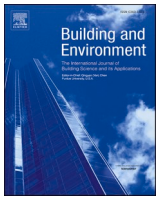
Building and Environment 218 (2022) 109081

Contents lists available at ScienceDirect

Building and Environment

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



Optimization and prediction in the early design stage of office buildings using genetic and XGBoost algorithms

Hainan Yan [a], Ke Yan [b], Guohua Ji [a,*]

[a] School of Architecture and Urban Planning, Nanjing University, Nanjing, 210093, China
[b] Department of the Built Environment, National University of Singapore, Singapore, 117566, Singapore

ARTICLE INFO

Keywords:
Office buildings
Early design stage
Building performance
XGBoost algorithm

ABSTRACT

Incorporating intelligent optimization algorithms in the early stages of office building design facilitates a better response to the local climate. The indoor and outdoor thermal performances of office buildings, such as solar radiation, indoor lighting, and outdoor thermal comfort, must be jointly evaluated during the conceptual design phase. Based on the technical framework of "performance-based generative architectural design", this study constructs a data-driven workflow for comprehensive performance assessment and rapid prediction of office buildings. The method was then applied to an office building in the hot summer and cold winter regions of China. Based on a total of 6000 data samples generated by the iterative process of genetic optimization, this study achieved a precision of 0.77, recall of 0.59, and F-1 score of 0.75 for categorical prediction by the XGBoost algorithm. The method facilitates the optimization potential of integrated solar and thermal performances in the early design phase of office buildings while significantly improving the efficiency of interaction and feedback between design decisions and their performance evaluation.

1. Introduction

Studies on building energy consumption in China in recent years show that the total building area in the country reached 63.487 billion square meters as of 2016, of which the public building area was approximately 11.506 billion square meters, accounting for 18.12% [1]. Public buildings are the most energy intensive per unit area and have maintained an upward trend in energy usage [2]. The evaluation, modeling, and prediction of the energy consumption of public and office buildings have significant potential for energy conservation [3]. To achieve the goals of energy saving and carbon emission reduction in office buildings, an ecological performance optimization-oriented approach to office building design is of significant relevance [4,5].

Building performance optimization and prediction studies are critical in the early design phase [6]. Because of the time cost of the calculation process and the highly specialized parameter settings, the previous environmental performance evaluation was often placed in the late design stage and left to professionals, that is, the post-evaluation paradigm [7]. Its environmental performance analysis and optimization process are gradually becoming detached from architectural design, even formally. Design decisions made early in the building design process are more efficient and affect the environmental performance of the design at a lower cost than those made later [8]. In a tight design cycle, architects often want quick feedback on design solutions. In recent years, a generative design (GD) approach based on performance simulations has emerged [9,10].

Over the last decade or more, performance simulation and algorithmic optimization to automate performance feedback have become common design tools [11]. An Italian study conducted by D'Agostino et al. presented a computational performance-driven early stage design optimization workflow using an educational building as a case study [12]. The procedure integrates parametric algorithmic modeling tools with a genetic-algorithm-driven optimization process that minimizes energy consumption, building costs, and natural illumination. Using a combination of self-organizing map clustering and the NSGA-II optimization algorithm, Dong et al. proposed a novel decision-making method that has been experimentally validated on a typical building in a severely cold region of China [13]. The intelligent optimization method proposed in this study is superior to the two commonly used methods of subjective decision design and simulation-aided design because it automates the multi-objective optimization process for lighting and energy efficiency, while improving the efficiency and accuracy of the design.

* Corresponding author. School of Architecture and Urban Planning, Nanjing University, Jianliang Building, Gulou Campus, Nanjing, 210093, China.
E-mail address: jgh@nju.edu.cn (G. Ji).

https://doi.org/10.1016/j.buildenv.2022.109081
Received 6 February 2022; Received in revised form 23 March 2022; Accepted 7 April 2022
Available online 14 April 2022

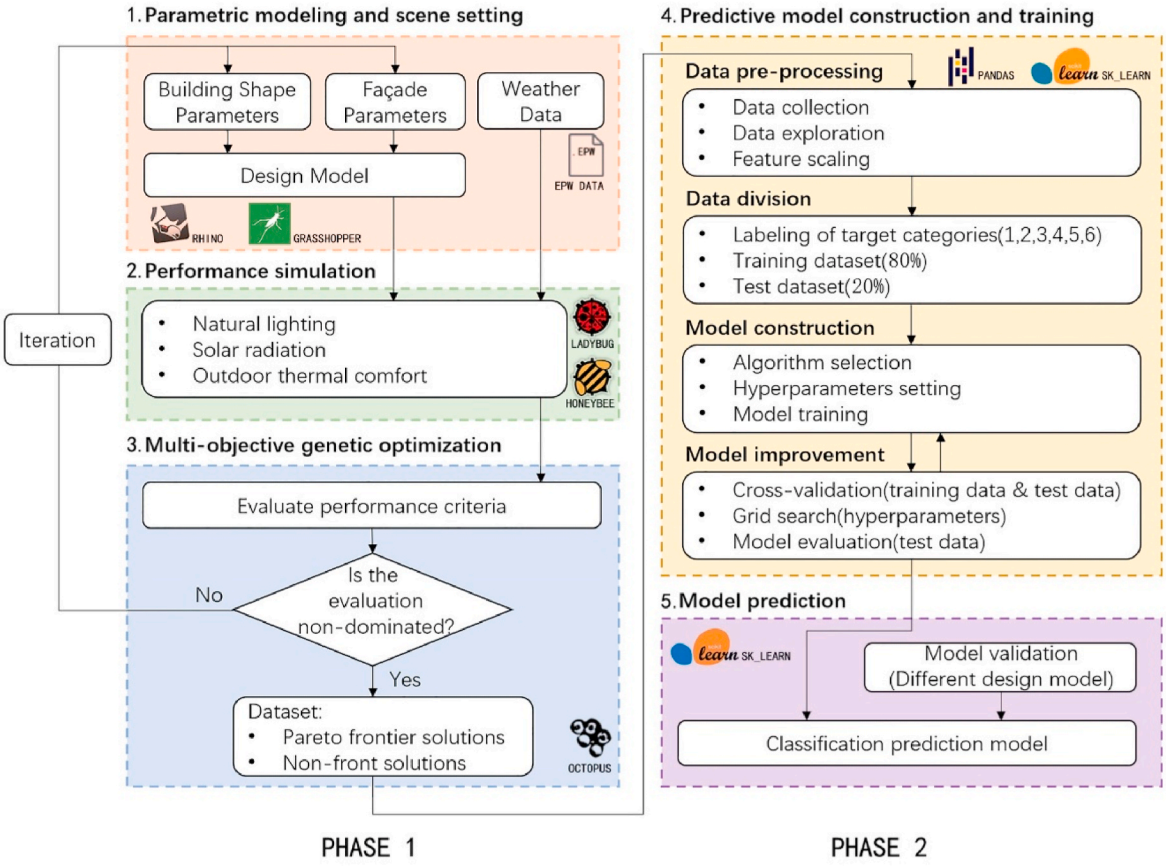


Fig. 1. Overview of the workflow.

Lin et al. [14] proposed a systematic approach to three broad issues in the implementation of building performance optimization in the early design stages: model integration, performance analysis in real-time, and interactive design. This led to the development of MOOSAS, a multi-objective building-performance optimization software. In a review article by Tian et al., building energy simulation and optimization (BESO) was examined for its application in passive building design [15]. Studies have specifically explored passive design strategies, including opaque envelopes, fenestrations, shadings, and natural ventilation. The findings suggest that BESO is an effective way to help designers explore various design areas.

It should be noted that the design framework constructed based on performance simulation and algorithm optimization still has significant shortcomings. First, the arithmetic and time costs of the method are high, particularly when an exhaustive search of the design solutions is applied [16]. Second, the method places high demands on architects to use software. Because current building performance simulation software, such as Radiance and Daysim, requires detailed parameter settings and a thorough understanding of building physics, it is difficult for architects to produce realistic optimization results by relying entirely on manual optimization [17]. Architects are concerned about the environmental performance of their building designs, however, they are not necessarily familiar with the building performance simulation and optimization methods needed to achieve higher building performance. Finally, the performance simulation and algorithm optimization collaboration constitute a complete closed-loop workflow that is not relatively flexible. In this closed-loop workflow, the architect has a limited scope for the subjective choice and adjustment of design solutions. If the performance results of the design solution are obtained after adjustments have been made by the architect, the design variables are unknown [18].

The use of machine-learning models to build fast prediction models has been proposed in recent years to fill these gaps [19–21]. Lin and Tsay trained a daylighting model based on artificial neural networks, which predicted the daylighting performance of different types of building facades [22]. Yan et al. introduced a collaborative deep learning framework using IoT data for building energy-consumption models and forecasting [23]. Mo et al. predicted the window behavior of occupants based on the XGBoost algorithm and demonstrated that the algorithm has significant advantages [24]. Han et al. developed a method for measuring the annual daylighting performance using artificial neural networks at an early design stage [25]. Building performance prediction models require a comprehensive focus on (1) the efficiency of prediction, (2) the accuracy of the prediction, and (3) the generalizability of the prediction model.

However, most current studies focus only on predicting and optimizing building performance indoors or outdoors, or on optimizing performance over a short period. Few studies have integrated indoor and outdoor building performance simulations and have focused on performance simulation and optimization over a year-round period. Second, studies on performance prediction often take a metamodel or a single room as the object of study, which excessively pursues prediction accuracy and loses room for error tolerance, which is not very useful for applying design practice. In other words, it is difficult to integrate the established prediction framework with actual building practices.

Given these shortcomings, this study constructs a prediction method for office building performance based on the XGBoost algorithm for the joint evaluation of the indoor and outdoor performance of office buildings at an early stage of the design process. In Nanjing, China, an office building was the experimental object. After obtaining the initial dataset through parametric performance simulation and multi-objective genetic optimization (MOGO), the performance labels can be rapidly predicted and evaluated through data preprocessing, model selection, and hyperparameter optimization. In this study, we present the

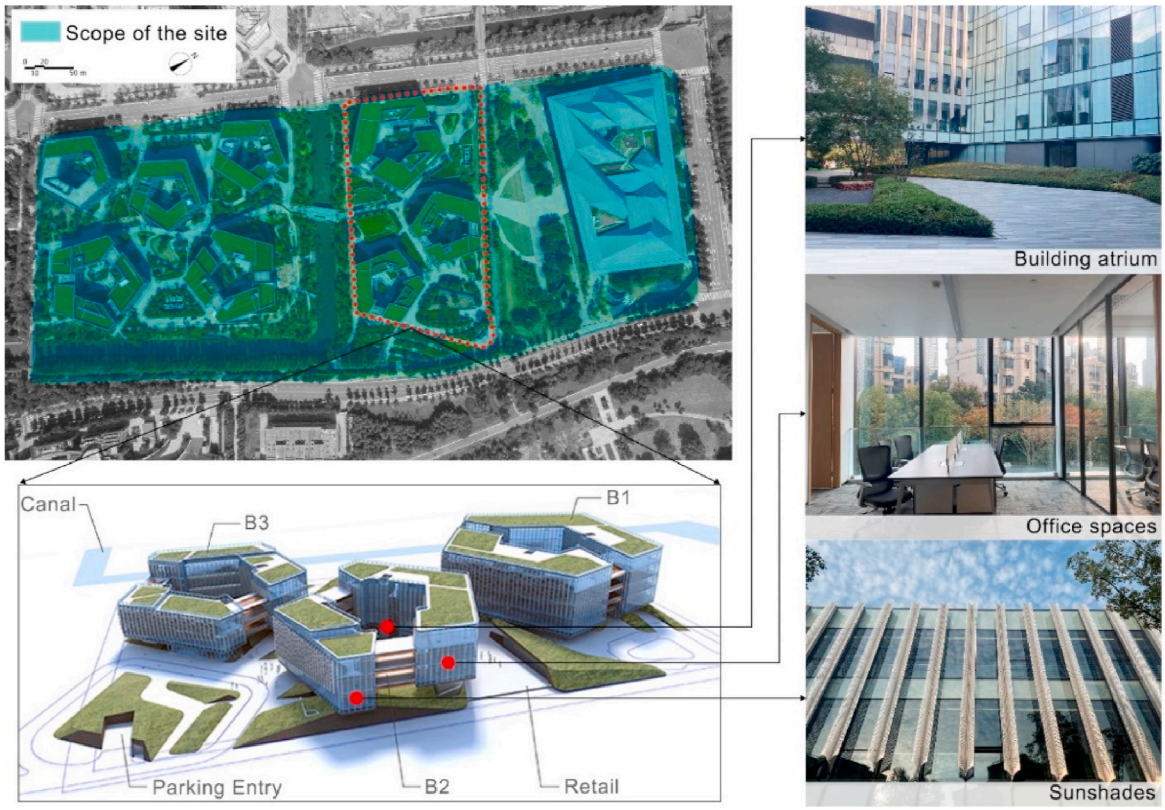


Fig. 2. Case site and detailed information on the office building.

following aspects as the main novelty of this study. First, the indoor and outdoor light and heat performances of office buildings were jointly evaluated and predicted. Second, it presents a detailed comparison of the prediction performance of the XGBoost algorithm with that of several other commonly used algorithms. Overall, a performance-based generative architectural design (PGAD) framework incorporating machine-learning algorithms was developed.

2. Methods

2.1. Overview workflow

Fig. 1 illustrates the proposed methodology, which has two phases. The main objective of the first phase can be summarized as the data generation for the different design solutions. The building shape and facade were initially generated parametrically, setting local climatic conditions and site surroundings. The study then moved on to the performance simulation section and considered three aspects of natural lighting, solar radiation, and outdoor thermal comfort to evaluate office buildings' indoor and outdoor performance. Subsequently, a multi-objective genetic optimization (MOGO) was used to develop the solution iteratively. The results of the MOGO method, including the Pareto frontier solutions and non-front solutions, were collected and aggregated to serve as raw data for the second phase. The second phase involved developing, training, and validating a categorical prediction model using the XGBoost algorithm. The validated and optimized XGBoost algorithm was used to predict the performance of office buildings at the early stage of development. The workflow is implemented through packages and plugins (Fig. 1), further described in subsequent sections. A personal computer (Intel Core i7 2.50 GHz and 16 GB of RAM) was used to conduct development and experiments.

Table 1
Building parameters.

Input	Range	Step	Unit	Explanation
RD	11.7–14.3	0.01	m	Room depth
GFH	4.0–6.0	0.01	m	Ground floor height (open floor)
SFH	4.0–5.0	0.01	m	Standard floor height
SS	0.5–1.5	0.01	m	Sunshade spacing
SW	0.5–1.5	0.01	m	Sunshade width
Orientation	0–359	1	degree	The angle formed with the original building orientation (north)

2.2. Case study and data generation

An office building within the Eco-Tech Island project located in Nanjing, China (Latitude: N32°01′, Longitude: E118°42′), was selected as the study case, as shown in Fig. 2. Designed by the NBBJ team and completed in 2020, the Nanjing Eco-Tech Island project covers 134,000 m^2 and comprises office buildings, residential buildings, and an exhibition hall [26]. This study focused on the office building, which has several proprietary energy-saving design strategies: (1) the pentagon-shaped floor plate in the office building provides maximum natural light penetration and maximizes leasable space; (2) the architects elevated the ground floor space to promote natural ventilation in the atrium, thereby improving outdoor comfort; and (3) with the custom vertical louvres, passive shading can be maximized while keeping views clear and open. The amount of heat dissipated in every office building was reduced to minimize the energy load through each of these strategies.

2.2.1. Parametric modeling and scene-setting

The office building selected for this study was located on the northern side of the Nanjing Eco-Tech Island project. It was surrounded by an exhibition hall and two other office buildings. Based on the

Table 2
Weather data for the Nanjing area on an annual basis.

Meteorological parameters	Values
Dry-bulb temperature	15.59 °C
Dew point temperature	10.92 °C
Relative humidity	75.67%
Direct normal radiation	84.58 Wh/m^2
Diffuse horizontal radiation	111.07 Wh/m^2
Global horizontal radiation	167.57 Wh/m^2
Horizontal infrared radiation	350.31 Wh/m^2
Wind speed	2.49 m/s
Wind direction	125.11°
Barometric pressure	101589.41 Pa

architect's design strategy and Chinese national office building design codes [27], parameters in terms of building orientation, room depth (RD), ground floor height (GFH) (open floor), standard floor height (SFH), and sunshade were set and maintained within a reasonable range of variation, which in turn controlled and generated a parametric office building model (Table 1). In this study, Grasshopper, a node-based programming plugin based on the 3D modeling software, Rhino, was used to perform parametric modeling [28]. A new parametric building model was generated when the design parameters were adjusted or updated. Parametric models can also produce quantified building shapes for simulations and computations.

The scene-setting, such as surrounding buildings and meteorological data, was also the input for subsequent simulations based on the Rhino-Grasshopper platform. To balance the integrity of the site environment with the simulation speed, two office buildings around the target office building were modeled in this study without considering other aspects such as trees and artificial heat sources. In addition, the climate in Nanjing is classified as 3A according to the ASHRAE Standard 90, 1–2020. The annual average meteorological data for the Nanjing climate region are presented in Table 2.

2.2.2. Performance simulation

As discussed in the previous sections, this study focused on indoor and outdoor light comfort and thermal comfort in office buildings. Three specific performance indicators were selected: daylight factor (DF), total solar radiation (TSR), and universal thermal climate index (UTCI). As DF is included in the Chinese building codes, it is used for indoor daylight performance [29]. Considering that sunlight positively improves comfort, health, and energy performance, this study calculated the TSR received by office building surfaces. This study measured outdoor thermal performance using the UTCI, a metric that incorporates temperature, humidity, solar radiation, and wind speed in one integrated measurement [30]. These measurements are discussed in detail in subsequent sections.

The open-source tools Ladybug and Honeybee, available in Grasshopper, were used to simulate daylighting, solar radiation, and outdoor comfort performance. Ladybug and Honeybee are proven simulation tools and have been widely used [31,32]. In this study, the Ladybug1.2.0 and Honeybee1.2.0 versions of these open-source tools were used.

Under overcast skies, the ratio of indoor illumination to outdoor illumination was measured using the DF [33]. In the early stages of design, the DF value can be easily calculated, helping architects and developers to quickly make informed decisions. While several dynamic daylight measurement methods have emerged in recent years, such as useful daylight illumination (UDI) and daylight autonomy (DA), this study selected DF primarily for its simplicity [34]. According to the Chinese building code, DF should be greater than 3% annually [35]. Experiments have also revealed that the productivity of office workers is more stable when the DF exceeds 3% in office buildings. Nevertheless, DF exceeding 10% can glare and reduce visual comfort [36]. Therefore, DF should be a two-tailed indicator with reasonable lower and upper bounds [37].

Table 3
Boundary condition settings for office buildings.

Boundary conditions	Reflectance
Wall	0.60
Floor	0.30
Ceiling	0.75
Window	0.80
Sunshade	0.84

The reflectance settings for the office building boundary conditions are listed in Table 3 and remained constant throughout the workflow. This study selected the same or similar building materials as the existing buildings in the performance simulation process. The experimental errors owing to material factors were also ignored. The surface was measured 0.75 m above the floor level, with a measuring grid size of 1 × 1 m. DF was calculated using Honeybee's DF calculation module.

Solar radiation is an essential factor in building design. In China, designers often obtain solar design strategies through simple formula calculations, even with the sole purpose of meeting the design specifications. This simple and rigid design approach is inaccurate and ineffective. Particularly in regions with hot summers and cold winters, designers need to consider a combination of reducing harmful radiation in summer and increasing beneficial radiation in winter [38]. The TSR received by a building during the year was measured by subtracting the TSR during summer from the TSR during winter. The equation is as follows:

$$TSR = TSR_{summer} - TSR_{winter} \quad (1)$$

TSR_{summer} indicates the TSR received by the building from June to August and TSR_{winter} indicates the TSR received from December to February. The Ladybug solar radiation calculation module was used to calculate TSR. The measurement grid was set to 1 × 1 m to ensure the accuracy of the results, and the other parameters were preset.

The UTCI describes an external thermal environment [39]. Studies have found that UTCI exhibits a higher level of reliability than other thermal comfort indices [40,41]. This study evaluated the outdoor comfort performance of office buildings during winter and summer using the UTCI.

UTCI is calculated using probing points 1.5 m above pedestrian level, with a measurement grid size of 2 × 2 m. According to the Nanjing EnergyPlus weather file (.epw), the summer measurement was noon of August 8, which is an extremely hot week. The winter measurement time was noon of January 18, which is an extremely cold week. Ladybug's outdoor thermal comfort module calculates the UTCI values. In this study, the average number of spatial grid results was used as the optimization objective to simplify the genetic optimization process. It should be noted that after the UTCI values of each grid are averaged, there were some limitations in reflecting the spatial distribution characteristics problem of thermal comfort performance.

2.2.3. MOGO

The genetic algorithms (GAs) were developed based on Darwin's theory of natural selection [42]. GAs, particularly MOGO, have proven suitable for optimizing building energy. In recent years, MOGO has been extensively used to optimize building forms, envelope HVAC, and renewable energy systems [43,44]. MOGO has the following main advantages: (1) it is a global optimization algorithm; (2) it can find multiple Pareto solutions for multi-objective optimization problems at once; and (3) the objective function of GAs need not be continuous. Select, crossover, and variation are the initial operators of MOGO. The benefits of this algorithm are evident in the case of several optimization parameters.

The Octopus is a component of Grasshopper, which was used to perform the optimization in this study [45]. It allows multi-objective optimization of competing objectives within a seamless workflow and

Table 4
The genetic algorithm settings.

Boundary conditions	Values
Elitism	0.5
Mutation Probability	0.2
Mutation Rate	0.9
Crossover Rate	0.8
Population Size	60
Max Generation	100

allows the entire optimization process to be visualized and controlled. Based on the relevant literature and the experience of several experiments [46,47], the GA settings are listed in Table 4.

2.3. Model training

This section provides a detailed workflow of a machine-learning algorithm for constructing a comprehensive performance rapid prediction and evaluation model for office buildings. It includes data acquisition and preprocessing, model selection, hyperparameter tuning, and model evaluation. This part of the study was developed based on the Scikit-learn library for the Python platform, which is a widely used machine-learning algorithm tool [48]. Each step is explained in detail in the following sections.

2.3.1. Data pre-processing

The initial dataset was obtained when the Octopus platform completed the optimization operation. Pareto-front solutions are the optimal solutions generated during the optimization process. The non-front solutions denote the set of successive solutions and are not on the Pareto front. As shown in Fig. 3 and Table 5, the data exported from the Octopus platform were classified into six categories according to the optimization operations and design requirements. Each class of solution carries a specific label and descriptive information. Relatively good solutions are all solution sets on the Pareto front, where the Pareto front solutions of the last few generations are ideal design solutions. The solution sets on the Pareto front that can satisfy the requirement of $3\% \leq DF \leq 10\%$ are labelled as A1, B2, C3, and the all-around performance decreases in order. The solution label set on the Pareto front does not meet the requirement of the lighting factor ($3\% \leq DF \leq 10\%$) D4, and the design parameters need to be adjusted to improve the lighting factor design solution. The solution sets not on the Pareto front were classified as E5 and F6, respectively. These design solutions have poor overall performance and require adjustment of the design parameters. Classification labels and descriptive information are set such that the machine learning model can learn the optimization and selection process of architects oriented towards comprehensive building performance improvement.

Using a min-max normalization function, all input values were normalized from 0 to 1 along the feature axis after the labels were set up on the dataset.

2.3.2. Algorithm selection and model setting

In the following step, the type and structure of the categorical prediction algorithm are defined as the result of obtaining the initial data and preprocessing it. The goal of the machine learning model constructed in this study was to predict and feed back the performance level of a new combination of architectural design solution parameters by learning from a training set of existing parameter sets with their corresponding performance. Based on the analysis and comparison of various machine-learning models [49,50], the XGBoost algorithm was used to perform the categorical prediction in this study.

The XGBoost algorithm was proposed in 2016 [51]. It has been applied in a variety of fields, such as the prediction of building cooling loads [52], prediction of window behavior in residential buildings [24], and automotive manufacturing [53]. Studies have shown that XGBoost is more accurate, stable, and efficient than conventional machine-learning methods [54,55].

XGBoost is a widely used boosting algorithm recognized for its efficiency and flexibility. The algorithm is used for supervised learning problems, where we use training data x to predict the target label y. XGBoost is equipped with an objective function that comprises two parts, training loss and regularization term.

$$obj(f) = \sum_{i=1}^{n} L(y_i, \quad \hat{y}_i) + \Omega(f) \tag{2}$$

Table 5
Description of the dataset label.

Label	Explanation
A1	The ideal solution with good building energy performance
B2	Sub-optimal solution with good energy-saving performance
C3	The feasible solution, but inferior energy savings
D4	Energy-saving performance is available, but the lighting factors need to be adjusted to the design
E5	It has energy-saving performance, but poor solar radiation and outdoor thermal comfort performance
F6	The combined energy savings are poor and not suitable as an alternative

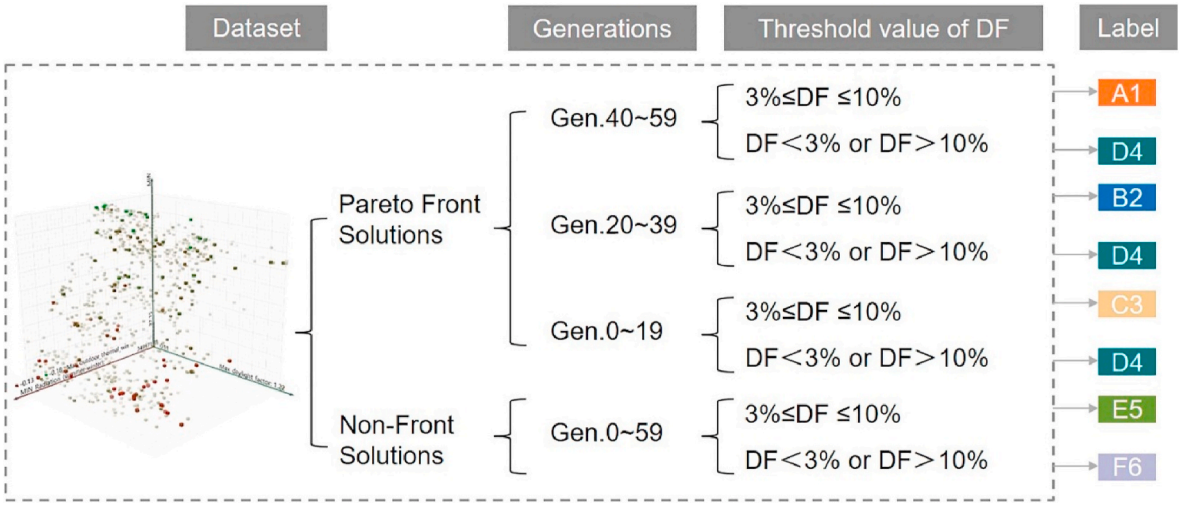


Fig. 3. Label settings for the dataset.

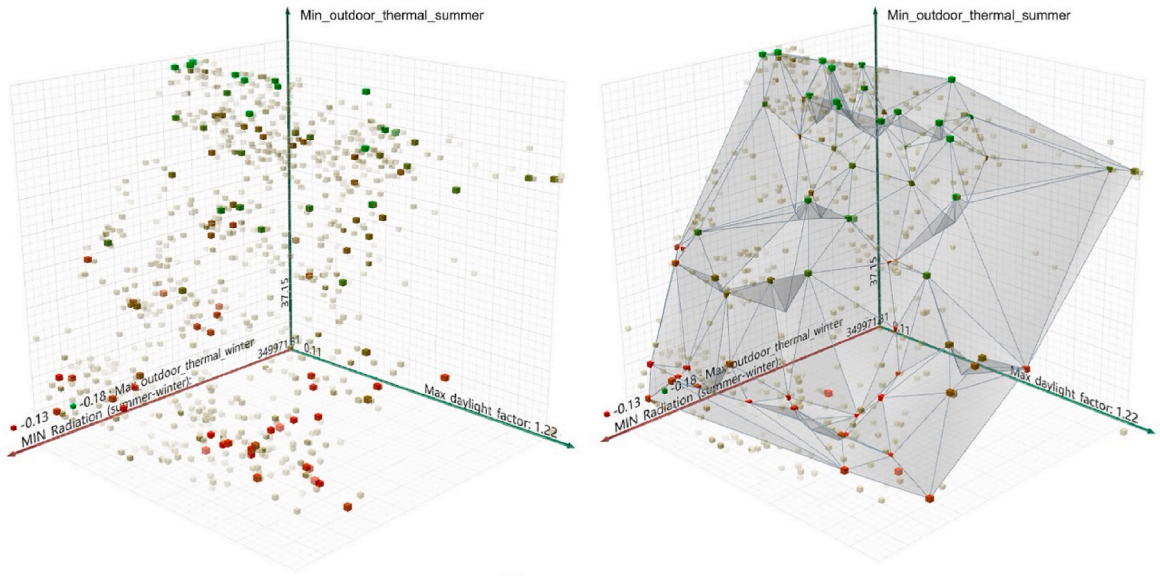


Fig. 4. Pareto frontier solutions and non-frontier solutions (left) and the three-dimensional grid surface composed by the Pareto front solution (right).

where n is the number of training examples, y_i is the real label, $\hat{y}_i$ is the estimated label, L is the training loss function, and Ω is a regularization term. The training loss measures how well the model predicts, whereas the regularization term avoids overfitting. XGBoost has an ensemble of K regression trees whose individual prediction outcomes are denoted as $(f_k(x_i))$. The final estimated outcome was the sum of all the trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

Substituting Eq. (3) into Eq. (2), the objective function of the Kth tree can be represented as

$$obj(f_K) = \sum_{i=1}^{n} L\left(y_i,\ \hat{y}_i^{K-1} + f_K(x_i)\right) + \Omega(f_K) + const \tag{4}$$

The constant in the equation is derived from the regularization term of the first K-1 trees. Using the Taylor expansion, Eq. (4) is converted to

$$obj(f_K) = \sum_{i=1}^{n} \left[L\left(y_i, \hat{y}_i^{K-1}\right) + g_k f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i) \right] + \Omega(f_K) + const \tag{5}$$

Where $g_i = \partial_{\hat{y}_i^{K-1}} L(y_i, \hat{y}_i^{K-1})$ and $h_i = \partial^2_{\hat{y}_i^{K-1}} L(y_i, \hat{y}_i^{K-1})$. To reduce the model complexity, the regularization term is calculated as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{6}$$

where T is the number of leaves, ω is the weight of the leaves, λ and γ are coefficients with default values of 1 and 0, respectively.

Machine-learning models are used for prediction and feedback of new data; therefore, their generalization ability and model accuracy are important moderating factors. This study examined the generalization ability of the XGBoost algorithm using 10-fold cross-validation, followed by the grid search method to optimize the hyperparameters that affect the generalization performance and model accuracy.

When using the 10-fold cross-validation method, 80% of the data were randomly divided into a training dataset, and the remaining 20% were used as the test dataset [56]. After the algorithmic model was trained on the dataset, cross-validation was used to determine its generalization ability, which directly influenced the predictive feedback of the machine-learning model.

Several user-defined parameters must be tuned using the XGBoost algorithm, including parameters in terms of tree depth and number. In this study, we focused on the three hyperparameters "learning_rate," "max_depth," and "n_estimators." The optimization search process of the algorithm finds new classifiers to gradually reduce the loss function. Variations in hyperparameter combinations resulted in highly variable prediction results. Comprehensive optimization of hyperparameters can reduce the instability of the algorithm and produce optimal results, which is preferable for the manual adjustment of hyperparameters. The grid search method was used in this study for hyperparametric integrated search.

2.3.3. Model evaluation

Classification performance can be evaluated using two critical metrics, precision and recall. Precision is usually used to assess the quality of the results, whereas recall is used to evaluate the completeness of the results. Precision and recall are contradictory metrics. In general, recall values tend to be low when precision is high, whereas recall values tend to increase when precision is low. Eqs. (7) and (8) were used to calculate precision and recall, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

TP is true positive, FP is false positive, and FN is false negative.

The F-1 score was proposed to consider precision and recall. Eq. (9) shows the calculation of the F-1 score.

$$F_1 = \frac{2 \bullet precision \bullet recall}{precision + recall} \tag{9}$$

The core idea of the F-1 score is to improve precision and recall as much as possible while keeping the difference between them as small as possible. The F-1 score applies to dichotomous problems, and for multiclassification problems, the F-1 score of dichotomous problems is generalized to have two metrics, Micro-F1 and Macro-F1. This study adopted the Macro-F1 calculation method using the Scikit-learn library on the Python platform.

Furthermore, sensitivity analysis helps designers assess the extent to which various characteristics influence the dependent variables so that they can identify the key variables to be controlled and accordingly suggest corresponding design strategies. For the XGBoost algorithm, the

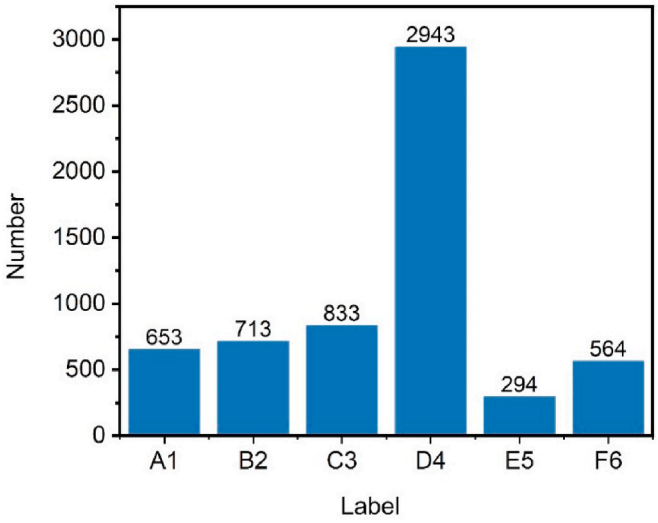


Fig. 5. Distribution of solutions attributed to different labels.

sensitivity of a feature is the sum of its occurrences in all trees based on the gain of the structure score. The more times an attribute is used in the model to build a decision tree, the higher is its importance score.

3. Results

3.1. Results of data acquisition and pre-processing

The dominant and non-dominated solutions generated by the MOGA after 100 generations of operation are shown on the left side of Fig. 4. The non-dominated solutions (i.e., optimized solutions) formed a Pareto front solution set and are represented by a three-dimensional grid surface (Fig. 4, right). These solutions have different positions in three-dimensional space, reflecting other performance characteristics. A total of 6000 solutions generated by the MOGA were given separate labels according to the lighting conditions in the above section. Fig. 5 shows the distribution of solutions attributed to different labels.

The data distribution for each design variable is shown in Fig. 6. The horizontal axis represents the values and the vertical axis represents the proportion of the data. The GFH was mainly 4–4.5 m, with an uneven overall distribution. The values of the remaining design variables were evenly distributed. The data distributions can help researchers understand the data obtained in this study while improving the accuracy and

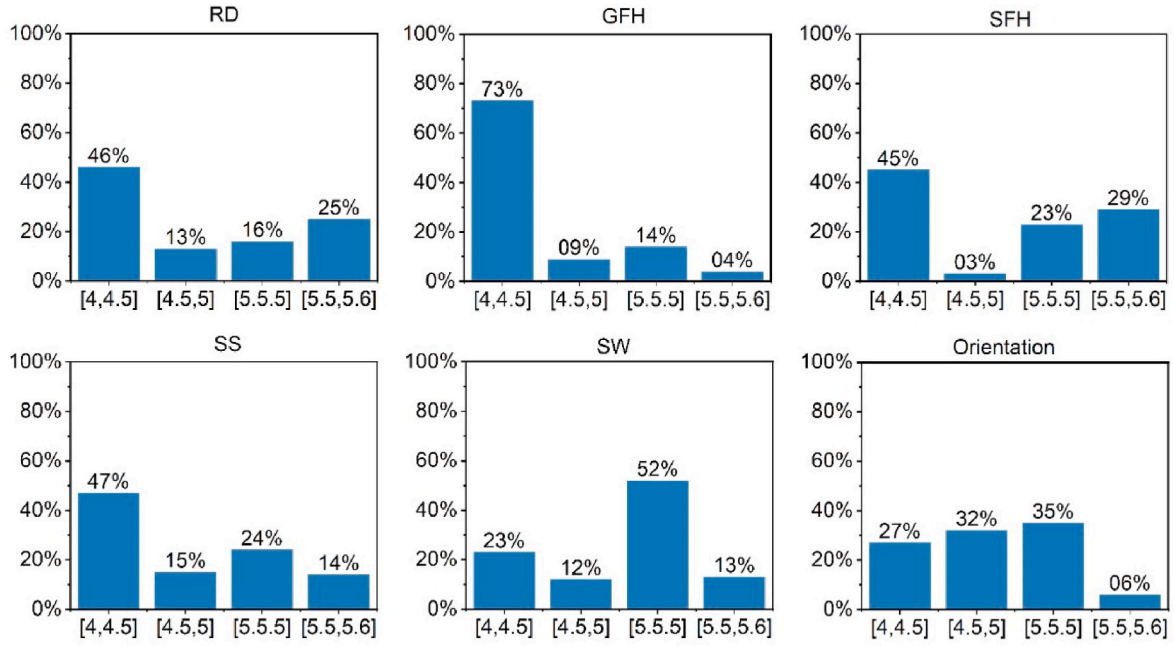


Fig. 6. Data distribution of the design variable.

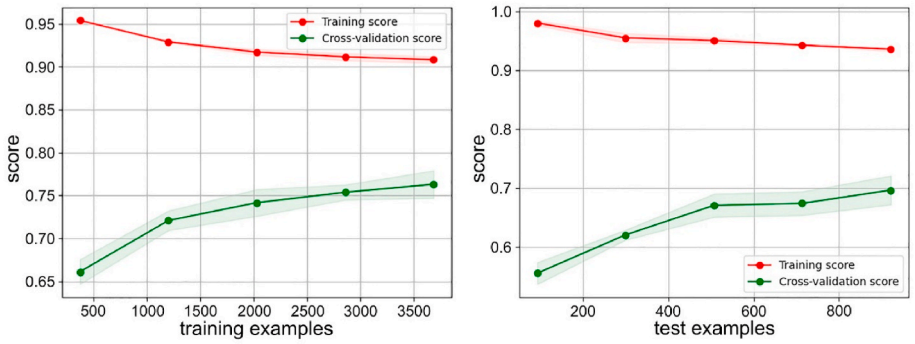


Fig. 7. Training curve (left) and learning curve (right).

Table 6
The tuned hyperparameters and evaluation metrics of XGBoost model.

n_estimators	max_depth	gamma	subsample	Precision	Recall	F-1 score
150	6	0.01	0.7	0.77	0.59	0.75

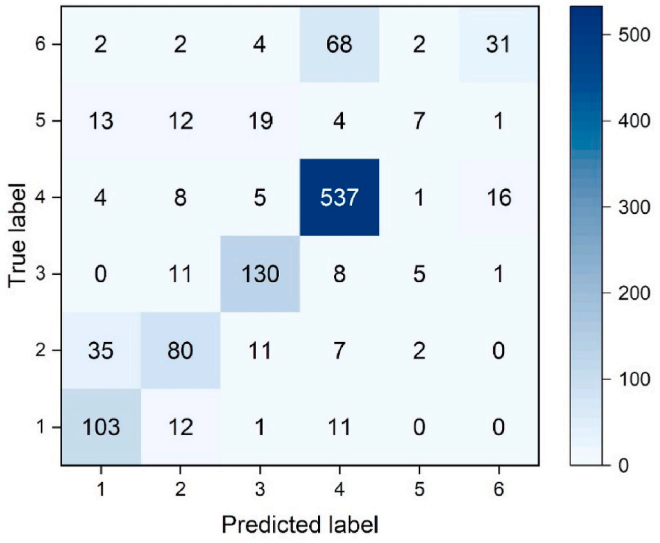


Fig. 8. Confusion matrix of the optimized XGBoost model.

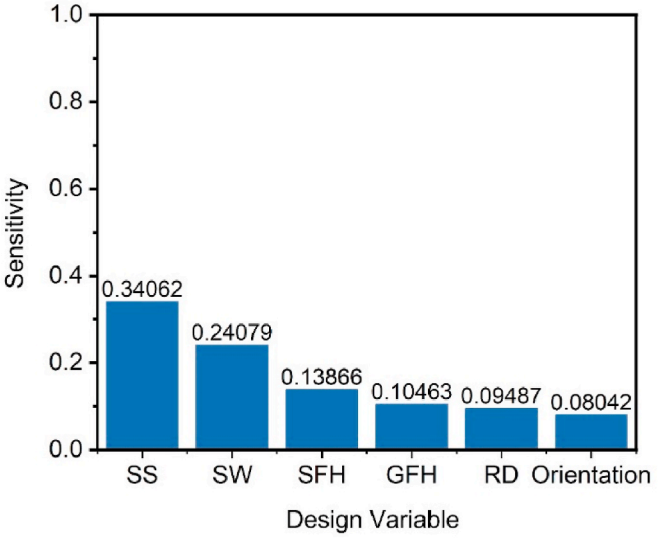


Fig. 9. Sensitivity analysis of label for all design variables.

applicability of the predictive models.

3.2. Results of hyperparameter tuning

When constructing the XGBoost model, several essential hyperparameters must be determined. The hyperparameters used in this study were as follows:

- "n_estimators": the number of base tree models, the higher the number of iterations, the higher the value;
- "max_depth": the maximum depth of the base tree model, with a higher value for more complex base tree models;
- 'gamma": the minimum loss reduction required to divide further on the leaf nodes of the tree, with higher values for more conservative models;
- "subsample": the subsample rate of the training instances.

The training process is shown on the left of Fig. 7, and the XGBoost model gradually stabilized its performance after 3500 epochs. The learning curve for the test dataset is shown on the right side of Fig. 7, which shows the relationship between the training sample size and the model performance. From the derivatives of the curves, the performance of XGBoost tended to be stable if the provided training data exceeded 800 samples. In Table 6 and Fig. 8, the tuned parameters and evaluation metrics from the XGBoost model are presented after construction.

Fig. 9 shows the sensitivity analysis for each design variable. Sunshade spacing (SS) has the most significant impact on the performance label, followed by sunshade width (SW) and SFH. The results indicate that the variables related to the shading elements and SFHs are closely related to the combined indoor and outdoor performances. The RD and orientation of the building were much less influential than the other parameters.

3.3. Validation of prediction results

Seven cases were randomly selected for the performance simulation and algorithm prediction to verify the accuracy of the categorical prediction algorithm in more detail and depth. Because the dataset with label D4 accounted for the largest number of cases, it accounted for two randomly selected cases. One case was selected for each of the other labels. We then compared the differences between the true and predicted labels, and the results are shown in Fig. 10 and Table 7.

Table 7 shows that the prediction results are accurate for most cases. However, the model proposed in this study sometimes predicts cases with a true label of D4 as a worse performance label. In addition, comparing the cases analyzed with labels A1 and F6, respectively, they differ significantly in SW and orientation design variables. Meanwhile, Section 3.2 shows that SW has a more significant impact on the final performance label. The case with the better overall performance (Label: A1) has a smaller SW of 0.7 m. The results indicate that the narrower shading elements ensure adequate interior lighting while performing well in other performance aspects.

4. Discussion

This paper proposes a study framework that combines MOGO with the XGBoost algorithm, which has advantages in terms of efficiency and practicality over studies based on GAs alone for better adoption by designers. To further verify the accuracy and benefits of the XGBoost algorithm, this study used the random forest (RF) [57], LGBM [58], and AdaBoost algorithms [59] for comparative analysis. In the training process, 80% of the 6000 sets of data were used for training and 20% for testing. In keeping with the application of XGBoost, this study combined a 10-fold cross-validation and grid search approach to train and optimize the hyperparameters for RF, LGBM, and AdaBoost. The tuned parameters for the RF, LGBM, and AdaBoost are listed in Table 8. The categorical prediction evaluation results for the four algorithms are presented in Table 9. The results showed that the XGBoost algorithm exhibited the highest F-1 score (0.75) and was the most desirable algorithm for categorical prediction. The best algorithm was the LGBM method, with an F-1 score of 0.73. The worst prediction algorithm was the AdaBoost method with an F-1 score of 0.63.

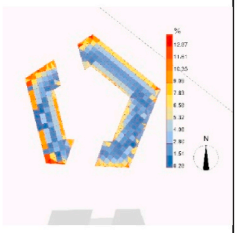
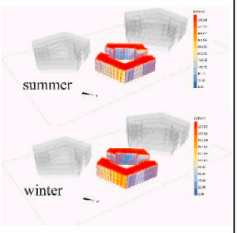
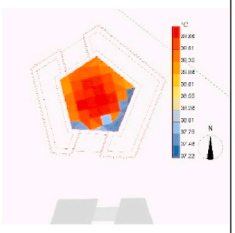
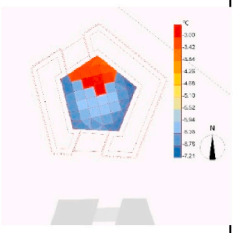
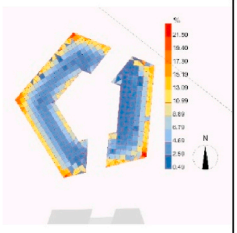
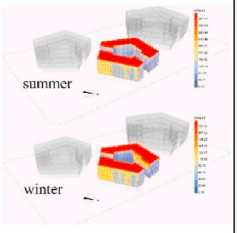
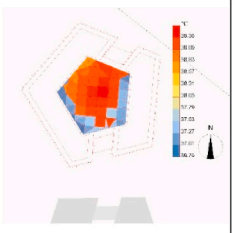
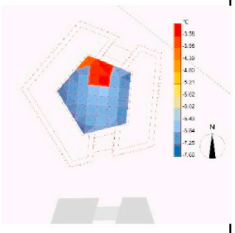
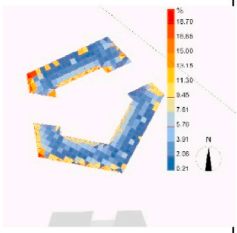
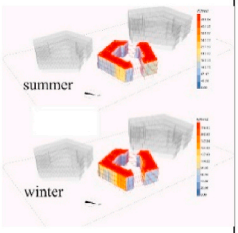
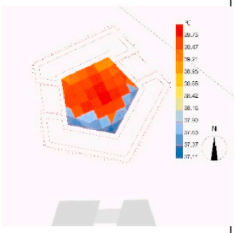
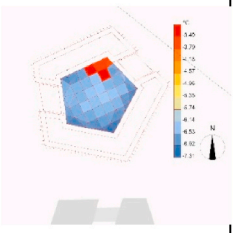
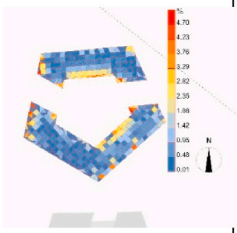
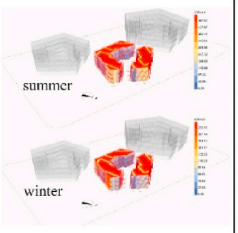
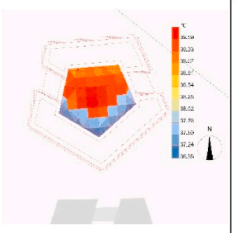
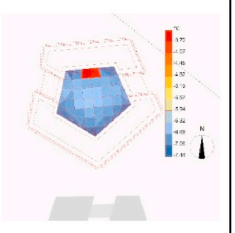
Index	Simulation results of DF	Simulation results of TSR	Simulation results of UTCI (summer)	Simulation results of UTCI (winter)
1				
	DF:3.63%	TSR:440061.23 kWh	UTCI:38.90 ℃	UTCI:-0.18 ℃
2				
	DF:5.91%	TSR:609431.30 kWh	UTCI:38.23 ℃	UTCI:-0.15 ℃
3				
	DF:4.50%	TSR:424576.39 kWh	UTCI:38.65 ℃	UTCI:-0.15 ℃
4				
	DF:0.85%		UTCI:38.38 ℃	UTCI:-0.14 ℃

Fig. 10. Performance simulation results for seven random cases.

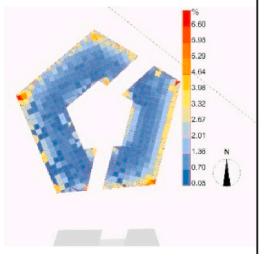
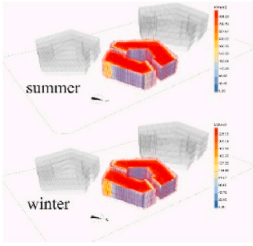
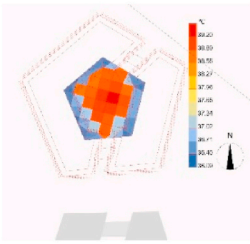
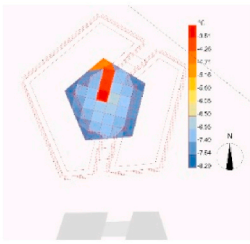
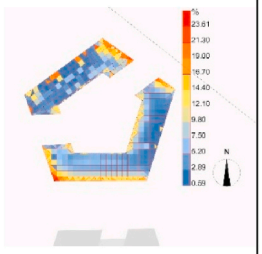
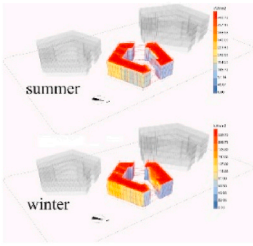
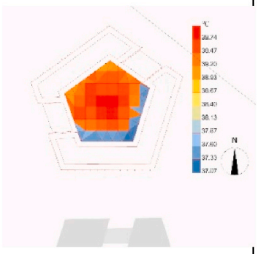
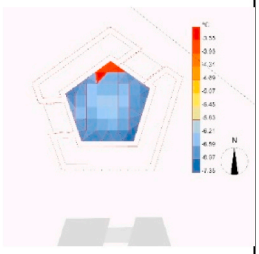
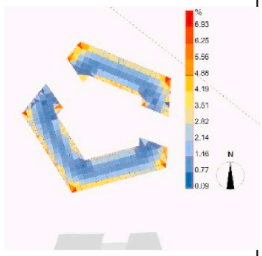
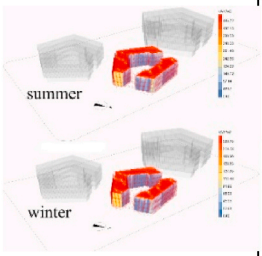
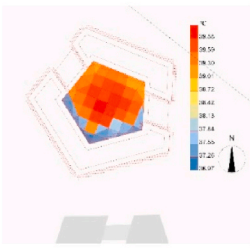
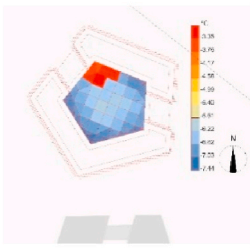
		TSR:440228.70 kWh		
5				
	DF:1.31%	TSR:568213.90 kWh	UTCI:37.34 ℃	UTCI:-0.14 ℃
6				
	DF:5.90%	TSR:441676.40 kWh	UTCI:38.74 ℃	UTCI:-0.15 ℃
7				
	DF:1.95%	TSR:399011.74 kWh	UTCI:38.61 ℃	UTCI:-0.15 ℃

Fig. 10. (*continued*).

In addition, the efficiency of the XGBoost algorithm is noteworthy. In contrast to the building performance simulation process or genetic optimization process, the XGBoost algorithm can evaluate and predict the performance of building design solutions based on the training dataset in a very short period, which significantly reduces computing power and time costs. Nevertheless, the generalizability of the categorical prediction model presented in this study should be explored further. If the design variables are replaced or the prediction target is changed, prediction accuracy may be reduced.

Finally, it is worth noting that this study used the DF to classify performance labels, which reflected the concentration of indoor lighting levels across the hot summer and cold winter regions of China. However, relevant studies can also use UTCI as a criterion for performance label classification [39]. UTCI has clear criteria for classifying thermal comfort levels; for example, when UTCI is in the range of 32–38 °C, the corresponding thermal comfort level is "hot." If the performance evaluation system is established using UTCI values, a new categorical prediction model is trained.

Table 7
Design variables and labels for seven random cases.

Index	RD	GFH	SFH	SS	SW	Orientation	True label	Predicted label
1	11.7 m	4.3 m	4.1 m	0.6 m	0.7 m	215.2°	A1	A1
2	13.1 m	5.3 m	5.0 m	1.5 m	0.8 m	16.1°	B2	B2
3	11.7 m	4.8 m	4.5 m	1.5 m	1.1 m	135.6°	C3	C3
4	12.2 m	5.0 m	4.7 m	0.5 m	1.3 m	116.2°	D4	D4
5	14.3 m	4.0 m	5.0 m	0.5 m	1.5 m	6.8°	D4	F6
6	11.8 m	4.1 m	5.0 m	1.4 m	0.5 m	147.3°	E5	E5
7	11.7 m	4.0 m	4.2 m	0.5 m	1.5 m	88.8°	F6	F6

Table 8
The tuned hyperparameters for Random Forest, LGBM, and AdaBoost.

Algorithm Models	Hyperparameters		
Random Forest	n_estimators	min_simples_leaf	min_simples_split
	200	30	5
LGBM	n_estimators	max_depth	learning_rate
	150	12	0.05
AdaBoost	n_estimators	algorithm	learning_rate
	150	SAMME.R	0.2

Table 9
Comparative analysis of the evaluation results of different categorical prediction algorithms.

Algorithm Models	Accuracy of the training set	Accuracy of the test set	Precision	Recall	F-1 score
XGBoost	0.90	0.77	0.77	0.59	0.75
Random Forest	0.73	0.71	0.44	0.48	0.64
LGBM	0.89	0.75	0.75	0.57	0.73
AdaBoost	0.69	0.69	0.69	0.46	0.63

5. Conclusion

To quickly predict and effectively optimize the comprehensive indoor and outdoor performance of office buildings, this study analyzed the feasibility of applying machine-learning methods to building design and constructed a corresponding study framework. Through field studies and data collation of a typical office building in Nanjing, this study used it as a study case for experimental application. First, the architectural design variables and simulation data, along with their corresponding performance labels, were incorporated as training data for machine learning. Furthermore, supervised learning methods were used to construct several classification prediction models. The XGBoost algorithm with better prediction accuracy was finally selected, and the model parameters were tuned until the F-1 score reached 0.75 using a grid search method with cross validation. This study clearly demonstrates that an early design approach based on machine learning for energy efficiency in office buildings can efficiently assist architects in generating comprehensive design solutions with excellent performance.

However, the present study has many limitations. First, this study focuses only on variables in the early design phase of office buildings and therefore does not consider more comprehensive design variables in detail. A design prototype is provided only for designers during the conceptual design phase of the building form and façade. Second, this study focused on solar thermal radiation, daylight coefficients, and outdoor thermal comfort performance objectives. Further studies are needed to extend other performance goals and appropriately match them. However, it is worth pointing out that when too many performance objectives are considered simultaneously, the algorithm becomes less efficient and the optimization direction becomes ambiguous during the optimization phase [60,61]. Finally, the study framework proposed in this study applies to specific studies and may serve as a reference for similar studies; however, its generalizability is limited.

In summary, further studies should integrate various design variables of office buildings and perform efficient design iterations and rapid interactive evaluation in a controlled manner based on generative logic, ultimately providing ample strategy space for architectural design. Furthermore, future studies should actively explore other machine learning algorithms as well as deep learning algorithms to solve practical and complex problems in the field of architecture and develop effective solutions.

CRediT authorship contribution statement

Hainan Yan: Writing – original draft. **Ke Yan:** Supervision. **Guohua Ji:** Supervision.

Declaration of competing interest

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

Acknowledgements

This work was funded by the National Natural Science Fundation of China (52178017) and the Opening Fund of Key Laboratory of Interactive Media Design Equipment Service Innovation, Ministry of Culture and Tourism (Grant No.20204).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2022.109081.

References

[1] CABEE, China building energy consumption research report 2020, available at: https://www.cabee.org/site/content/23568.html, 2020. (Accessed 20 January 2022).
[2] P. Anand, C. Deb, K. Yan, J. Yang, D. Cheong, C. Sekhar, Occupancy-based energy consumption modelling using machine learning algorithms for institutional buildings, Energy Build. 252 (2021), 111478.
[3] J. Li, C. Zhang, Y. Zhao, W. Qiu, Q. Chen, X. Zhang, Federated learning-based short-term building energy consumption prediction method for solving the data silos problem, Build. Simulat. 15 (2022) 1145–1159.
[4] M. Rastegari, S. Pournaseri, H. Sanaieian, Daylight optimization through architectural aspects in an office building atrium in Tehran, J. Build. Eng. 33 (2021), 101718.
[5] S. Yang, M.P. Wan, W. Chen, B.F. Ng, S. Dubey, Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization, Appl. Energy 271 (2020), 115147.
[6] M. Röck, A. Hollberg, G. Habert, A. Passer, Lca, Bim, Visualization of environmental potentials in building construction at early design stages, Build. Environ. 140 (2018) 153–161.
[7] P.F. Yuan, Y. Song, Y. Lin, H.S. Beh, Y. Chao, T. Xiao, S. Huang, J. Zheng, Z. Wu, An architectural building cluster morphology generation method to perceive, derive, and form based on cyborg-physical wind tunnel (CPWT), Build. Environ. 203 (2021), 108045.
[8] M. Bracht, A. Melo, R. Lamberts, A metamodel for building information modeling-building energy modeling integration in early design stage, Autom. ConStruct. 121 (2021), 103422.

[9] R. Danhaive, C.T. Mueller, Design subspace learning: structural design space exploration using performance-conditioned generative modeling, Autom. ConStruct. 127 (2021), 103664.
[10] S. Zhao, E. De Angelis, Performance-based generative architecture design: a review on design problem formulation and software utilization, J. Integrated Des. Process Sci. 22 (3) (2018) 55–76.
[11] X. Li, B. Yang, Y. Liu, L. Chen, R. Guo, F. Wang, K. Yan, Comparison of models for predicting winter individual thermal comfort based on machine learning algorithms, Build. Environ. (2022), 108970.
[12] D. D'Agostino, P. D'Agostino, F. Minelli, F. Minichiello, Proposal of a new automated workflow for the computational performance-driven design optimization of building energy need and construction cost, Energy Build. 239 (2021), 110857.
[13] Y. Dong, C. Sun, Y. Han, Q. Liu, Intelligent optimization: a novel framework to automatize multi-objective optimization of building daylighting and energy performances, J. Build. Eng. 43 (2021), 102804.
[14] B. Lin, H. Chen, Q. Yu, X. Zhou, S. Lv, Q. He, Z. Li, MOOSAS–A systematic solution for multiple objective building performance optimization in the early design stage, Build. Environ. 200 (2021), 107929.
[15] Z. Tian, X. Zhang, X. Jin, X. Zhou, B. Si, X. Shi, Towards adoption of building energy simulation and optimization for passive building design: a survey and a review, Energy Build. 158 (2018) 1306–1316.
[16] N. Somu, G.R. Mr, K. Ramamritham, A deep learning framework for building energy consumption forecast, Renew. Sustain. Energy Rev. 137 (2021), 110591.
[17] C.W. Kwon, K.J. Lee, Integrated daylighting design by combining passive method with daysim in a classroom, Energies 11 (11) (2018) 3168.
[18] Y. Xu, Building performance optimization for university dormitory through integration of digital gene map into multi-objective genetic algorithm, Appl. Energy (2021), 118211.
[19] Z. Luo, C. Sun, Q. Dong, X. Qi, Key control variables affecting interior visual comfort for automated louver control in open-plan office–a study using machine learning, Build. Environ. 207 (2022), 108565.
[20] Q. He, Z. Li, W. Gao, H. Chen, X. Wu, X. Cheng, B. Lin, Predictive models for daylight performance of general floorplans based on CNN and GAN: a proof-of-concept study, Build. Environ. 206 (2021), 108346.
[21] N. Jin, F. Yang, Y. Mo, Y. Zeng, X. Zhou, K. Yan, X. Ma, Highly accurate energy consumption forecasting model based on parallel LSTM neural networks, Adv. Eng. Inf. 51 (2022), 101442.
[22] C.-H. Lin, Y.-S. Tsay, A metamodel based on intermediary features for daylight performance prediction of façade design, Build. Environ. 206 (2021), 108371.
[23] K. Yan, X. Zhou, J. Chen, Collaborative deep learning framework on IoT data with bidirectional NLSTM neural networks for energy consumption forecasting, J. Parallel Distr. Comput. 163 (2022) 248–255.
[24] H. Mo, H. Sun, J. Liu, S. Wei, Developing window behavior models for residential buildings using XGBoost algorithm, Energy Build. 205 (2019), 109564.
[25] Y. Han, L. Shen, C. Sun, Developing a parametric morphable annual daylight prediction model with improved generalization capability for the early stages of office building design, Build. Environ. 200 (2021), 107932.
[26] NBBJ, Nanjing ecological science and technology Island, available at: http://www.nbbj.com/work/nanjing-eco-hi-tech-island/, 2017. (Accessed 26 December 2021).
[27] X. Shiwen, Standard for Design of Office Building JGJ/T 67-2019, 2019.
[28] S. Davidson, Grasshopper-algorithmic Modeling for Rhino, Lynnwood: United States, 2013.
[29] P. Tregenza, The daylight factor and actual illuminance ratios, Light. Res. Technol. 12 (2) (1980) 64–68.
[30] K. Błażejczyk, G. Jendritzky, P. Bröde, D. Fiala, G. Havenith, Y. Epstein, A. Psikuta, B. Kampmann, An introduction to the universal thermal climate index (UTCI), Geogr. Pol. 86 (1) (2013) 5–10.
[31] M.S. Roudsari, M. Pak, A. Smith, Ladybug: a parametric environmental plugin for grasshopper to help designers create an environmentally-conscious design, in: Proceedings of the 13th International IBPSA Conference Held in Lyon, France Aug, 2013, pp. 3128–3135.
[32] G. Evola, V. Costanzo, C. Magrì, G. Margani, L. Marletta, E. Naboni, A novel comprehensive workflow for modelling outdoor thermal comfort and energy demand in urban canyons: results and critical issues, Energy Build. 216 (2020), 109946.
[33] K. Kensek, J.Y. Suk, Daylight factor (overcast sky) versus daylight availability (clear sky) in computer-based daylighting simulations, J. Creative. Sustain. Architect. Built. Environ. 1 (2011) 3–14.
[34] A. Nabil, J. Mardaljevic, Useful daylight illuminance: a new paradigm for assessing daylight in buildings, Light. Res. Technol. 37 (1) (2005) 41–57.
[35] Y. Bian, T. Luo, Investigation of visual comfort metrics from subjective responses in China: a study in offices with daylight, Build. Environ. 123 (2017) 661–671.
[36] B.N. Mohapatra, M.R. Kumar, S.K. Mandal, Analysis of daylighting using daylight factor and luminance for different room scenarios, Int. J. Civ. Eng. Technol. 9 (2018) 949–960.
[37] L.E. Mavromatidis, X. Marsault, H. Lequay, Daylight factor estimation at an early design stage to reduce buildings' energy consumption due to artificial lighting: a numerical approach based on Doehlert and Box–Behnken designs, Energy 65 (2014) 488–502.
[38] L. Zhang, L. Zhang, Y. Wang, Shape optimization of free-form buildings based on solar radiation gain and space efficiency using a multi-objective genetic algorithm in the severe cold zones of China, Sol. Energy 132 (2016) 38–50.
[39] P. Bröde, D. Fiala, K. Błażejczyk, I. Holmér, G. Jendritzky, B. Kampmann, B. Tinz, G. Havenith, Deriving the operational procedure for the universal thermal climate index (UTCI), Int. J. Biometeorol. 56 (3) (2012) 481–494.
[40] A. Matzarakis, S. Muthers, F. Rutz, Application and comparison of UTCI and PET in temperature climate conditions, Finisterra 49 (98) (2014) 11–21.
[41] S. Wang, Y.K. Yi, N. Liu, Multi-objective optimization (MOO) for high-rise residential buildings' layout centered on daylight, visual, and outdoor thermal metrics in China, Build. Environ. 205 (2021), 108263.
[42] I. Dincer, M.A. Rosen, P. Ahmadi, Optimization of Energy Systems, John Wiley & Sons, 2017.
[43] N. Hashempour, R. Taherkhani, M. Mahdikhani, Energy performance optimization of existing buildings: a literature review, Sustain. Cities Soc. 54 (2020), 101967.
[44] B. Kiss, Z. Szalay, Modular approach to multi-objective environmental optimization of buildings, Autom. ConStruct. 111 (2020), 103044.
[45] R. Vierlinger, K. Bollinger, Acommodating Change in Parametric Design, 2014.
[46] F. Rosso, V. Ciancio, J. Dell'Olmo, F. Salata, Multi-objective optimization of building retrofit in the Mediterranean climate by means of genetic algorithm application, Energy Build. 216 (2020), 109945.
[47] P. Satrio, T.M.I. Mahlia, N. Giannetti, K. Saito, Optimization of HVAC system energy consumption in a building using artificial neural network and multi-objective genetic algorithm, Sustain. Energy Technol. Assessments 35 (2019) 48–57.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[49] V.M. Barthelmes, Y. Heo, V. Fabi, S.P. Corgnati, Exploration of the Bayesian Network framework for modelling window control behaviour, Build. Environ. 126 (2017) 318–330.
[50] R. Markovic, E. Grintal, D. Wölki, J. Frisch, C. van Treeck, Window opening model using deep learning methods, Build. Environ. 145 (2018) 319–329.
[51] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[52] C. Fan, F. Xiao, Y. Zhao, A short-term building cooling load prediction method using deep learning algorithms, Appl. Energy 195 (2017) 222–233.
[53] K. Chen, H. Chen, L. Liu, S. Chen, Prediction of weld bead geometry of MAG welding based on XGBoost algorithm, Int. J. Adv. Manuf. Technol. 101 (9) (2019) 2283–2295.
[54] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China, Energy Convers. Manag. 164 (2018) 102–111.
[55] Y.-H. Chen, J.-L. Chen, Ai@ ntiphish—machine learning mechanisms for cyber-phishing attack, IEICE Trans. Info Syst. 102 (5) (2019) 878–887.
[56] L. Yao, M. Cai, Y. Chen, C. Shen, L. Shi, Y. Guo, Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning, Epilepsy Behav. 96 (2019) 92–97.
[57] A. Paul, D.P. Mukherjee, P. Das, A. Gangopadhyay, A.R. Chintha, S. Kundu, Improved random forest for classification, IEEE Trans. Image Process. 27 (8) (2018) 4012–4024.
[58] M. Saha, S. Nayak, N. Mohanty, V. Baral, I. Rout, Preterm Delivery Prediction Using Gradient Boosting Algorithms, Communication and Intelligent Systems, Springer, 2021, pp. 59–68.
[59] T.-K. An, M.-H. Kim, A New Diverse AdaBoost Classifier, 2010 International Conference on Artificial Intelligence and Computational Intelligence, IEEE, 2010, pp. 359–363.
[60] S. Attia, M. Hamdy, W. O'Brien, S. Carlucci, Assessing gaps and needs for integrating building performance optimization tools in net zero energy buildings design, Energy Build. 60 (2013) 110–124.
[61] J. Zhang, N. Liu, S. Wang, Generative design and performance optimization of residential buildings based on parametric algorithm, Energy Build. 244 (2021), 111033.