Build Simul (2023) 16: 2049–2061
https://doi.org/10.1007/s12273-023-1045-x

Ensemble machine learning framework for daylight modelling of various building layouts

Rashed Alsharif[1,2], Mehrdad Arashpour[1] (✉), Emad Golafshani[1], Milad Bazli[3], Saeed Reza Mohandes[4]

1. Department of Civil Engineering, Monash University, Melbourne, Australia
2. Department of Construction Engineering at AlQunfudah, Umm Al-Qura University, Saudi Arabia
3. College of Engineering, IT, and Environment, Charles Darwin University, Australia
4. Department of Mechanical, Aerospace and Civil Engineering, University of Manchester, UK

Abstract

The application of machine learning (ML) modelling in daylight prediction has been a promising approach for reliable and effective visual comfort assessment. Although many advancements have been made, no standardized ML modelling framework exists in daylight assessment. In this study, 625 different building layouts were generated to model useful daylight illuminance (UDI). Two state-of-the-art ML algorithms, eXtreme Gradient Boosting (XGBoost) and random forest (RF), were employed to analyze UDI in four categories: UDI-*f* (fell short), UDI-*s* (supplementary), UDI-*a* (autonomous), and UDI-*e* (exceeded). A feature (internal finish) was introduced to the framework to better reflect real-world representation. The results show that XGBoost models predict UDI with a maximum accuracy of $R^2 = 0.992$. Compared to RF, the XGBoost ML models can significantly reduce prediction errors. Future research directions have been specified to advance the proposed framework by introducing new features and exploring new ML architectures to standardize ML applications in daylight prediction.

Keywords

artificial intelligence
indoor environment
machine learning
parametric building layout
sunlight
visual comfort

Article History

Received: 12 January 2023
Revised: 05 May 2023
Accepted: 15 May 2023

1 Introduction

The main purpose of buildings is to provide comfortable indoor environments to achieve specific needs, such as housing and working. A comfortable indoor environment involves thermal, acoustic, visual, and air quality comfort (Chen et al. 2022). Often, comfort assessment is complex in the early design phase due to limitations in quantifying comfort (Pérez-Fargallo et al. 2018). However, building modelling tools have emerged to close this gap and improve the performance of buildings in the aforementioned comfortable aspects (Ghobad and Glumac 2018; Lv et al. 2019; Peng et al. 2020). The advancements in simulating thermal comfort and the air quality of indoor environments have been significant. Unfortunately, simulating the visual comfort aspect in these modelling tools has not developed at the same pace in terms of ease of use (Yngvesson and Adolfsson 2018).

The European standard EN 12665 defines visual comfort as "a subjective condition of visual well-being induced by the visual environment" (Michael and Heracleous 2017). One important factor is the indoor environment's distribution and illuminance (Carlucci et al. 2015). For example, if the indoor space is exceedingly illuminated, the occupant might be visually uncomfortable because of glare potential and vice versa. The principal source of illumination in buildings is by allowing daylight for passive lighting, which helps significantly in reducing the energy consumption that lighting fixtures use (Day et al. 2019). Visual comfort is an important factor for improved cognitive conditions of occupants, ultimately improving their performance in workplaces or residential buildings. Exceeded or absent illumination of indoor spaces drives occupants to put a cognitive load on spatial awareness processing, which might distract or exhaust them and become unproductive as they should be (Shi et al. 2021; Liu et al. 2022). Therefore, designing energy-efficient and thermally comfortable buildings is as important as

E-mail: mehrdad.arashpour@monash.edu

List of symbols

ANN	artificial neural network	UDI	useful daylight illuminance
d	distance from a light sensor to a corner of a window	UDI-a	useful daylight illuminance (autonomous)
		UDI-e	useful daylight illuminance (exceeded)
I	internal finish	UDI-f	useful daylight illuminance (fell-short)
MAE	mean absolute error	UDI-s	useful daylight illuminance (supplementary)
ML	machine learning	w	rotation of a window with correspondence to a light sensor
MLM	machine learning model		
RF	random forest	x	distance from a light sensor to a perimeter obstacle
RMSE	root mean square error		
R^2	coefficient of determination	XGBoost	eXtreme Gradient Boosting

designing them for visually comfortable indoor spaces (Carlucci et al. 2015).

Recently, machine learning modelling (MLM) has become the mainstream in many scientific fields (Manfren et al. 2022). In this modelling approach, an algorithm is trained on an established dataset of inputs and objectives (Arashpour et al. 2022). In data-driven modelling, the computational effort is significantly reduced compared to traditional mathematical modelling (Arashpour et al. 2021; Thrampoulidis et al. 2021). Moreover, MLM is a robust technique that enables processing the complex computation of daylight engineering problems with minimal data and computational effort, especially in huge early stage planning (Ayoub 2020). For example, He et al. (2021) developed surrogate MLM to replace traditional daylight simulation tools using pixel-to-pixel visualisation datasets. Their findings reveal that the developed MLM can be 84 times faster than the standard DAYSIM/Radiance approach in handling layouts with 8732 light sensors. Finally, newly developed daylight metrics can be predicted by MLMs, which is not possible with current daylight modelling tools using a standard set of metrics (Chi 2022).

ML method in the daylight and visual comfort domain has been used in different ways but without creating standard approaches for this domain (Ngarambe et al. 2022). Arbab et al. (2021) developed four MLMs, including an artificial neural network (ANN) model to predict the illuminance inside a test room using a synthetically generated dataset. The MLMs were trained to predict the raw illuminance in (lux) by changing the louvres design only. Their findings revealed that the ANN model was the most accurate MLM to replace typical simulation approaches of louver designs. In addition, Lin and Tsay (2021) proposed a new concept of replacing typical geometrical design characteristics of test rooms with "intermediary features" to be the key features for ML development. These features were correlated with the indoor daylighting conditions of the test room. The results showed that the proposed MLM predicts daylight availability with an accuracy of R^2 = 0.91 with 90% savings in time compared with typical ray tracing simulation tools.

ML has also been used to optimize daylighting and visual comfort during the operational phase of buildings' lifecycle. Gunay et al. (2017) developed a discrete-time Markov logistic regression model approximation using a recursive algorithm to predict light fixture switching and blind control patterns inside a controlled building. Their approach minimized lighting energy consumption by around 25% without compromising the occupants' visual comfort in office and laboratory environments. Moreover, Luo et al. (2022) developed an ML-assisted model for automated louvres control. Their model-based control strategy was based on an efficient-compact set of variables that have been identified using a three-phase identification process, i.e., filtering features, embedding ML algorithms, and wrapping the model by trimming the least important features until the desired performance is reached. Their findings revealed that spatiotemporal features, such as the distance between occupant grid and each louver, dominate other features in terms of importance in developing MLM to replace repetitive typical simulation techniques.

The abovementioned literature review shows that there are potentials for standardizing ML application in daylight and visual comfort assessment. Hence, this study advances the current approach and elevates the domain towards a standardized stream. Deconstruction of building spatial layout components is obtained from the literature, and an important building characteristic (internal finish) is introduced as a training feature. In addition, a recently developed ML training technique (eXtreme Gradient Boosting) (XGBoost) is tested on the presented spatial components approach. Answers to the following research questions are of this paper's concern:

- How accurate an XGBoost ML model is in predicting daylight?

- How does the XGBoost ML model perform against another popular decision tree ML model in predicting daylight?
- What is the potential scalability of standardizing this approach in daylight ML modelling?

The contribution of this study to the literature lies in three main aspects. First, a new feature is introduced to an established daylight ML modelling approach. Second, the application of state-of-the-art ML algorithms (i.e., XGBoost) in daylight ML modelling is explored. Finally, the daylighting conditions of a southern hemisphere region are used to expose this approach to new horizons.

The content structure of the paper is as follows: Section 2 highlights the theory behind the spatial component deconstruction approach and the application of ML modelling in the daylight and visual comfort domain. Section 3 presents the detailed methodology used to apply the presented theories in a simulated environment. Section 4 provides the results of this study with discussions. Finally, Section 5 presents the conclusions of this research.

2 Theory

2.1 Daylight

Daylight is the main factor influencing occupants' visual comfort (Davoodi et al. 2020). Many metrics have been explored to interpret how comfortably is the indoor space lit. The literature is not unanimous about which metric is best (Wagiman et al. 2021). The useful daylight illuminance (UDI) is one of the most interpretive metrics for daylight performance, refined in 2012 by Mardaljevic et al. after being firstly introduced in 2005 as a daylight metric (Nabil and Mardaljevic 2005; Mardaljevic et al. 2012). This metric, widely served in the literature, is calculated using hourly sky conditions (including the sun movement) from an existing dataset and has shown a robust assessment of indoor passive illuminance (Fang et al. 2022; Khidmat et al. 2022; Montaser Koohsari and Heidari 2022). In general, UDI provides a fraction of the time when the illuminance of a specific spot is within a nominated range. The illuminance of a specific spot is measured in *lux*, equal to the illumination of a 1 m^2 surface that is 1 m away from a single light source (Blackwell 2000). Because the daylight illuminance range includes desirable and undesirable levels of illuminance, UDI is introduced as four bins levels, including $UDI_{fell\text{-}short}$ (UDI-*f*) with an illuminance of less than 100 lux, $UDI_{supplementary}$ (UDI-*s*) for the values between 100 lux to 300 lux, $UDI_{autonomous}$ (UDI-*a*) for the values between 300 lux to 3000 lux, and $UDI_{exceeded}$ (UDI-*e*) for the values of more than 3000 lux (Mardaljevic et al. 2012). Figure 1 shows a visualization of the UDI four bins' categories. The classification

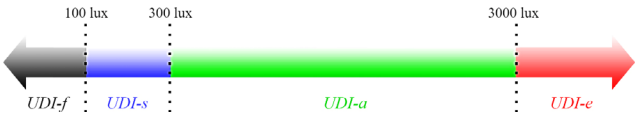

Fig. 1 UDI range with four bins of illuminance levels

of the UDI range helps achieve visually comfortable indoor environments in an early design phase and in determining possible glare and underlit spaces.

2.2 Building layout generation

Building layout generation is an essential process for generating a synthetic training dataset. The 4-square method presented by Le-Thanh et al. (2022) is a recent technique to generate different building layouts with a simple concept. There are four equal squares stacked to represent one large square. Each square moves towards a specific direction within a specific range to make a 4-square clockwise shift process. This process enables the modeler to generate a different building layout each time a slight movement of any square occurs (details in Appendix A). In addition, a random population of windows on one or more sides of the layout can be done to allow daylight illuminance. Several thresholds can be done to regulate the population of windows so they do not take the unusual window-to-wall ratios or be populated on undesirable sides of the layout. The detailed movements and directions of the 4-square method are shown in Table 1.

2.3 Decision trees ensemble models

Ensemble trees ML models combine weak ML models, such as decision trees, to generate a superior ML model that performs better than a weak ML model (Belitz and Stackelberg 2021; Arashpour et al. 2023). The two most popular techniques for developing decision tree-based ensemble models are bagging and boosting (González et al. 2020). Each decision tree is built using a randomly selected subset of the training dataset in bagging. The average prediction of decision trees for a given data point is the estimation of the bagging ensemble ML model (Zhang et al. 2022). A well-known representation of the bagging technique

Table 1 4-square method movements and directions to generate different building layouts

Square	Movement range	Unit	Axis of movement (to the *x*-axis)
A	0–2000	mm	0°
B	0–2000	mm	270°
C	0–2000	mm	180°
D	0–2000	mm	90°

is random forest (RF), in which each subset is chosen through a random selection process with replacement. RF handles higher dimensionality and missing data very well; however, since it ultimately takes the average of multiple decision trees, it might not be exact in objectives' values (Wang et al. 2019).

On the other hand, boosting technique organizes weak ML models differently. In decision tree-based boosting, decision trees are trained sequentially to minimize prediction error (Lou et al. 2016; Oyedele et al. 2021). Although boosting generates highly accurate models, it might be prone to overfitting if hyperparameters are mistuned (Arashpour 2023). eXtreme Gradient Boosting (XGBoost) is the cutting-edge representation of the boosting training techniques of ML models (Chen and Guestrin 2016). Figure 2 illustrates the training concept for both bagging and boosting.

2.4 Training data typology

Enabling ML models to predict daylight illuminance in different building layouts depends on many variables. The prediction of UDI cannot be made for the whole building layout at once, and it must be done using a sensor-based followed by collective gathering. Initially, the layout floor surface is deconstructed into a mesh of sensors that capture daylight illuminance in an hourly-based routine, as shown in Figure 3. Then, the annual amount of illuminance is estimated to identify the sensors' four bins' values, i.e., UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*. These objectives are used to develop every sensor's ML model separately based on several variables (Le-Thanh et al. 2022).

First, the perimeter distance from the sensor to every surrounding obstacle, e.g., walls, is calculated. The sensor becomes the source of 60 rays in 360° that measure how far each obstacle is from the sensor and in which direction. This information is stored in distance variables x_1, …, x_{60} (details in Appendix A). Second, the distance from every sensor to each corner of the 4 windows' is calculated by a set of 4 distance rays generated from the sensor to the windows' corners. This information is stored in distance variables d_{n1}, …, d_{n4}, where n is the window number.

It should be noted that the maximum number of windows is set to 4 in this study. Third, the position of the sensor in accordance with the window is determined by the variable w. It is the angle between a north y-axis generated from every sensor and the beginning or the end of every window. Because the maximum number of windows is set to 4, this information is stored in variables w_{n1}, …, w_{n2}, where n is the window number. Detailed information about these variables can be found in reference (Le-Thanh et al. 2022). Finally, we have introduced a variable to this approach called I. It is the total reflectance of the internal finish of a building layout. Internal finishes (or reflectance by internal surfaces) significantly influence the distribution of UDI within the internal space (Brembilla et al. 2022; Montaser Koohsari and Heidari 2022). Because this is not a sensor-based generated variable, sensors of the same building layouts are assigned with the same I.

Therefore, the structure of the training dataset is a matrix of m rows and 89 columns, where m is the number of sensors, and 89 is the sum of x, d, w, and I variables, in addition to the objectives UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*.

3 Research method

3.1 Model development

In this study, a new approach is made by performing daylight illuminance ML predictions using a weather dataset for a southern hemisphere region. Unlike regions in the northern hemisphere, the sun's path is tilted towards the north, making the northside façade more exposed to daylight illuminance than the south side (Alsharif et al. 2022). Melbourne, Victoria, is the location for the case study and all daylight simulations. The exact weather dataset

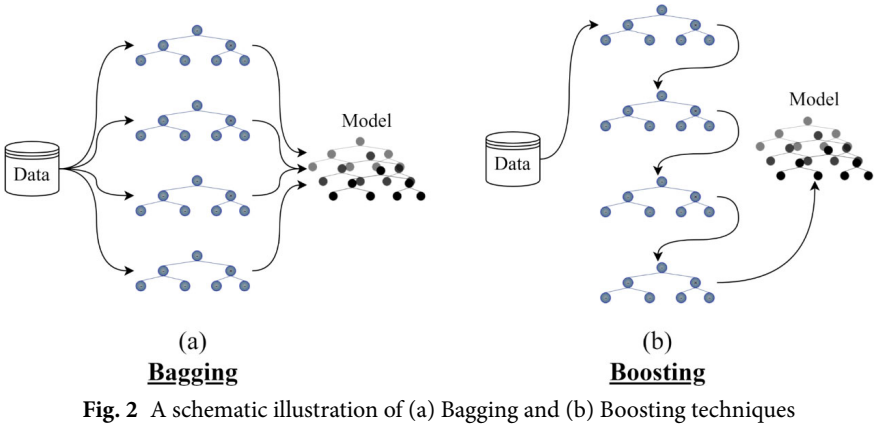

Fig. 2 A schematic illustration of (a) Bagging and (b) Boosting techniques

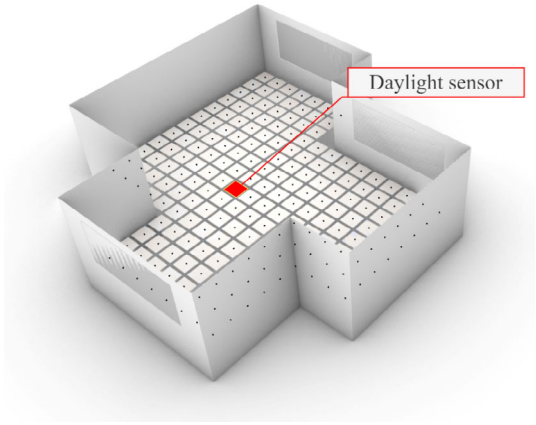

Fig. 3 Customized mesh of sensors to capture UDI

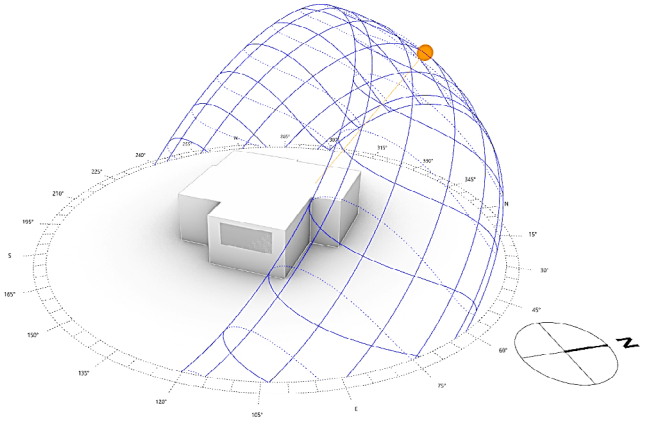

Fig. 4 The sun's path in the southern hemisphere

is "AUS_VIC.Melbourne.948680_(RMY)". Therefore, the machine learning models (MLMs) generated from this study cannot be generalized for use in other regions. The sun's path in the southern hemisphere is shown in Figure 4.

Figure 5 shows the workflow that generates and evaluates four ML models for UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*. As a beginning, 625 different building layouts are generated based on the 4-square method mentioned in Section 2.2. In this step, Grasshopper is used within Rhino 7 environment to code the building layout generation module. In parallel to the building layout generation, windows are populated randomly on one or more sides of each generated layout with several thresholds. The height of the building layout is fixed to 2700 mm, and the sill height is constrained to 1250 mm, with a window height of 1200 mm. Then, the working level plane (750 mm) is deconstructed into a mesh of sensors varying from 184 to 256 depending on the layout's size. It should be noted that the glazing system used in this study is a double-glazed system with 80% transmittance.

After generating the building layouts, ClimateStudio software is incorporated in Grasshopper to perform daylight simulation of all 625 cases to obtain the objectives UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e* of every sensor in every building layout. This process will generate a matrix of 126,967 × 89, where 126,967 is the number of total sensors, and 89 are the variables and objectives explained in Section 2.4. The obtained matrix is the dataset used for developing the ML models.

The dataset is divided into the training dataset (80%) and the testing dataset (20%). The division is based on the building layout and a total of 497 building layouts (101,630 sensors) for training and 128 building layouts (25,337 sensors) for testing. The training dataset is used to develop four ML models for four different predictions, i.e., UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*, using the XGBoost algorithm for decision tree-based boosting models. After the training phase is complete, the testing dataset is fed to the generated MLMs using only the variables *x, d, w,* and *I,* while holding out the objectives UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*. Finally, the predicted results are compared against the testing dataset to evaluate the performance of the MLMs.

3.2 Decision trees algorithms and hyperparameters tuning

Ensemble MLMs require hyperparameter tuning to predict objectives precisely, especially in the case of boosting models (Veloso et al. 2021). RF and XGBoost models are developed using the same dataset to compare their performances. In addition, hyperparameters tuning is conducted to maximize the predictions' accuracy.

Hyperparameters in MLM determine the learning process (Yang and Shami 2020). The tuning of these parameters changes the performance of MLMs. In this study, the tuned hyperparameters include the number of estimators, maximum depth, and learning rate.

The number of estimators is the number of decision trees used to generate the ensemble model. In some cases, more decision trees are preferred depending on the complexity of the interrelated variables. However, this comes with the cost of the high computational effort needed. Also, it may overfit the model to the training data if an excessive number of decision trees is assigned (Papadopoulos et al. 2018).

The maximum depth hyperparameter determines how many branches each decision tree has. This is a highly critical parameter due to its role in controlling overfitting the model. Higher depth causes a decision tree to overlearn relations of a specific sample and make it inaccurate to make generalizations with new datasets, while the opposite happens if shallow decision trees are assigned to the model (Shekar and Dagnew 2019).

The learning rate is the amount of shrinkage assigned to model features to make the model conservative. The algorithm assigns weight to every decision tree during the training process. Reducing the learning rate hyperparameter

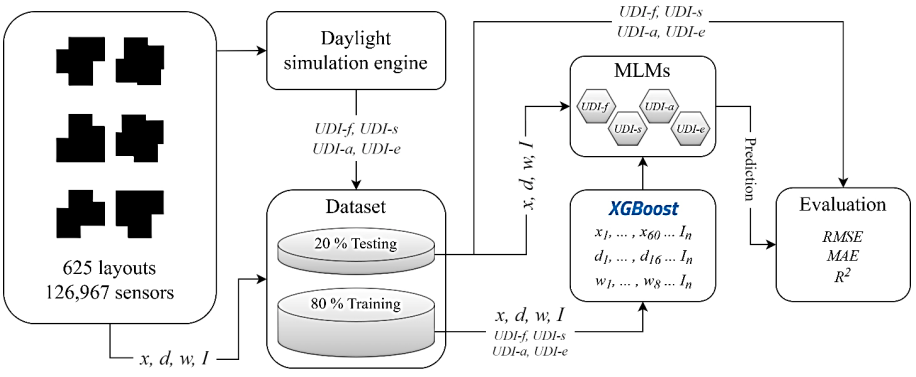

Fig. 5 Overall workflow used in the study

makes the model less flexible to learn new complexities throughout the dataset, while increasing it makes the model oscillate around ideal values and minimum errors (Park and Ho 2021).

The hyperparameters' tuning process involves establishing a range of values and training the models using a job list of all possible combinations of hyperparameters. Finally, the hyperparameters combination with minimum error is nominated as the tuned model (Veloso et al. 2021). The proposed ranges start with minimum values used in similar approaches and end with assumed values with the plan of expanding the range if the MLMs do not reach their best performance by that end. The scoring metric for assessing the accuracy of hyperparameters tuning is the root mean square error (RMSE) and is calculated using a cross-validation approach. The dataset is divided into (k = 10) folds, and the MLMs are developed using (k – 1 = 9) folds. The generated models are scored based on the remaining fold, and RMSE is calculated. The pseudocode explaining the hyperparameters tuning process is expressed in Figure 6.

Algorithm: Pseudocode for MLMs hyperparameters tuning

Input:
dataset = matrix$_{101630\times89}$
n = number of estimators (= 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000)
eta = learning rate (= 0.005, 0.01, 0.05, 0.1, 0.5, 1)
d = maximum depth (= 5, 7, 10, 12, 15, 20, 25, 30)
j = number of possible combinations of n, *eta*, d (j= 432)
XGBoost = f (n, *eta*, d)

start
for *run* = 1 to j **do**
  $n = n_{(1\text{-}run)}$, eta = $eta_{(1\text{-}run)}$, $d = d_{(1\text{-}run)}$
  split *dataset* into k (=10) folds
  for i = 1 to k **do**
    testing_data = fold i
    training_data = all the data except those in i_{th} fold
    develop *XGBoost* using *training_data*
    calculate the test error e using *testing_data*
  save e
  end
  save *XGBoost*
end
save the best *XGBoost* (minimum e)
report parameters n, *eta*, d for the best *XGBoost*
finish

Fig. 6 Pseudocode for MLMs hyperparameters tuning

3.3 ML models evaluation

To evaluate the accuracy of the MLMs models, error metrics consisting of RMSE, mean absolute error (MAE), and coefficient of determination (R^2) are used, defined as follows in Eq. (1)–(3):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,\mathrm{real}} - X_{i,\mathrm{mdl}}\right)^2} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|X_{i,\mathrm{real}} - X_{i,\mathrm{mdl}}\right| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(X_{i,\mathrm{mdl}} - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_{i,\mathrm{real}} - \bar{X}\right)} \quad (3)$$

where $X_{i,\mathrm{real}}$ is the actual simulation result, $X_{i,\mathrm{mdl}}$ is the prediction by MLMs models, $\bar{X}$ is the average of results, and n is the number of data records. Better performance of the MLMs is indicated when lower values of RMSE, MAE, and higher values of R^2 are obtained and vice versa. It should be noted that these metrics are calculated for eight models, each model of the four models (UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*) using the two algorithms, RF and XGBoost.

4 Results

4.1 Hyperparameters tuning

Figure 7 demonstrates the performance of the MLMs (i.e., UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*) with tuning the number of estimators (or trees) using the RMSE as the scoring metric for both models, RF and XGBoost. All models improved significantly before reaching 1000 decision trees in size. XGBoost models almost always perform better than RF models. The exception is for models UDI-*a* and UDI-*e* before reaching the 1000 decision trees size. This might be attributed to the lack of enough boosting due to the limited number of decision trees.

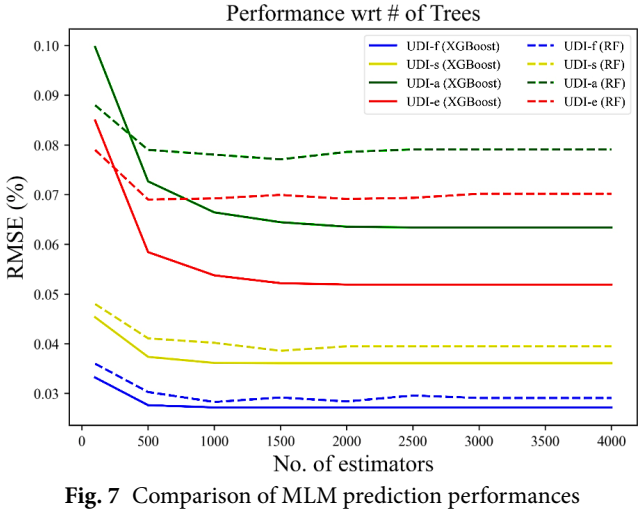

Fig. 7 Comparison of MLM prediction performances

The best performing values of all hyperparameters (i.e., No. of estimators, maximum depth, learning rate) are nominated for developing the MLMs and evaluated in the testing phase. Table 2 shows the best values after tuning these hyperparameters. The best number of estimators is not different for the UDI-*f*, UDI-*s*, and UDI-*e* models when using RF or XGBoost of 1000, 1500, and 2000 trees, respectively. However, it is different for the UDI-*a* model with 2500 trees in XGBoost against 1500 trees in RF. For the maximum depth of trees, the RF model shows its best performance with a depth of 20 in all models, while 10 is the best depth for trees in the XGBoost model.

The learning rate is a hyperparameter only for XGBoost models. The best learning rate value is 0.05 for UDI-*f*, UDI-*s*, and UDI-*a* models, while 0.01 is the optimum learning rate value for the UDI-*e* model.

4.2 Models' evaluation

The testing dataset is used to predict UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e* by feeding the variables *x, d, w,* and *I* to the developed MLMs, as in Figure 5. Then, the predicted UDI values are evaluated against the UDI values preserved in the testing dataset. Table 3 shows the RMSE, MAE, and R^2 for the MLMs, i.e., RF and XGBoost.

RF models show competitive performance in predicting UDI with a minimum R^2 of 0.88 in the UDI-*f* model. The most accurate RF model is the UDI-*e* model with RMSE, MAE, and R^2 of 6.91, 4.39, and 0.96, respectively. The UDI-*s* and UDI-*a* come in second and third in terms of accuracy with RMSE, MAE, and R^2 of 3.86%, 1.56%, and 0.94 for the UDI-*s* model, and 7.71, 5.36, and 0.94 for the UDI-*a* model respectively.

On the other hand, XGBoost models deliver excellent performing MLMs in predicting UDI values with a minimum R^2 of 0.972 in the UDI-*f* model. UDI-*a* is the most accurate MLM with RMSE, MAE, and R^2 of 2.72, 1.48, and 0.992, respectively. The second and third most accurate models are the UDI-*e* and UDI-*s,* respectively, with RMSE, MAE, and R^2 of 3.24, 2.07, and 0.991 for the UDI-*e*, and 1.71, 0.67, and 0.988 for the UDI-*s* respectively.

Figure 8 illustrates a scatter plot of the predicted UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e* values against the simulated values for the same sensors in the testing dataset using the XGBoost models. It can be noticed that UDI-*a* and UDI-*e* models have an excellent distribution of UDI values that helps the MLMs learn better the interrelationship between variables. Differently, UDI-*f* model has a poor distribution, which is potentially the reason behind this model being the

Table 2 Hyperparameters after tuning for both models RF and XGBoost

	UDI-*f*		UDI-*s*		UDI-*a*		UDI-*e*	
Hyperparameter	XGBoost	RF	XGBoost	RF	XGBoost	RF	XGBoost	RF
No. of estimators	1000	1000	1500	1500	2500	1500	2000	2000
Maximum depth	10	20	10	20	10	20	10	20
Learning rate	0.05	—	0.05	—	0.05	—	0.01	—

Table 3 Error metrics of both MLMs RF and XGBoost

	UDI-*f* (%)		UDI-*s* (%)		UDI-*a* (%)		UDI-*e* (%)	
Metric	XGBoost	RF	XGBoost	RF	XGBoost	RF	XGBoost	RF
RMSE	1.41	2.83	1.71	3.86	2.72	7.71	3.24	6.91
MAE	0.36	0.76	0.67	1.56	1.48	5.36	2.07	4.39
R^2	0.972	0.88	0.988	0.94	0.992	0.94	0.991	0.96

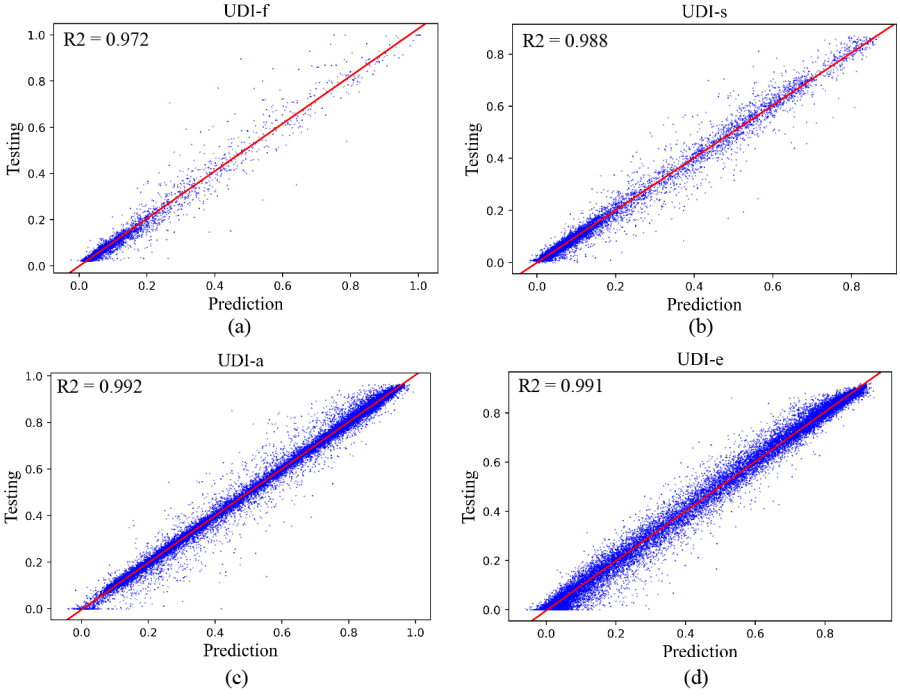

Fig. 8 Comparison between predicted and simulated samples for the testing dataset (XGBoost): (a) UDI-*f* model, (b) UDI-*s* model, (c) UDI-*a* model, (d) UDI-*e* model

least accurate. This distribution is attributed to the nature of UDI-*f* narrow threshold of 0–100 lux, which is rare in the dataset.

4.3 ML models performance

The MLMs perform very well, especially the XGBoost models, as shown in Table 3. However, a significant improvement can be noticed when looking at the UDI-*f* model. The UDI-*f* model represents the fell-short areas in providing adequate lighting over a year. It is different from other models because its nature is almost always insignificant except in narrow corners that light cannot always access. In XGBoost models, the accuracy of the UDI-*f* model has increased significantly from the RF models due to the sequential learning provided in XGBoost. Unlike in RF models, decision trees in XGBoost are not trained until their predecessors are trained. Therefore, the pattern of these unusual underlit areas is easier to be captured by such ML models. In RF models, decision trees are trained in parallel, which makes them more prone to miss the rare existence of the UDI-*f* model being significant.

Differently, the UDI-*e* model is highly accurate when using either of the training algorithms. As mentioned in Section 2.1, the UDI-*e* model determines when the sensors are exceedingly lit over a year. Usually, daylight exists the

most in areas beside windows. Therefore, the high accuracy in predicting UDI-*e* can be attributed to the rational correlation between the location of windows and exceedingly lit areas.

Another interesting observation is the absence of a correlation between models' prediction errors (RMSE and MAE) and models' accuracy (R^2). For example, the UDI-*f* model has a lower error in predicting the percentage of time a lighting condition is than the UDI-*a* model. However, the UDI-*a* has a higher prediction accuracy than the UDI-*f* model. This may be attributed to the ubiquity of patterns in the information of each model. In the same example, the UDI-*a* model has more patterns within its dataset than the UDI-*f* model. The range of illuminance of the UDI-*f* model is narrower than in the UDI-*a* model, as demonstrated in Figure 1. This makes the accuracy of prediction more possible despite the prediction error. The UDI-*e* model has the advantage of being accurate due to the reason mentioned before of having a direct relationship with windows location.

4.4 Comparison between predicted and simulated results

Figure 9 illustrates a 3-dimensional representation of

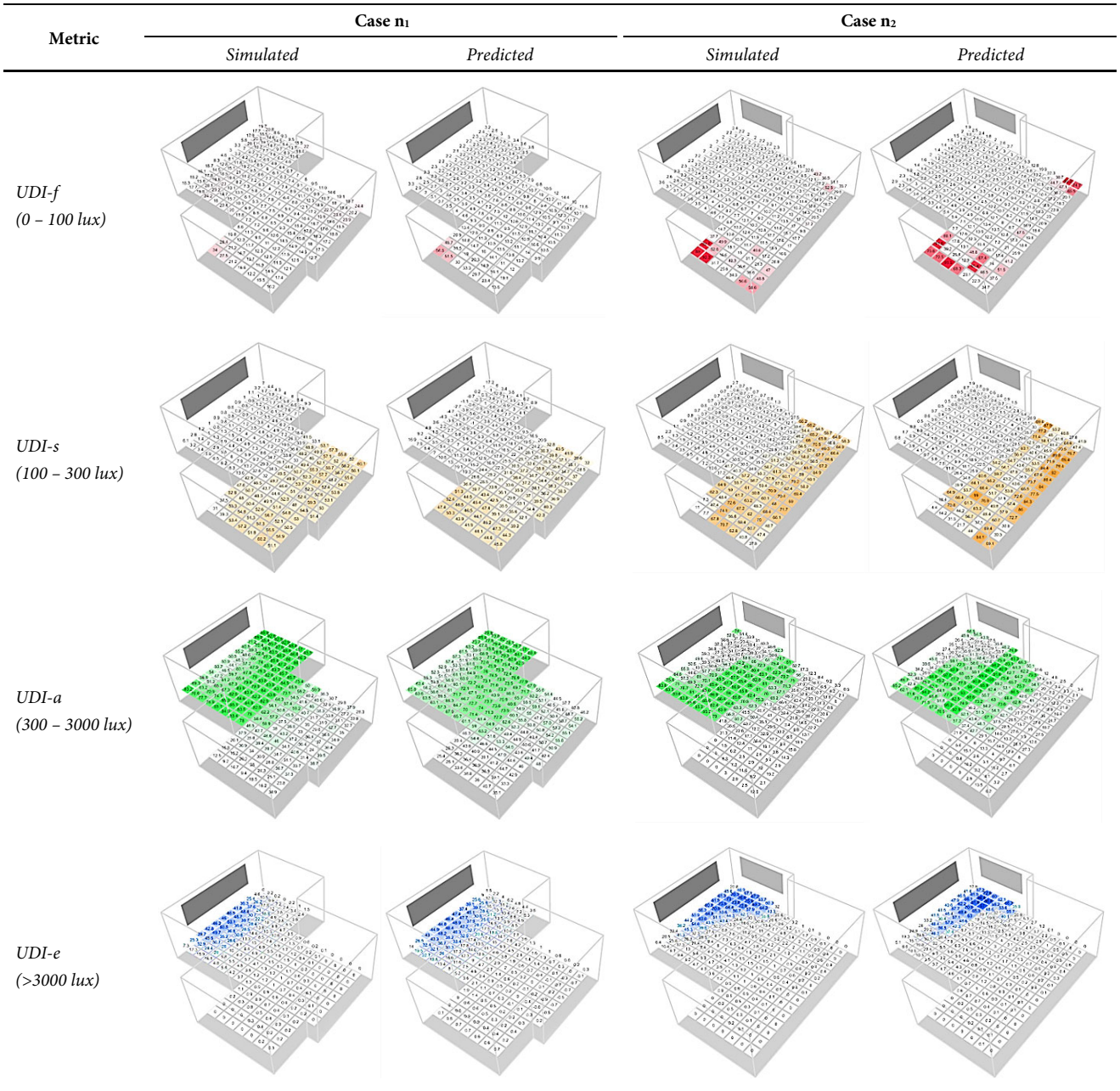

Fig. 9 3D illustration of predicted and simulated UDI values

randomly generated 2 cases, n_1 and n_2. The illustration compares these cases when simulated and predicted for the four models, UDI-*f*, UDI-*s*, UDI-*a*, and UDI-*e*. Case n_1 has 1 window oriented towards the west and has 184 sensors. Case n_2 has 2 windows oriented toward the west and north and has 211 sensors. The four UDI ranges are illustrated with four different colors. Sensors with low UDI values have brighter shades and become gradually saturated with the specified color as UDI values increase. The variable *I* denoting the internal finishing is determined using the total visible reflectance (R_{vis}) of the internal finish; it is 0.65 and 0.5 for the cases n_1 and n_2, respectively.

When looking at predictions of the UDI-*f* model illustrated in red color in Figure 9, the MLM can capture the general lack of illuminance pattern in the layout. In case n_1, almost all sensors are not "fell-short" in daylight illuminance throughout the year, except for the far south-eastern corner of the layout. This may be attributed to the inability of daylight to access this pocketed corner away from the only window available. The prediction of the developed XGBoost model was accurate enough to capture this pattern and provide the under-lit area. In case n_2, the same pattern exists in addition to the eastern corner of the layout. The predictions in case n_2 tend to be slightly exaggerated compared to case n_1. However, the collective pattern of under-lit areas is similar to the simulated model.

Next, the UDI-*s* model denotes areas with an illuminance of 100–300 lux throughout the year. In the case of n_1, the simulated model shows the far areas from the window as supplementarily lit but not under or autonomously lit. This pattern is captured successfully but with counting sensors previously captured by the UDI-*f* model. This contradiction can be avoided by introducing a new framework in future research that enables UDI models to correct each other in a hierarchical method and exclude already counted sensors in predecessor models. The current framework develops each model on a separate objective and makes predictions independent of other models. In case n_2, the MLM also captures the general pattern of the simulated model. However, areas close to walls seem to be either overestimated or underestimated. Luckily, the outcome is not considered as individual sensors but as a whole layout that enables the modeler to notice outlier sensors in the mesh of sensors.

The UDI-*a* model represents the autonomously lit places where mostly desirable illumination exists. In the case of n_1, the simulated model illuminance is within the UDI-*a* levels in the areas around the window but not the closest. This is an expected pattern as these areas are exposed to daylight during most of the day to provide 300–3000 lux illumination levels over the year. The predicted results of sensors follow a similar pattern collectively with excellent predictions. This is attributed to the verity of patterns within the dataset to enable detailed development of the UDI-*a* XGBoost model. The wide range of UDI-*a* illuminance helps in providing more results in the dataset for better development. Similarly, the simulated case n_2 shows UDI-*a* ranges around the two windows but not just under them. The prediction also shows an outstanding ability to generate a similar pattern, especially in the corner between the two windows with detailed and complex geometrical characteristics.

Finally, the UDI-*e* model presenting the exceedingly illuminated areas within the layout is color-coded blue in Figure 9. In this model, the direct relationship between these areas and windows location helps the model easily predict using variables *w* and *d* from the dataset. In the case of n_1, the simulated model shows exceedingly illuminated areas close to windows. This is due to the continuous exposure of these sensors to daylight throughout the year. The MLM model predicts these sensors accurately. In case n_2, the exceedingly illuminated sensors exist in the common area between the two windows. Similar to case n_1, the MLM predicts these sensors efficiently for the reasons above. In models UDI-*a* and UDI-*e*, the XGBoost generated accurate models for different reasons, including various patterns within the dataset and direct correlation (high sensitivity) with specific variables.

5 Conclusion

The presented study develops ensemble machine learning (EML) models for useful daylight illuminance (UDI) predictions. The development advances ML daylight modelling approaches in different fronts. A new feature to the ML training dataset *I* (internal finish) is introduced, and state-of-the-art EML algorithms, eXtreme Gradient Boosting and Random Forest, are employed. The XGBoost models are compared with another random forest (RF) algorithm-generated model set. The framework of this study consists of four main stages: synthetic dataset generation, dataset preparation, ML model training, and evaluation.

Four ML models describing the condition of UDI were developed to predict the visual comfort of building layouts, namely, UDI-*f* (fell short), UDI-*s* (supplementary), UDI-*a* (autonomous), and UDI-*e* (exceeded). Deconstruction of building layout to standardized spatial components is performed for a customized mesh of sensors using variables *x* (distance from a sensor to obstacles), *d* (distance from a sensor to corners of windows), *w* (orientation of windows with correspondence to sensors), and the new variable *I* (internal finish). The following are the main findings of this study:

- All generated EML models performed very well with a

minimum coefficient of determination of R^2 = 0.88 for RF models, and R^2 = 0.972 for XGBoost models. These models are chosen after tuning the hyperparameters.

- The UDI-*a* model is the best-performing model among all with R^2 = 0.992. This is due to the completeness of the dataset, including a wide range of illuminance values. In addition, UDI-a performs the second best (R^2 = 0.991), which can be explained by the unique relationship between this model and windows. The areas immediately around windows are exposed to daylight almost all day. Hence, the pattern of this model is efficiently captured by the EML models.
- The developed framework is generalizable since it is open to introducing new features in different cases and the ability to choose efficient ML algorithms that need reasonable computational resources.

Some limitations of the current study should be highlighted. Firstly, specific weather data for a southern hemisphere region has been used to generate the training dataset. Therefore, different locations need other datasets generated based on their weather data. Second, the developed frameworks generated test rooms that are limited in area, windows, and number of zones. Third, the training dataset used to develop the models are synthetic data (simulated). Finally, the outcomes of the developed models in their current state are only useful for early-phase qualitative judgments. Designers can infer general illuminance patterns and potential discomfort situations using the proposed framework. The framework is not ready for precise illuminance predictions of a single sensor.

Future research can consider a higher level of complexity in generating building layouts. The utilized method in this study forms standard layouts by overlapping four squares. Moreover, additional features can be introduced to the training dataset to improve prediction performance in complex designs.

Appendix A

Figure A1 demonstrates the 4-square method employed in this study. In this method, 4 squares (A, B, C, and D) move clockwise in an increment of 1 mm with a range of [0–2000 mm]. This movement enables the formation of any regular plan (2 examples are shown where red dots are the original positions centres, and blue dots are the centres after random movements are applied). Random number of windows ranging of [1–4] is populated to the output building layouts at random locations (1 window on the left side example, 4 windows on the right side example). Combined with 4-square method, a total of 625 unique building layouts were generated.

Figure A2 illustrates the variable *x* "a spatial component".

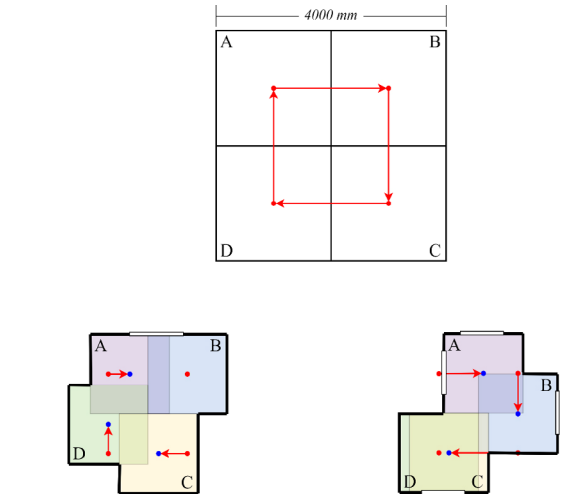

Fig. A1 4-square method to generate building layouts with random number of windows (1–4)

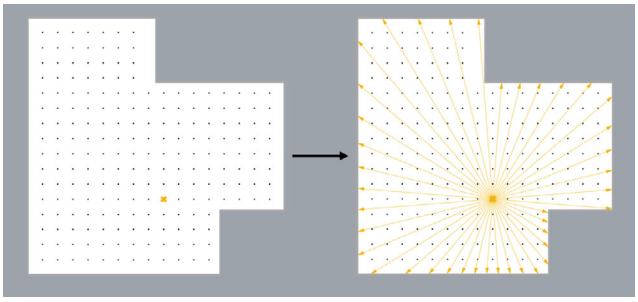

Fig. A2 An illustration of the variable x for a sensor in a building layout

It denotes 60 distances on a 360° space plan from the light sensor to any obstacle surrounds it.

Data availability

Data will be made available on reasonable request.

Acknowledgements

The authors are grateful for support from the Australian Research Council (ARC) through the Linkage Infrastructure, Equipment and Facilities (LE210100019). The assistance of the ASCII Lab members at Monash University is greatly appreciated.

Funding note: Open Access funding enabled and organized by CAUL and its Member Institutions.

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.

Author contribution statement

Alsharif R., Arashpour M., and Golafshani E. conceived and planned the experiments. Alsharif R., Arashpour M., and Golafshani E. carried out the experiments. Alsharif R. and Golafshani E. planned and carried out the machine learning phase. Alsharif R. and Arashpour M. contributed to sample preparation. Alsharif R., Arashpour M., Golafshani E., Bazli M., and Mohandes S. contributed to the interpretation of the results. Alsharif R. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

References

Alsharif R, Arashpour M, Golafshani EM, et al. (2022). Machine learning-based analysis of occupant-centric aspects: critical elements in the energy consumption of residential buildings. *Journal of Building Engineering*, 46: 103846.

Arashpour M, Ngo T, Li H (2021). Scene understanding in construction and buildings using image processing methods: a comprehensive review and a case study. *Journal of Building Engineering*, 33: 101672.

Arashpour M, Kamat V, Heidarpour A, et al. (2022). Computer vision for anatomical analysis of equipment in civil infrastructure projects: Theorizing the development of regression-based deep neural networks. *Automation in Construction*, 137: 104193.

Arashpour M (2023). AI explainability framework for environmental management research. *Journal of Environmental Management*, 342: 118149.

Arashpour M, Golafshani EM, Parthiban R, et al. (2023). Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education*, 31: 83–99.

Arbab M, Rahbar M, Arbab M (2021). A comparative study of artificial intelligence models for predicting interior illuminance. *Applied Artificial Intelligence*, 35: 373–392.

Ayoub M (2020). A review on machine learning algorithms to predict daylighting inside buildings. *Solar Energy*, 202: 249–275.

Belitz K, Stackelberg PE (2021). Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling & Software*, 139: 105006.

Blackwell B (2002). Light exposure to sensitive artworks during digital photography. *Spectra*, 26(2): 24–28.

Brembilla E, Drosou NC, Mardaljevic J (2022). Assessing daylight performance in use: A comparison between long-term daylight measurements and simulations. *Energy and Buildings*, 262: 111989.

Carlucci S, Causone F, De Rosa F, et al. (2015). A review of indices for assessing visual comfort with a view to their use in optimization processes to support building integrated design. *Renewable and Sustainable Energy Reviews*, 47: 1016–1033.

Chen T, Guestrin C (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.

Chen Y, Chen B, Deng J, et al. (2022). The integration model of objective and subjective data of residential indoor environment quality in Northeast China based on structural equation modeling. *Building Simulation*, 15: 741–754.

Chi DA (2022). Solar energy density as a benchmark to improve daylight availability and energy performance in buildings: A single metric for a single-objective optimization. *Solar Energy*, 234: 304–318.

Davoodi A, Johansson P, Aries M (2020). The use of lighting simulation in the evidence-based design process: A case study approach using visual comfort analysis in offices. *Building Simulation*, 13: 141–153.

Day JK, Futrell B, Cox R, et al. (2019). Blinded by the light: Occupant perceptions and visual comfort assessments of three dynamic daylight control systems and shading strategies. *Building and Environment*, 154: 107–121.

Fang J, Zhao Y, Tian Z, et al. (2022). Analysis of dynamic louver control with prism redirecting fenestrations for office daylighting optimization. *Energy and Buildings*, 262: 112019.

Ghobad L, Glumac S (2018). Daylighting and energy simulation workflow in performance-based building simulation tools. In: Proceedings of the 2018 Building Performance Analysis Conference and Simbuild, Chicago, IL, USA.

González S, García S, Del Ser J, et al. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64: 205–237.

Gunay HB, O'Brien W, Beausoleil-Morrison I, et al. (2017). Development and implementation of an adaptive lighting and blinds control algorithm. *Building and Environment*, 113: 185–199.

He Q, Li Z, Gao W, et al. (2021). Predictive models for daylight performance of general floorplans based on CNN and GAN: A proof-of-concept study. *Building and Environment*, 206: 108346.

Khidmat RP, Fukuda H, Paramita B, et al. (2022). Investigation into the daylight performance of expanded-metal shading through parametric design and multi-objective optimisation in Japan. *Journal of Building Engineering*, 51: 104241.

Le-Thanh L, Nguyen-Thi-Viet H, Lee J, et al. (2022). Machine learning-based real-time daylight analysis in buildings. *Journal of Building Engineering*, 52: 104374.

Lin C-H, Tsay Y-S (2021). A metamodel based on intermediary features for daylight performance prediction of façade design. *Building and Environment*, 206: 108371.

Liu G, Qu G, Ren L, et al. (2022). The influence mechanism of daylight visual evaluation in college classrooms under visual field physiological characteristics of student group: Case study. *Building and Environment*, 209: 108655.

Lou S, Li DHW, Lam JC, et al. (2016). Prediction of diffuse solar irradiance using machine learning and multivariable regression. *Applied Energy*, 181: 367–374.

Luo Z, Sun C, Dong Q, et al. (2022). Key control variables affecting interior visual comfort for automated louver control in open-plan office—A study using machine learning. *Building and Environment*, 207: 108565.

Lv Y, Peng H, He M, et al. (2019). Definition of typical commercial building for South China's Pearl River Delta: local data statistics and model development. *Energy and Buildings*, 190: 119–131.

Manfren M, James PA, Tronchin L (2022). Data-driven building energy modelling - An analysis of the potential for generalisation through interpretable machine learning. *Renewable and Sustainable Energy Reviews*, 167: 112686.

Mardaljevic J, Andersen M, Roy N, et al. (2012). Daylighting metrics: Is there a relation between useful daylight illuminance and daylight glare probabilty? In: Proceedings of the 1st Building Simulation and Optimization Conference, Loughborough, UK.

Michael A, Heracleous C (2017). Assessment of natural lighting performance and visual comfort of educational architecture in Southern Europe: The case of typical educational school premises in Cyprus. *Energy and Buildings*, 140: 443–457.

Montaser Koohsari A, Heidari S (2022). Subdivided venetian blind control strategies considering visual satisfaction of occupants, daylight metrics, and energy analyses. *Energy and Buildings*, 257: 111767.

Nabil A, Mardaljevic J (2005). Useful daylight illuminance: a new paradigm for assessing daylight in buildings. *Lighting Research & Technology*, 37: 41–57.

Ngarambe J, Adilkhanova I, Uwiragiye B, et al. (2022). A review on the current usage of machine learning tools for daylighting design and control. *Building and Environment*, 223: 109507

Oyedele A, Ajayi A, Oyedele LO, et al. (2021). Deep learning and Boosted trees for injuries prediction in power infrastructure projects. *Applied Soft Computing*, 110: 107587.

Papadopoulos S, Azar E, Woon WL, et al. (2018). Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *Journal of Building Performance Simulation*, 11: 322–332.

Park Y, Ho JC (2021). Tackling overfitting in boosting for noisy healthcare data. *IEEE Transactions on Knowledge and Data Engineering*, 33: 2995–3006.

Peng H, Li M, Lou S, et al. (2020). Investigation on spatial distribution and thermal properties of typical residential buildings in South China's Pearl River Delta. *Energy and Buildings*, 206: 109555.

Pérez-Fargallo A, Rubio-Manzano C, Martínez-Rocamora A, et al. (2018). Linguistic descriptions of thermal comfort data for buildings: Definition, implementation and evaluation. *Building Simulation*, 11: 1095–1108.

Shekar BH, Dagnew G (2019). Grid search-based hyperparameter tuning and classification of microarray cancer data. In: Proceedings of 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India.

Shi L, Zhang Y, Wang Z, et al. (2021). Luminance parameter thresholds for user visual comfort under daylight conditions from subjective responses and physiological measurements in a gymnasium. *Building and Environment*, 205: 108187.

Thrampoulidis E, Mavromatidis G, Lucchi A, Orehounig K (2021). A machine learning-based surrogate model to approximate optimal building retrofit solutions. *Applied Energy*, 281: 116024.

Veloso B, Gama J, Malheiro B, et al. (2021). Hyperparameter self-tuning for data streams. *Information Fusion*, 76: 75–86.

Wagiman KR, Abdullah MN, Hassan MY, et al. (2021). A new metric for optimal visual comfort and energy efficiency of building lighting system considering daylight using multi-objective particle swarm optimization. *Journal of Building Engineering*, 43: 102525.

Wang Z, Yu H, Luo M, et al. (2019). Predicting older People's thermal sensation in building environment through a machine learning approach: Modelling, interpretation, and application. *Building and Environment*, 161: 106231.

Yang L, Shami A (2020). On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*, 415: 295–316.

Yngvesson L, Adolfsson E (2018). The impact of scale when using models of daylight analysis. Jönköping University, Sweden.

Zhang K, Yang J, Sha J, et al. (2022). Dynamic slow feature analysis and random forest for subway indoor air quality modeling. *Building and Environment*, 213: 108876.