Energy Reports 7 (2021) 5003–5012

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Enhancing building energy efficiency using a random forest model: A hybrid prediction approach

Yang Liu [a,b,*,1], Hongyu Chen [b,c,1], Limao Zhang [c], Zongbao Feng [d,**]

[a] *Wuhan Univ., Zhongnan Hosp. of Wuhan University, Wuhan 430071, PR China*
[b] *School of Economics and Management, Wuhan University, Wuhan 430072, PR China*
[c] *School of Civil and Environmental Engineering, Nanyang Technological University, 639798, Singapore*
[d] *Huazhong University of Science & Technology, School of Civil and Hydraulic Engineering, Wuhan 430074, PR China*

ARTICLE INFO

Article history:
Received 25 January 2021
Received in revised form 21 July 2021
Accepted 29 July 2021
Available online 17 August 2021

Keywords:
Building energy prediction
Building envelope
BIM
DesignBuilder
Random forest

ABSTRACT

The building envelope considerably influences building energy consumption. To enhance the energy efficiency of buildings, this paper proposes an approach to predict building energy consumption based on the design of the building envelope. The design parameters of the building envelope include the comprehensive heat transfer coefficient and solar radiation absorption coefficient of exterior walls, comprehensive heat transfer coefficient and solar radiation absorption coefficient of the roof, comprehensive heat transfer coefficient of outer windows, and window-wall ratio. The approach is applied to optimize the design parameters of the building envelope structure of a university teaching building in northern China. First, a building information model of a teaching building is established in *Revit* and imported into DesignBuilder energy consumption analysis software. Subsequently, a data set of the abovementioned 6 parameters is obtained by performing orthogonal testing and energy consumption simulations. On this basis, an RF model is used to predict building energy consumption and rank the importance of each parameter, and the Pearson function is used to evaluate the corresponding correlations. The results show that the most important parameters with the highest correlations to building energy consumption are the comprehensive heat transfer coefficients of the exterior walls and outer windows and the window-wall ratio. Finally, the RF prediction results are compared to the prediction results of a BP artificial neural network (BP-ANN) and support vector machine (SVM). The findings indicate that the RF model exhibits notable advantages in building energy consumption prediction and is the optimal prediction model among the compared models.

1. Introduction

As one of three major energy-intensive fields, the building sector is an essential source of energy consumption and greenhouse gas emissions in urban areas worldwide (Tian et al., 2020b); notably, this sector contributes to 40% of the total global primary energy consumption and one-third of the total amount of CO_2-eq emissions (Liu et al., 2020a; He et al., 2014). For instance, in China, approximately 21.1% of energy consumption and 19.5% of CO_2-eq emissions were associated with the building sector in 2017 (CABEE, 2018). Therefore, minimizing this energy demand is important for ensuring sustainability in the building sector and achieving energy conservation and pollutant emission reduction goals (He, 2019; Belussi et al., 2019). It is necessary to accurately predict building energy loads to enhance the operational performance of building energy systems due to the key role of such predictions in implementing building energy efficiency measures (Abdelkader, 2020); certain related tasks include building energy operation system measurement and verification (Gallagher et al., 2019), system fault diagnosis and detection (Li et al., 2016), demand response control (Pedersen et al., 2017), and building energy benchmarking (Zhao and Magoulès, 2012).

Before predicting energy consumption, the energy consumption characteristics of a building must be analyzed. A building and its surrounding environment are a complex system, and all subsystems of the building affect its energy efficiency. Furthermore, the interdependence among systems is critical. These factors increase the complexity of the thermal processes and interactions in buildings, and it is often challenging to calculate energy consumption (Ceballos-Fuentealba et al., 2019). To effectively understand the complex interactions that occur in such

---

[*] Corresponding author at: Wuhan Univ., Zhongnan Hosp. of Wuhan University, Wuhan 430071, PR China.
[**] Corresponding author.

E-mail addresses: dabailiu@whu.edu.cn (Y. Liu), m201873595@hust.edu.cn (Z. Feng).

[1] Yang Liu and Hongyu Chen contributed equally to the work.

https://doi.org/10.1016/j.egyr.2021.07.135

Nomenclature

Abbreviation

RF	Random forest
CTF	Conduction transfer function
BIM	Building information modeling
SVM	Support vector machine
BP-ANN	Backpropagation artificial neural network
BESTEST	Building energy simulation test
IEA	International Energy Agency

Nomenclature

OBB	Out-of-bag data
E^*	Generalization error
X	Probability P covering the X space
Y	Probability P covering the Y space
$\overline{\rho}$	Average correlation coefficient of the decision tree
s	Average strength of the decision tree
N	Number of random forest decision trees
ntree	Number of regression trees
mtry	Number of random features
RMSE	Root mean square error
R^2	Certainty coefficient
j	Number of samples
MSE_{ij}	Mean square residual of the *j*th sample
S_E	Standard error
$Cov\,(X, Y)$	Covariance of variables *X* and *Y*
μ_X	Average value of variable *X*
μ_Y	Average value of variable *Y*
σ_X	Deviation of variable *X*
σ_Y	Deviation of variable *Y*
p	Correlation coefficient
A_1	Comprehensive heat transfer coefficient of exterior walls
A_2	Solar radiation absorption coefficient of exterior walls
A_3	Comprehensive heat transfer coefficient of the roof
A_4	Solar radiation absorption coefficient of the roof
A_5	Comprehensive heat transfer coefficient of outer windows
A_6	Window-wall ratio

systems, building designers generally rely on computational modeling and simulation programs to support the analysis of building energy patterns (Elbeltagi et al., 2017). A modeling approach is typically selected based on the characteristics of the modeling scenario, applied building technology and building envelope (Ascione et al., 2019). The heat exchange between a building and the external environment is influenced by the performance of the building envelope, which therefore affects the building energy consumption level for general-purpose buildings. More than 50% of energy consumption can be attributed to heat dissipation through the building shell (Fan and Xia, 2017); thus, increasing the energy efficiency of building envelopes is a priority for enhancing predictions of building energy consumption (Aslani et al., 2019). Consequently, several researchers have focused on effective building envelope designs/retrofits, and many relevant strategies have been investigated to reduce the environmental impacts of buildings. Additionally, the evolution of the design approach, supported by the availability of new tools (e.g., building information modeling (BIM)), has contributed to effective building envelope design and retrofitting (Zhou et al., 2019).

In this paper, we develop a hybrid approach that considers the actual conditions of buildings based on BIM and apply an ad hoc tool that incorporates simulation and optimization models based on DesignBuilder. A random forest (RF)-based approach is adopted for building energy prediction. Model training and prediction are performed to validate the feasibility and accuracy of using the RF approach in estimating building energy consumption. Moreover, a comparative analysis is performed to compare the prediction performance of RF with that of a BP-ANN model and an SVM. In addition, this paper provides insight into the sorting of the importance of each building envelope design parameter used in RF; the result can assist researchers in identifying pivotal impact variables and gaining a comprehensive understanding of the energy behavior of a given building.

The remainder of this paper is organized as follows. Section 2 reviews the literature regarding methods to acquire building energy consumption data and methods to predict the energy consumption of buildings. Section 3 describes the proposed methodology used to perform building energy consumption simulation and data set acquisition with BIM and DesignBuilder; additionally, the application of the RF algorithm to predict building energy consumption is introduced. Finally, Section 5 discusses the results obtained from the case study, presents the key conclusions and highlights several possible directions for future research.

2. Literature review

2.1. Acquisition of building energy consumption data

Building energy consumption data are the basis of the research on building energy consumption. The methods used to obtain building energy consumption data can be divided into three categories.

(1) Building load calculations. To calculate the load associated with building energy consumption, static (Wang et al., 2014) and dynamic (Wan et al., 2017) calculation methods based on stable and unstable heat transfer theories, respectively, can be applied. The static simulation method is a simplified approach that assumes that the internal and external environments of a building remain unchanged and ignores the effects of building heat storage, leading to a simple and fast calculation process. However, compared with that of the dynamic calculation method, the accuracy of the static simulation method is insufficient. Although the dynamic simulation method can overcome the shortcomings of the static simulation method, the dynamic calculation of building energy consumption is complex and time consuming. With the development of computer-based numerical simulations, building load calculation methods are gradually being replaced.

(2) Numerical simulation. In numerical simulations of building energy consumption, building energy consumption is simulated through computer simulation technology, and building energy consumption models are established based on dynamic calculations of energy consumption. There are many types of building energy simulation software, and software packages have been developed and improved for over 30 years. Currently, the most popular building energy simulation software packages include DOE-2 (Andolsun et al., 2012), DeST (Zhang et al., 2008), EnergyPlus (Al-janabi et al., 2019a), and DesignBuilder (Anwar et al., 2020), among others.

(3) Monitoring and measurement. Building energy consumption monitoring and measurement refers to the use of instruments to measure the energy consumption of a building in a certain period of time and to identify key parameters that affect building energy consumption (Chen et al., 2021; Gatt et al., 2020). In particular, energy consumption monitoring and measurement is a valuable and authoritative approach to obtain energy consumption data and can guide building energy consumption audits and energy-saving renovations.

Among the three methods used to obtain building energy consumption data, the building load calculation method is suitable for the early stage of building plan design, as the building energy consumption is preliminarily predicted before the building design plan is formed. The numerical simulation approach is suitable for determining the energy consumption level after the specific building design plan has been completed. In the context of predicting the energy consumption for building energy-saving renovation programs, monitoring and measurement represent the most accurate techniques that can be used to obtain building energy consumption data during actual building use. When examining the energy efficiency of a building, the combination of numerical simulations and monitoring and measurement results is the most commonly used approach to obtain building energy consumption data, and this method is used in this paper.

2.2. Building energy consumption prediction

Based on the methods for collecting energy consumption data discussed in the previous section, the future energy consumption distribution of buildings can be predicted. Predicting energy consumption is important for building energy efficiency evaluation, multienergy scheduling and energy-saving strategy formulation (Tian et al., 2020c). In the literature pertaining to the prediction of energy consumption in the construction industry, two approaches have been widely applied by researchers: statistical methods and empirical learning algorithms (Lei et al., 2021; Zhang et al., 2020).

Statistical methods are also called gray box models, and large amounts of energy consumption data are typically needed for regression analysis. Aranda et al. (2012) developed a power model based on energy consumption data from 55 banks in Spain and later performed multiple linear regression for prediction; a comparison verified the excellent performance of the model in predicting energy consumption. Braun et al. (2014) used multiple regression analysis to investigate a supermarket in northern England based on gas and electricity data for 2012 to estimate consumption during the climate period of 2030–2059. Ayoub (2019) constructed a data set of several influential factors to develop multivariate linear regression models and found that the simulation results were in agreement with the actual values, suggesting that building performance can be successfully predicted in the early stages of design. However, the establishment of engineering models is often challenging and time consuming for complex building energy systems. In addition, a recent study showed that certain methods may be associated with large differences between the predicted and actual values of building energy consumption (Ryan and Sanquist, 2012).

With the development of monitoring technologies and the growing popularity of building automation systems, massive amounts of construction data have been generated (Pan and Zhang, 2020). These data provide an opportunity to adopt empirical learning algorithms to achieve building energy prediction. Park et al. (2018) used artificial neural network models to predict energy costs by combining historical operational data and climate data. Liu et al. (2020) presented a public building energy consumption prediction model based on SVM theory. Lu et al. (2020) predicted building energy consumption based on extreme gradient boosting and considered the daily energy consumption of the City of Bloomington Intake Tower as the simulation object to verify the prediction accuracy of the method; compared with an engineering-based linear regression method, the proposed method was more practical for building energy prediction because the required data, such as building energy consumption data, environmental data and occupancy data, are relatively easy to obtain.

In this section, the applications of engineering-based linear regression methods and empirical learning algorithms for building energy prediction are described. Based on the literature review, it can be concluded that the following issues still need to be addressed.

- Building energy consumption prediction models based on statistical methods provide satisfactory generalization ability and interpretability; however, the prediction accuracy is inadequate, and the calculations are complicated.
- The use of certain empirical learning algorithms, e.g., extreme gradient boosting and artificial neural networks, may yield unstable results; notably, a small change in the input data may result in a significant change in the output data, thereby affecting the prediction accuracy.

Compared to conventional two-dimensional (2D) drawings, BIM provides geometrically accurate three-dimensional (3D) representations, which provide a realistic and enriched model that benefits all phases of the building life cycle (Zhang et al., 2016). For example, Autodesk Revit provides powerful tools for supporting architectural design, architectural services, engineering design, and structural engineering design. DesignBuilder is a comprehensive software based on a built-in database of building structures, which are used to simulate and analyze building thermal performance, the annual load distribution and other factors (Etxebarria-Mallea et al., 2021). As a representative multifaceted learning algorithm, an RF is used to train the base models. An RF model provides excellent prediction and classification performance because the prediction results are derived from the comprehensive predictions of multiple decision trees, and no overfitting occurs with increasing number of trees (Stephan et al., 2015). Thus, we imported the BIM model into DesignBuilder energy consumption analysis software. Then, case studies were explored using the simulation tool to evaluate the current thermal performance of the studied building envelope. The objective was to effectively increase the accuracy of building energy consumption prediction based on physical models and reduce the complexity of model construction. Furthermore, the proposed method can help architects and decision makers in assessing preliminary building performance by considering environmental factors and facade configurations without the need to perform exhaustive analyses.

3. Methodology

3.1. Building energy simulation

3.1.1. Designbuilder

DesignBuilder is a user graphical interface software that was specially developed for EnergyPlus to enhance the visualization performance and ease of use of the software (Liu et al., 2019). The software is divided into five modules: a visual modeling module, an EnergyPlus dynamic thermal simulation core, an HVAC module, a CFD module, and a sunshine environment module. Through the visual modeling module, a user can not only use the default modeling functions but also import existing CAD drawings and building information models. In addition, DesignBuilder includes many authoritative databases and supports user-defined databases (Ascione et al., 2020). Various input parameters are templated to provide a high degree of freedom, enhance the simulation accuracy and reduce the modeling duration. The simulation results can be output through various plots and can also be displayed at monthly or daily scales (Al-janabi et al., 2019b).

Among the numerical simulation methods described in Section 2.2, DesignBuilder and EnergyPlus provide the most complete functionalities and yield the best simulation results. DesignBuilder has a graphical interface that shows the general structure of the building model during the energy consumption analysis. Thus, DesignBuilder software is used to simulate building energy consumption in this paper. The energy consumption calculations are based on the thermal balance method. A method based on the conduction transfer function (CTF) is used to calculate the heat transfer of the outer enclosure structure. The CTF is related to reaction coefficients. The core step in this method is solving the thermal equilibrium equation for air in a given region. The temperature variable in the equation is based on the temperature of the inner surface of the wall rather than the temperature of indoor air. Therefore, the calculation results are similar to the actual thermal levels (Solla et al., 2019).

3.1.2. Building energy consumption simulation based on BIM and designbuilder

Compared to conventional 2D drawings, BIM provides more realistic and enriched models for all phases of the building life cycle. In particular, BIM offers a geometrically accurate 3D representation and the capability to affiliate attributes and data to different components and objects inside a model. Since BIM technology provides visualization, coordination, simulation and optimization functionalities, various software programs with BIM technology can be used to realize the automation, intelligent design and analysis of green buildings (Solla et al., 2019). The primary objective is to use the relevant BIM information-based architectural design software to establish a 3D building information model based on the relevant physical and geometric information for a given construction project. The steps required to apply the building information model in DesignBuilder software are as follows (Cheng et al., 2020).

(1) When constructing a building information model in Revit software with reference to CAD drawings, the building walls, doors, windows, floors, roofs and other components are drawn sequentially, and the building material, size, heat transfer coefficient and other attributes are input.
(2) The rooms are divided on each floor, the room numbers are marked, and the room heights are set to facilitate the export of the building model.
(3) After converting the *rvt* file generated by Revit to *gbXML* format, the file is imported into the design builder software to generate an *skh*-format model.
(4) By setting the corresponding parameters in the model in DesignBuilder software, the energy consumption of the building model can be calculated.
(5) Then, DesignBuilder is used to simulate energy consumption and perform orthogonal tests, and a sample data set is obtained.

3.2. Overview of the RF algorithm

The foundation of the RF model is classification tree theory. The greatest advantage of an RF as a predictive model is its high generalization ability, which can effectively prevent overfitting in data prediction (Wang et al., 2018). In the process of extracting a sample set, the probability of not being selected is $\left(1-\frac{1}{N}\right)^N$. When $N \to \infty$,

$$\left(1-\frac{1}{N}\right)^N \approx \frac{1}{e} \approx 0.368 \tag{1}$$

Approximately 36.8% of the samples are retained in the original sample set in each extraction step. These retained samples are known as out-of-bag (OBB) data (Guo et al., 2020), and the generalization error can be calculated as follows:

$$E^* = P_{X,Y}(M(X,Y) \lhd 0) \tag{2}$$

where X and Y indicate the probability P of covering the X and Y spaces, respectively. In RF, when the number of decision trees is sufficiently large (Tian et al., 2020a), E^* converges to the following form as the number of trees increases:

$$P_{X,Y}(P_\theta(h(X,\theta)=Y) - \max_{\mathrm{j}\neq Y} P_\theta(\mathrm{h}(X,\theta)=\mathrm{j}) \lhd 0) \tag{3}$$

This formula indicates that the generalization error of the model does not cause overfitting as the number of decision trees increases (Kim et al., 2020), and the error approaches a finite upper bound. Thus, the following equation can be established:

$$E^* = \frac{\overline{\rho}\left(1-s^2\right)}{s^2} \tag{4}$$

where $\overline{\rho}$ and s denote the average correlation coefficient and average strength of a decision tree, respectively. The finite upper bound of the RF generalization error is positively correlated with a decrease in $\overline{\rho}$ and negatively correlated with s; therefore, the generalization error can be effectively controlled.

3.3. Establishment of an RF prediction model

The process flow of the building energy consumption prediction model constructed in this paper based on an RF approach is shown in Fig. 1. First, the initial index system of building energy consumption and the original data set are established. Second, the RF model and Pearson algorithm are used to analyze the importance of each influential factor and the correlations among the factors. Next, an RF training model is established, and regression predictions are obtained. Finally, an error analysis of the predicted results is performed (Cao et al., 2020).

This study examines the influence of the design parameters on building energy consumption. Based on the simulation results of building energy consumption, six factors closely related to building energy consumption are selected as the input indexes of the prediction model (An et al., 2020; Nasir and Hassan, 2020; Marino et al., 2017): the comprehensive heat transfer coefficient and solar radiation absorption coefficient of the exterior walls, the comprehensive heat transfer coefficient and solar radiation absorption coefficient of the roof, the comprehensive heat transfer coefficient of the outer windows and the window-wall ratio. The simulated values of building energy consumption are selected to establish the evaluation index, and the initial index system is constructed. Different influential factors in the index system of building energy consumption are used as variables in the RF framework, and related experiments are performed to obtain the original data set.

3.4. Parameter setting and RF modeling

(1) Sample diversity. The training sample set of the RF model includes approximately two-thirds of the data in the original data set, and the remaining data, or one-third of the original data set, constitute the test sample set of the RF model.

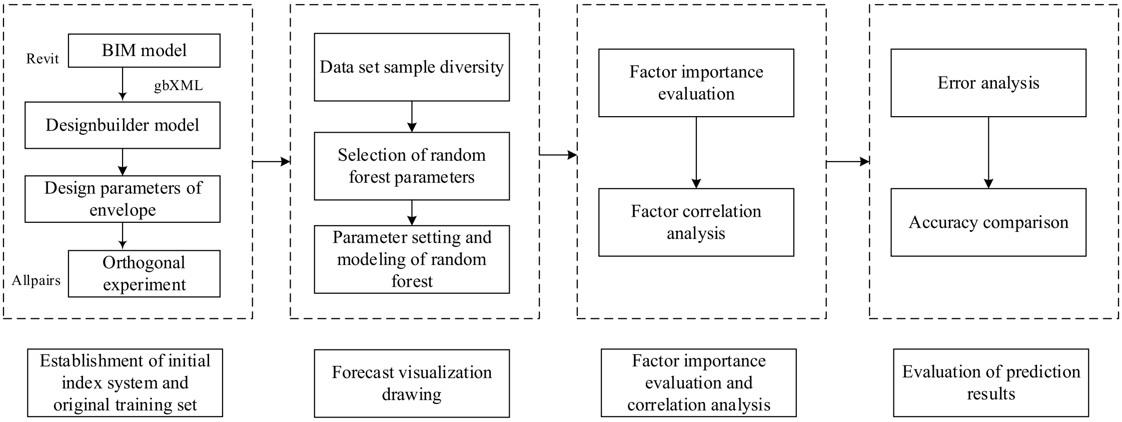

Fig. 1. Technical roadmap.

(2) Parameter selection. In the modeling process, two parameters are mainly involved: the number of regression trees, *ntree*, and the number of random features, *mtry*. The value of ntree affects the training degree and accuracy of the RF model. Generally, a large ntree corresponds to high model accuracy. The value of mtry controls the degree of influence of the RF model attributes. mtry is generally determined according to empirical formulas. The number of variables is P, and by default, $mtry = P$ (classification model) or $mtry = P/3$ (regression model). After determining the two parameters, the RF algorithm can be used to establish a random Sen regression model based on the training and testing sample sets.

3.5. Importance evaluation and correlation analysis

(1) Importance evaluation. The variable importance score can be used to evaluate the degree of influence of the characteristic variables on the dependent variables to determine the key variables to be controlled. In this paper, the importance of the feature variables is calculated according to Eq. (5); that is, the OBB error is used to evaluate the importance of the feature variables. The reduction in the mean square residual error is measured after the random replacement and model accuracy calculations. The degree of decrease is used to evaluate the importance of the feature variables. The specific calculation steps are as follows.

- A regression decision tree is established for each training sample set to obtain b mean square residuals: $MSE_1, MSE_2 \ldots MSE_b$.
- After constructing the regression decision tree to be split, the characteristic variables are randomly selected. By randomly replacing the variable Xi in b samples of OOB data, a new OOB test set can be formed. Predictions are repeated to obtain a new OOB residual mean square MSE_{ij}, and matrix A is generated as follows:

$$A = \begin{bmatrix} MSE_{11} & \ldots & MSE_{1b} \\ \ldots & \ldots & \ldots \\ MSE_{p1} & \ldots & MSE_{pb} \end{bmatrix} \tag{5}$$

  where p is the number of influential factors and b is the number of training sample sets.
- The mean square residual reduction in the variable Xi can be obtained by subtracting the corresponding row of MSE_1, $MSE_2 \ldots MSE_b$ and matrix B and dividing it by the standard error. $VIM_i\,(MSE)$ is the importance score of the characteristic variables, which can be expressed as follows:

$$VIM_i\,(MSE) = \frac{\left(\frac{1}{b}\sum_{j=1}^{b}\left(MSE_j - MSE_{ij}\right)\right)}{S_E}, (1 \le i \le p) \tag{6}$$

where MSE_{ij} is the mean square residual of the jth sample and S_E is the standard error. The RF change factor is adjusted at various locations to add noise to the model and assess the resulting accuracy. A high variation in model accuracy corresponds to the feature variable having a large impact on the output variable and being important to the model.

(2) Correlation analysis. This paper uses the Pearson correlation coefficient to analyze the linear correlation between each mix ratio factor and building energy consumption, and the corresponding correlations are assessed. The Pearson correlation coefficient is calculated as the product of the mean difference and sum of squares of the mean difference between two variables; therefore, this coefficient is also known as the correlation coefficient of the product difference. The overall Pearson correlation coefficient can be expressed as:

$$p = \frac{\mathrm{cov}\,(X, Y)}{\sigma_X \sigma_Y} = \frac{E\,(X - \mu_X)\,(Y - \mu_Y)}{\sigma_X \sigma_Y} \tag{7}$$

where $Cov\,(X, Y)$ is the covariance of variables X and Y, μ_X and μ_Y are the average values of variables X and Y, respectively, and σ_X and σ_Y are the deviations of variables X and Y, respectively. The value of the correlation coefficient p lies within the interval $[-1, 1]$. A strong linear relationship between two variables corresponds to a value of p that is close to 1 or -1. When $p > 0$ or $p < 0$, a positive or negative correlation exists between two variables, respectively. When $p = 0$, no linear correlation exists between the two variables, although the variables may be related in other ways.

3.6. Evaluation of the prediction results

To evaluate the prediction accuracy of the RF model, the root mean square error ($RMSE$) and goodness of fit (R^2) are selected as the evaluation parameters. and R^2 can be used to describe the deviation and degree of correlation between the predicted and measured values, respectively. Additionally, the prediction results obtained using the RF model are compared with those obtained using the BP-ANN and SVM models to verify the superiority of the RF model. The expressions of the mean square error and goodness of fit are shown in Eqs. (8) and (9), respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{obs} - y^{pred}\right)}{n}} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y^{obs} - y^{pred}\right)^2}{\sum_{i=1}^{n}\left(y^{obs} - \overline{y}^{obs}\right)^2} \tag{9}$$

Table 1
Basic settings for the energy consumption simulation.

Basic applied settings	Value
Mean outdoor temperature (°C)	7
Heating temperature setting (°C)	25
Heating energy efficiency ratio	2.5
Indoor lighting power (W/m^2)	5
Teaching equipment power (W/m^2)	4.5
Occupant density (People/m^2)	0.3

3.7. Case background

A new teaching building is to be built at a university in northern China. The teaching building is located in the southwest corner of the university. The main part of the teaching building is expected to adopt a nonshaded reinforced concrete frame structure with a floor area of 33 350 m^2. The building footprint covers an area of 11 573 m^2, the total building area is 40 318 m^2, and the volume ratio is 1.217. The teaching building has five floors, the layout of each floor is the same, and each floor is divided into four main areas: a classroom area, a teacher rest area, a bathroom area and a corridor area. The energy source of the teaching building is electric power, and a central air conditioning system is used to adjust the room temperatures. The classroom area and teacher rest area have the same heating area, and the bathroom area and corridor area have no heating areas. The heating temperature of the classroom area and teacher rest area is set as 25 °C, and the starting temperature for heating is 10 °C (i.e., if the temperature is less than this value, automatic heating is initiated). According to the relevant specifications and requirements of the actual project, several basic applied settings in the model must be determined before using DesignBuilder for energy consumption simulations, as summarized in Table 1.

4. Results and discussion

4.1. Building energy consumption simulation and data set acquisition based on BIM and DesignBuilder

To establish the model of building energy consumption for the teaching building, first, a building information model of the building is built in Revit according to CAD drawings. Next, the model is imported into DesignBuilder for regional division. Finally, the simulation model of building energy consumption is obtained, as shown in Fig. 2.

After establishing the building energy consumption model based on BIM, an orthogonal experimental design was used to obtain reliable and representative building energy consumption data.

In orthogonal experimental design, optimal design schemes are obtained based on probability theory, mathematical statistics and practical experience (Huang and Li, 2017). This approach considers not only the scientific nature of the analysis but also the workload and complexity of the experiment, thereby ensuring the uniformity and broad applicability of the scheme to a certain extent. The design parameters of the outer envelope more notably influence the energy consumption of the building than those of the inner envelope. Therefore, in this paper, the comprehensive heat transfer coefficient of the exterior walls (A_1) (Goncalves et al., 2020), solar radiation absorption coefficient of the exterior walls (A_2) (Fantucci et al., 2020), comprehensive heat transfer coefficient of the roof (A_3) (Liu et al., 2021), solar radiation absorption coefficient of the roof (A_4) (Li et al., 2015), comprehensive heat transfer coefficient of the outer windows (A_5) (Heydari et al., 2021) and window-wall ratio (A_6) (Phillips et al., 2020) are used as the building energy simulation parameters; the recommended value range of each envelope structure design parameter is rationally divided into factor classes. Allpairs software is used to design an orthogonal table, as shown in Table 2.

Table 2
Orthogonal design.

Index factor	A_1 W/(m^2 K)	A_2 W/(m^2 K)	A_3	A_4 W/(m^2 K)	A_5	A_6
1	1.192	0.276	0.3	0.228	0.3	0.15
2	1.474	0.35	0.4	0.293	0.4	0.25
3	1.96	0.487	0.5	0.326	0.5	0.35
4	2.529	0.567	0.6	0.385	0.6	0.45
5	2.72	0.648	0.7	0.436	0.7	0.55
6	3.189	0.755	0.75	0.486	0.75	0.65

Table 3
Orthogonal test results.

Index factor	A_1 W/(m^2 K)	A_2 W/(m^2 K)	A_3	A_4 W/(m^2 K)	A_5	A_6	*Energy consumption* (kWh)
1	1.192	0.276	0.3	0.228	0.3	0.15	8204.71
2	3.189	0.567	0.7	0.228	0.5	0.65	9282.515
3	1.474	0.276	0.3	0.436	0.75	0.25	8453.471
…	…	…	…	…	…	…	…
52	2.529	0.755	0.7	0.326	0.3	0.25	8967.419
53	2.72	0.487	0.3	0.486	0.75	0.55	8394.225
54	1.192	0.755	0.4	0.385	0.3	0.55	8260.067

DesignBuilder is used to simulate the building energy consumption level for 54 sets of tests. Each test is individually input into DesignBuilder, and the software is used to simulate the energy consumption of the building. The results of the orthogonal test of building energy consumption are presented in Table 3.

4.2. Prediction of building energy consumption based on an RF model

(1) Establishment of the original training set. Using the simulation described in Section 3.2, 54 sets of simulation results for building energy consumption are obtained and used to establish an indicator system for the design parameters of the initial envelope structure. The specific input indicators are A_1, A_2, A_3, A_4, A_5, and A_6.

(2) Sample diversity. The training sample set of the RF model includes approximately 80% of the original data, and the test sample set includes the remaining 20% of the original data. This case involves 54 sets of sample data, and thus, 43 training sample subsets and 11 test subsets are used. The training subsets are applied to determine the RF parameters, build the RF model and obtain importance scores, and the test subsets are used to evaluate and verify the predictive performance of the model.

(3) Parameter selection. The number of random features *mtry* in the regression model is generally set to one-third the number of variables in the data set. The evaluation index system in this case contains 6 influential factors. Notably, the result is optimal when $mtry = 2$. According to theoretical research, the generalization error of an RF regression model gradually converges as *ntree* increases. Therefore, a sufficiently large *ntree* value should be selected to build the model and ensure that the training error stabilizes. An OOB data error estimation method is used to determine the value of *ntree*, and the result is shown in Fig. 3. When the number of decision trees $ntree > 400$, the OBB error rate tends to stabilize, and thus, $ntree = 600$ is selected in this case.

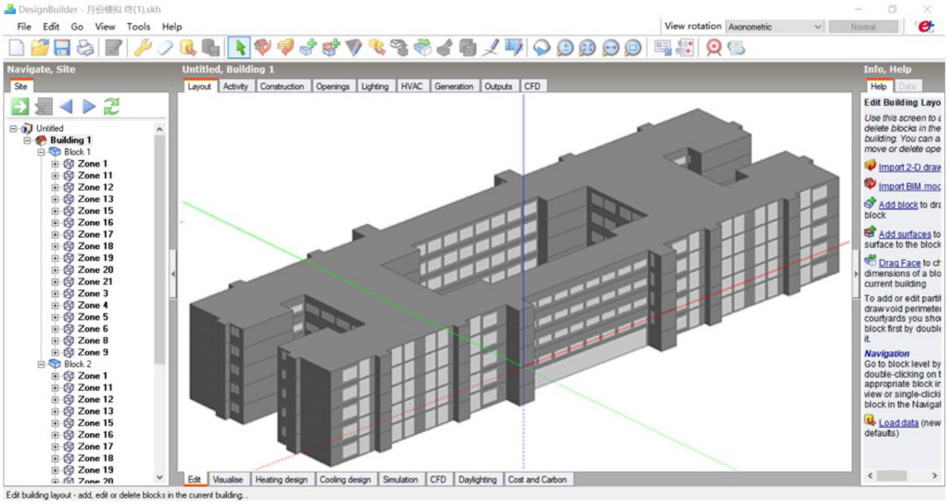

Fig. 2. Diagram of the energy consumption simulation model for the teaching building based on BIM.

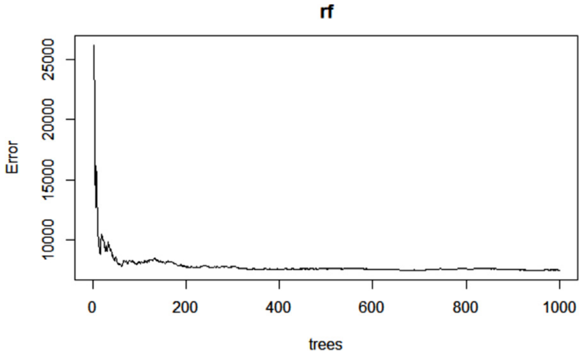

Fig. 3. Selection of the number of decision trees.

4.3. Importance evaluation and correlation analysis of the variables

(1) Importance calculation. An importance evaluation can be performed to determine the degree of importance of the envelope design parameters in the estimation of building energy consumption. The importance score of each factor can be calculated according to Eq. (6) presented in Section 3.5. Subsequently, the importance scores of the influential factors are arranged in descending order, and the importance of each variable in the training model is evaluated, as shown in Fig. 4. A high range of variation of a variable, based on node purity, corresponds to the large influence of the variable on the evaluation index and the high importance of the variable. Fig. 4 shows that the importance levels of variables such as A_1, A_5 and A_6 are relatively large, indicating that these variables exert a notable impact on building energy consumption and must be considered in engineering practice.

(2) Correlation calculation. The correlations between the influential factors and building energy consumption can be analyzed by using the Pearson function. The Pearson correlation coefficients can be calculated using Eq. (8) presented in Section 3.6, and a correlation diagram for the variables can be obtained using R software, as shown in Fig. 5. The Pearson correlation coefficient ranges from −1 to 1; blue and red represent positive and negative correlations between two variables, respectively. The darker the color and larger the diameter of a circle are, the greater the absolute value of the Pearson correlation coefficient between two variables and stronger the correlation, respectively. As shown in Fig. 5, the correlations can be sorted in decreasing order as A_1, A_5, A_6, A_3, A_4, and A_2; these results are similar to the importance rankings.

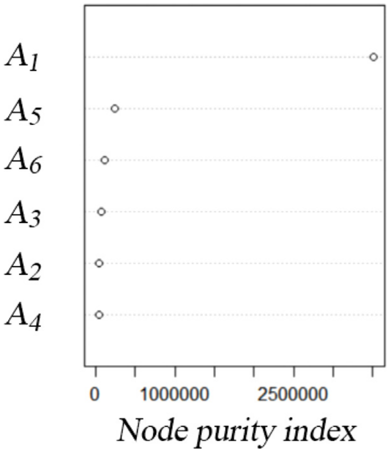

Fig. 4. Importance ranking of influential factors.

4.4. Model training and prediction

This paper randomly selects 43 subsets of data from the original data set to establish training samples for the RF model, and the remaining 11 subsets of data are used as test samples based on the design parameters of the 6 building envelope structures. The two parameters of the model are set as $mtry = 2$ and $ntree = 600$. The model is trained using the training samples. The regression fitting curve for the training sample set is shown in Fig. 6, and the trained RF model is used for regression fitting based on the test sample set. The prediction result of regression fitting based on the test sample set is shown in Fig. 7.

Fig. 6 shows that the fitting values of the RF model for the training sample set are similar to the true values, and the fitting effect is satisfactory. Fig. 7 shows that the predicted values of the RF model based on the test sample set are similar to the true values, indicating that the prediction error is small and the fitting accuracy is high.

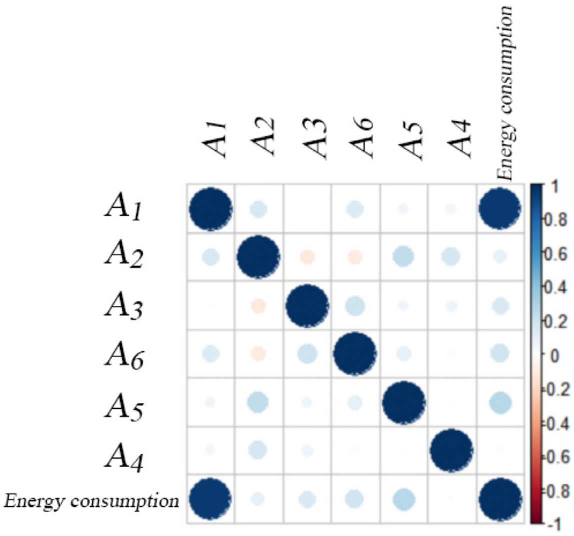

Fig. 5. Correlations among influential factors.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

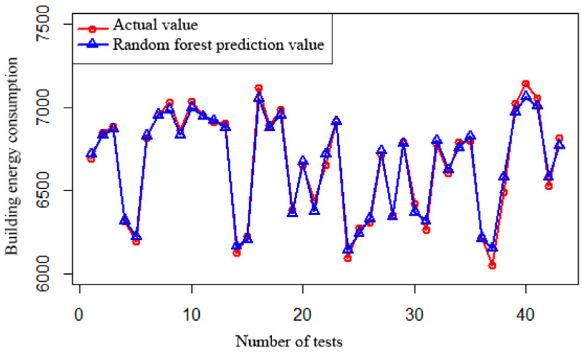

Fig. 6. Training sample fitting results.

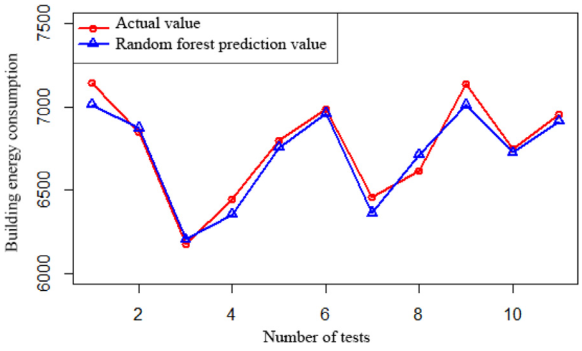

Fig. 7. Test sample fitting results.

4.5. Evaluation of the regression prediction results

To further verify the reliability of the RF model, SVM and BP-ANN models are considered to predict building energy consumption, and the key hyperparameters, namely, SVM, *C* and *gamma*, are adjusted. The prediction results are compared with the results of the RF model, and R^2 are adopted to assess the models. R^2 reflects the fit of a model and ranges from 0 to 1. An R^2 value close to 1 corresponds to an optimal model fit. An *RMSE* value close to zero corresponds to a statistically ideal fit between the predicted values and actual data. A comparison of the errors in the prediction results obtained using different models is shown in Table 4.

Table 4
Error comparison for different methods.

Performance index	*RMSE*	R^2
RF	0.015	0.993
SVM	0.426	0.929
BP-ANN	0.991	0.893

The *RMSE* for the prediction results of the RF model is considerably smaller than that for the SVM and BP-ANN models, which indicates that the regression prediction results of the RF model are closer to the actual values and the prediction accuracy of the proposed approach is higher than that of the other models (Yokoyama and Yamaguchi, 2020). Among the three methods, the RF model yields the R^2 value closest to 1, indicating that the model provides the highest fitting degree and the most satisfactory prediction effect. In summary, prediction results with high accuracy and reliability can be obtained by predicting building energy consumption based on the RF prediction model.

5. Conclusions

Rapid urbanization and growing global awareness regarding the environmental impact of buildings have motivated researchers to explore innovative energy-saving solutions. BIM provides a platform for related activities in the construction field and can help implement collaborative practices in design and construction. In this context, our research contributes to the prediction of energy consumption based on building structures from several perspectives. First, we develop an energy consumption simulation model involving BIM (Revit) and energy simulation software (DesignBuilder). The visual building model established using BIM software is imported into DesignBuilder to extensively analyze engineering scenarios and identify the appropriate strategies. Moreover, the comprehensive heat transfer coefficient and solar radiation absorption coefficient of the external walls, comprehensive heat transfer coefficient and solar radiation absorption coefficient of the roof, comprehensive heat transfer coefficient of the outer windows, and window-wall ratio are selected as input variables for DesignBuilder to perform numerical simulations. Next, to overcome the instability of the empirical learning algorithms and increase the prediction accuracy, this paper introduces RF ensemble learning prediction. Finally, the prediction results obtained using the RF are modified and optimized in BIM-DesignBuilder software, and architectural drawings in CAD format are directly output from the model.

Second, the RF technique can consider the interactions among the corresponding influential factors and rank the factors in order of importance, thereby enabling managers to implement targeted energy-saving measures. This research provides a novel approach for identifying the key factors that influence building energy prediction and achieving an enhanced understanding of building energy patterns. The variable importance analysis conducted in this study indicates that the most influential factors for the considered educational building are A_1, A_5 and A_6. In addition, after predicting building energy consumption based on the RF model, the predicted RMSE is 0.0115, and R^2 is 0.933, which reflect the high accuracy of the RF prediction model.

Finally, to further verify the reliability of the RF prediction model, the building energy consumption prediction results of the RF model, an SVM model and a BP-ANN model are compared. The results show that in comparison with the SVM and BP-ANN models, the RF model yields the highest goodness of fit and smallest root mean square error. Thus, the proposed method can aid architects and decision makers in assessing preliminary building performance by considering the key factors in heavily obstructed environments and for various facade configurations without performing exhaustive analyses. The results obtained in this study can assist researchers and building professionals in understanding and managing the energy optimization design and usage scenarios of public buildings considering cost optimality and energy efficiency.

6. Limitations and future work

Importing Revit data into DesignBuilder involves certain challenges, including changes in the room volume based on the different drawing formats used in Revit and DesignBuilder, and overcoming the issues related to the exported model will be a focus of future work. In addition, the predictions of the absolute energy values in an energy simulation, given the relevant assumptions, are rarely accurate. Usually, various validation tests, such as the Building Energy Simulation Test (BESTEST) developed by the International Energy Agency (IEA), are implemented to validate building energy simulation tools (IEA 2007). In addition, validation tests comparing actual measurements from test buildings can be performed.

This paper considers only a single case study in northern China to predict building energy consumption. However, China has a vast territory that spans five climatic zones, and the energy consumption demands in different climate zones vary. In addition, different types of public buildings have different energy consumption characteristics. Therefore, the energy-saving potential of the envelope structure design parameters in different climate zones and different methodologies for different building types should be considered. These factors can be considered in future research. Additionally, future work can focus on enhancing model prediction performance and expanding the proposed RF model to obtain energy predictions for other types of public buildings, such as hospitals and commercial buildings.

CRediT authorship contribution statement

Yang Liu: Conceptualization, Methodology, Software. **Hongyu Chen:** Data collection, Writing – original draft. **Limao Zhang:** Writing – review & editing. **Zongbao Feng:** Writing – review & editing.

Declaration of competing interest

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

Funding

This work was supported by the Zhongnan Hospital of Wuhan University Science, Technology and Innovation Seed Fund, Project CXPY2020013, the Construction of Science and Technology Plan Project of Hubei Province (Grant No. 202041), and the National Natural Science Foundation of China (Grant No. 72031009).

References

Abdelkader, O.A., 2020. Impact of perception on willingness and behavior of individuals toward switching to sustainable energy practices in buildings. Energy Rep. 6, 2119–2125.

Al-janabi, A., Kavgic, M., Mohammadzadeh, A., Azzouz, A., 2019a. Comparison of energyPlus and IES to model a complex university building using three scenarios: Free-floating, ideal air load system, and detailed. J. Build. Eng. 22 (2019), 262–280.

Al-janabi, A., Kavgic, M., Mohammadzadeh, A., Azzouz, A., 2019b. Comparison of energyPlus and IES to model a complex university building using three scenarios: Free-floating, ideal air load system, and detailed. J. Build. Eng. 22, 262–280.

An, J., Yan, D., Guo, S., Gao, Y., Peng, J., Hong, T., 2020. An improved method for direct incident solar radiation calculation from hourly solar insolation data in building energy simulation. Energy Build. 227, 110425.

Andolsun, S., Culp, C.H., Haberl, J.S., Witte, M.J., DOE, EnergyPlus.vs., 2012. EnergyPlus vs DOE- 2.1e: The effect of ground coupling on cooling/heating energy requirements of slab-on-grade code houses in four climates of the US. Energy Build. 52, 189–206.

Anwar, M., Rasul, M.G., Khan, M.M.K., 2020. Performance analysis of rooftop greenery systems in Australian subtropical climate. Energy Rep. 6, 50–56.

Aranda, A., Ferreira, G., Mainar-Toledo, M.D., Scarpellini, S., Llera Sastresa, E., 2012. Multiple regression models to predict the annual energy consumption in the spanish banking sector. Energy Build. 49, 380–387.

Ascione, F., Bianco, N., Iovane, T., Mastellone, M., Mauro, G.M., 2020. Is it fundamental to model the inter-building effect for reliable building energy simulations? Interaction with shading systems. Build. Environ. 183, 107161.

Ascione, F., Bianco, N., Maria Mauro, G., Napolitano, D.F., 2019. Building envelope design: Multi-objective optimization to minimize energy consumption, global cost and thermal discomfort. Application to different Italian climatic zones. Energy 174, 359–374.

Aslani, A., Bakhtiar, A., Akbarzadeh, M.H., 2019. Energy-efficiency technologies in the building envelope: Life cycle and adaptation assessment. J. Build. Eng. 21, 55–63.

Ayoub, M., 2019. A multivariate regression to predict daylighting and energy consumption of residential buildings within hybrid settlements in hot-desert climates. Indoor Built Environ. 28 (6), 848–866.

Belussi, L., Barozzi, B., Bellazzi, A., Danza, L., Devitofrancesco, A., Fanciulli, C., Ghellere, M., Guazzi, G., Meroni, I., Salamone, F., Scamoni, F., Scrosati, C., 2019. A review of performance of zero energy buildings and energy efficiency solutions. J. Build. Eng. 25, 100772.

Braun, M.R., Altan, H., Beck, S.B.M., 2014. Using regression analysis to predict the future energy consumption of a supermarket in the UK. Appl. Energy 130, 305–313.

CABEE, 2018. In: Cai, W.G. (Ed.), Chinese Building Energy Consumption Report (2018). Beijing, PR China. Available at: https://mp.weixin.qq.com/s/UtefuqsikpZ6arqwQjtlGA(in Chinese).

Cao, L., Li, Y., Zhang, J., Jiang, Y., Han, Y., Wei, J., 2020. Electrical load prediction of healthcare buildings through single and ensemble learning. Energy Rep. 6, 2751–2767.

Ceballos-Fuentealba, I., Alvarez-Miranda, E., Torres-Fuchslocher, C., Luisa del Campo-Hitschfeld, M., Diaz-Guerrero, J., 2019. A simulation and optimisation methodology for choosing energy efficiency measures in non-residential buildings. Appl. Energy 256, 113953.

Chen, B., Liu, Q., Chen, H.Y., Wang, L., Deng, T.T., Zhang, L.M., Wu, X.G., 2021. Multiobjective optimization of building energy consumption based on BIM-DB and LSSVM-NSGA-II *. J. Cleaner Prod. 294, 126153.

Cheng, B.Q., Li, J.W., Tam, V.W.Y., Yang, M., Chen, D., 2020. A BIM-LCA approach for estimating the greenhouse gas emissions of large-scale public buildings: A case study. Sustainability 12 (2), 685.

Elbeltagi, E., Wefki, H., Abdrabou, S., Dawood, M., Ramzy, A., 2017. Visualized strategy for predicting buildings energy consumption during early design stage using parametric analysis. J. Build. Eng. 13, 127–136.

Etxebarria-Mallea, M., Oregi, X., Grijalba, O., Hernandez-Minguillon, R., 2021. The impact of energy refurbishment interventions on annual energy demand, indoor thermal behaviour and temperature-related health risk. Energy Policy 153, 112276.

Fan, Y.L., Xia, X.H., 2017. A multi-objective optimization model for energy-efficiency building envelope retrofitting plan with rooftop PV system installation and maintenance. Appl. Energy 189, 327–335.

Fantucci, S., Serra, V., Carbonaro, C., 2020. An experimental sensitivity analysis on the summer thermal performance of an opaque ventilated faade. Energy Build. 225, 110354.

Gallagher, C.V., Leahy, K., O'Donovan, P., Bruton, K., O'Sullivan, D.T.J., 2019. IntelliMaV: A cloud computing measurement and verification 2.0 application for automated, near real-time energy savings quantification and performance deviation detection. Energy Build. 185, 26–38.

Gatt, D., Yousif, C., Cellura, M., Camilleri, L., Guarino, F., 2020. Assessment of building energy modelling studies to meet the requirements of the new Energy Performance of Buildings Directive. Renewable Sustain. Energy Rev. 127, 109886.

Goncalves, J.E., van Hooff, T., Saelens, D., 2020. Understanding the behaviour of naturally-ventilated BIPV modules: A sensitivity analysis. Renew. Energy 161, 133–148.

Guo, X.F., Zhao, Q.N., Zheng, D., Ning, Y., Gao, Y., 2020. A short-term load forecasting model of multi-scale CNN-LSTM hybrid neural network considering the real-time electricity price. Energy Rep. 6, 1046–1053.

He, B.J., 2019. Towards the next generation of green building for urban heat island mitigation: Zero UHI impact building. Sustainable Cities Soc. 50, 101647.

He, B.J., Yang, L., Ye, M., Mou, B., Zhou, Y.A., 2014. Overview of rural building energy efficiency in China. Energy Policy 69, 385–396.

Heydari, A., Sadati, S.E., Gharib, M.R., 2021. Effects of different window configurations on energy consumption in building: Optimization and economic analysis. J. Build. Eng. 35, 102099.

Huang, K.T., Li, Y.J., 2017. Impact of street canyon typology on building's peak cooling energy demand: A parametric analysis using orthogonal experiment. Energy Build. 154 (2017), 448–464.

Kim, W., Han, Y., Kim, K.J., Song, K.W., 2020. Electricity load forecasting using advanced feature selection and optimal deep learning model for the variable refrigerant flow systems. Energy Rep. 6, 2604–2618.

Lei, L., Chen, W., Wu, B., Chen, C., Liu, W., 2021. A building energy consumption prediction model based on rough set theory and deep learning algorithms. Energy Build. 240, 110886.

Li, D., Hu, G., Spanos, C.J., 2016. A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis. Energy Build. 128 (2016), 519–529.

Li, D., Zheng, Y.M., Liu, C.Y., Wu, G.Z., 2015. Numerical analysis on thermal performance of roof contained PCM of a single residential building. Energy Convers. Manage. 100, 147–156.

Liu, Y., Chen, H., Xian, j., 2020a. Research on green renovations of existing public buildings based on a cloud model –TOPSIS method. J. Build. Eng. 101930.

Liu, Y., Chen, H., Zhang, L., Wu, X., Wang, X.-j., 2020. Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China. J. Cleaner Prod. 272 (2020b), 122542.

Liu, Y., Wang, X.J., Zhou, S., Chen, H.Y., 2021. Enhancing public building energy efficiency using the response surface method: An optimal design approach. Environ. Impact Assess. Rev. 87, 106548.

Liu, Y., Zou, S.Q., Chen, H.Y., Wu, X.G., Chen, 2019. Simulation analysis and scheme optimization of energy consumption in public buildings. Adv. Civil Eng. 2019, 6326138.

Lu, H.F., Cheng, F.F., Ma, X., Hu, G., 2020. Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower. Energy 203, 117756.

Marino, C., Nucara, A., Pietrafesa, M., 2017. Does window-to-wall ratio have a significant effect on the energy consumption of buildings? A parametric analysis in Italian climate conditions. J. Build. Eng. 13, 169–183.

Nasir, M.H.A., Hassan, A.S., 2020. Thermal performance of double brick wall construction on the building envelope of high-rise hotel in Malaysia. J. Build. Eng. 31, 101389.

Pan, Y., Zhang, L.M., 2020. Data-driven estimation of building energy consumption with multi-source heterogeneous data. Appl. Energy 268, 114965.

Park, B.R., Choi, E.J., Hong, J., Lee, J.H., Moon, J.W., 2018. Development of an energy cost prediction model for a VRF heating system. Appl. Therm. Eng. 140, 476–486.

Pedersen, T.H., Hedegaard, R.E., Petersen, S., 2017. Space heating demand response potential of retrofitted residential apartment blocks. Energy Build. 141, 158–166.

Phillips, R., Troup, L., Fannon, D., Eckelman, M.J., 2020. Triple bottom line sustainability assessment of window-to-wall ratio in US office buildings. Build. Environ. 182, 107057.

Ryan, E.M., Sanquist, T.F., 2012. Validation of building energy modeling tools under idealized and realistic conditions. Energy Build. 47, 375–382.

Solla, M., Ismail, L.H., Shaarani, A.S.M., Milad, A., 2019. Measuring the feasibility of using of BIM application to facilitate GBI assessment process. J. Build. Eng. 25, 100821.

Stephan, J., Stegle, O., Beyer, A., 2015. A random forest approach to capture genetic effects in the presence of population structure. Nature Commun. 6 (7432).

Tian, Z.C., Wei, S., Shi, X., 2020a. Developing data-driven models for energy-efficient heating design in office buildings. J. Build. Eng. 32, 101778.

Tian, Y., Yu, J., Zhao, A., 2020b. Predictive model of energy consumption for office building by using improved GWO-BP. Energy Rep. 6, 620–627.

Tian, Y., Yu, J.Q., Zhao, A.J., 2020c. Predictive model of energy consumption for office building by using improved GWO-BP. Energy Rep. 6, 620–627.

Wan, H., Xu, X.H., Gao, J.J., Yan, T., 2017. Development of a wet-bulb temperature-based heat balance control method for a hybrid ground source heat pump system. In: Cui, P., Liu, J., Zhang, W. (Eds.), 10th International Symposium on Heating, Ventilation and Air Conditioning, Ishvac2017. pp. 3251–3258.

Wang, Z.X., Ding, Y., Geng, G., Zhu, N., 2014. Analysis of energy efficiency retrofit schemes for heating, ventilating and air-conditioning systems in existing office buildings based on the modified bin method. Energy Convers. Manage. 77, 233–242.

Wang, Z.Y., Wang, Y.R., Zeng, R.C., Srinivasan, R.S., Ahrentzen, S., 2018. Random forest based hourly building energy prediction. Energy Build. 171, 11–25.

Yokoyama, A., Yamaguchi, N., 2020. Comparison between ANN and random forest for leakage current alarm prediction. Energy Rep. 6, 150–157.

Zhang, C.B., Li, J.Y., Zhao, Y., Li, T.T., Chen, Q., Zhang, X.J., 2020. A hybrid deep learning-based method for short-term building energy load prediction combined with an interpretation process. Energy Build. 225, 110301.

Zhang, L., Wu, X., Ding, L., Skibniewski, M.J., Lu, Y., 2016. Bim-based risk identification system in tunnel construction. J. Civil Eng. Manage. 22, 529–539.

Zhang, X.L., Xia, J.J., Jiang, Z.Y., Huang, J.Y., Qin, R., Zhang, Y., Liu, Y., Jiang, Y., 2008. DeST-An integrated building simulation toolkit part.: Applications. Build. Simul. 1 (3), 193–209.

Zhao, H.-x., Magoulès, F., 2012. A review on the prediction of building energy consumption. Renew. Sustain. Energy Rev. 16 (6), 3586–3592.

Zhou, J., Wan, I., Wang, Y., Roskilly, A., 2019. Investigating the impact of building's facade on the building's energy performance – a case study. Energy Procedia 158, 3144–3151.