artificial neural network for predicting building energy performance: a surrogate energy retrofits decision support framework. Buildings 2022;12:829. https://doi.org/10.3390/buildings12060829.
[83] Chegari B, Tabaa M, Simeu E, Moutaouakkil F, Medromi H. An optimal surrogate-model-based approach to support comfortable and nearly zero energy buildings design. Energy 2022;248:123584. https://doi.org/10.1016/j.energy.2022.123584.
[84] Bre F, Roman N, Fachinotti V. An efficient metamodel-based method to carry out multi-objective building performance optimizations | Elsevier Enhanced Reader n.d. https://doi.org/10.1016/j.enbuild.2019.109576.
[85] Keivan Bamdad, Michael E. Cholette, John Bell. Building energy optimization using surrogate model and active sampling n.d.
[86] Gonçalves D, Sheikhnejad Y, Oliveira M, Martins N. One step forward toward smart city Utopia: smart building energy management based on adaptive surrogate modelling. Energy Build 2020;223:110146. https://doi.org/10.1016/j.enbuild.2020.110146.
[87] He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowl Base Syst 2021;212:106622. https://doi.org/10.1016/j.knosys.2020.106622.
[88] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting Items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264–9. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.