BUILD SIMUL (2023) 16: 915–925
https://doi.org/10.1007/s12273-022-0976-y

Machine learning approach for estimating the human-related VOC emissions in a university classroom

Jialong Liu[1], Rui Zhang[1], Jianyin Xiong[1,2,3] (✉)

1. School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China
2. Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA
3. State Key Laboratory of Green Building in Western China, Xi'an University of Architecture and Technology, Xi'an 710055, China

Abstract

Indoor air quality becomes increasingly important, partly because the COVID-19 pandemic increases the time people spend indoors. Research into the prediction of indoor volatile organic compounds (VOCs) is traditionally confined to building materials and furniture. Relatively little research focuses on estimation of human-related VOCs, which have been shown to contribute significantly to indoor air quality, especially in densely-occupied environments. This study applies a machine learning approach to accurately estimate the human-related VOC emissions in a university classroom. The time-resolved concentrations of two typical human-related (ozone-related) VOCs in the classroom over a five-day period were analyzed, i.e., 6-methyl-5-hepten-2-one (6-MHO), 4-oxopentanal (4-OPA). By comparing the results for 6-MHO concentration predicted via five machine learning approaches including the random forest regression (RFR), adaptive boosting (Adaboost), gradient boosting regression tree (GBRT), extreme gradient boosting (XGboost), and least squares support vector machine (LSSVM), we find that the LSSVM approach achieves the best performance, by using multi-feature parameters (number of occupants, ozone concentration, temperature, relative humidity) as the input. The LSSVM approach is then used to predict the 4-OPA concentration, with mean absolute percentage error (MAPE) less than 5%, indicating high accuracy. By combining the LSSVM with a kernel density estimation (KDE) method, we further establish an interval prediction model, which can provide uncertainty information and viable option for decision-makers. The machine learning approach in this study can easily incorporate the impact of various factors on VOC emission behaviors, making it especially suitable for concentration prediction and exposure assessment in realistic indoor settings.

Keywords

indoor air quality;
human-related VOCs;
machine learning;
interval prediction;
least squares support vector machine (LSSVM);
kernel density estimation (KDE)

Article History

Received: 30 September 2022
Revised: 17 November 2022
Accepted: 06 December 2022

1 Introduction

Modern people spend most of their time in indoor environments, and the work-from-home policy caused by the COVID-19 epidemic further increases the time people spend indoors (Klepeis et al. 2001; Galanti et al. 2021). Currently, poor indoor air quality is a severe threat to human health (WHO 2007; Salthammer et al. 2010; Landrigan et al. 2018; Abouleish 2021; Bu et al. 2021; Cui et al. 2022). Previous studies on indoor pollution tended to focus on pollutants introduced by outdoor air, or from building materials, such as particulate matter, formaldehyde, benzene, and some other volatile organic compounds (VOCs), while relatively few studies have been conducted on human-related emissions (Weschler 2009; Tang et al. 2015; He et al. 2019; Tian et al. 2021; NASEM 2022; Zhao et al. 2022). Amann et al. (2014) reported a variety of VOCs appeared in exhaled breath and skin emanations. A study by Tang et al. (2016) showed that human-emitted VOCs accounted for about 60% of the total VOC mass in a well-ventilated classroom. Occupants can also pollute indoor air due to the use of personal care products (Yang et al. 2018a). Wisthaler and Weschler (2010) found the reaction of squalene in human skin with indoor ozone can generate some primary products

E-mail: xiongjy@bit.edu.cn

such as 6-methyl-5-hepten-2-one (6-MHO), and secondary products such as 4-oxopentanal (4-OPA). 6-MHO and 4-OPA have been confirmed to irritate the human digestive tract and respiratory tract, causing allergies, while some other ozone/squalene products can lead to comedogenic skin, inflammatory acne, and other skin diseases (Fruekilde et al. 1998; Jarvis et al. 2005; Anderson et al. 2012; Pham et al. 2015; Wolkoff et al. 2016). These studies all show that the presence of occupants can significantly impact indoor air quality, especially in densely-populated indoor settings.

At present, the emissions of VOCs from building materials and furniture are well characterized (Little et al. 1994; Yang et al. 2001; Xiong et al. 2011; Liu et al. 2013; Zhang et al. 2016; Zhou et al. 2018; Wang et al. 2022; Hu et al. 2023). As for the ozone-initiated human VOC emissions, the most common method is to use mechanistic-based models for concentration prediction. Lakey et al. (2017) developed a kinetic multilayer model to describe the reaction characteristics of ozone with squalene, and achieved good prediction results in a simulated office. This model was further improved by incorporating the impact of clothing (Lakey et al. 2019). Moreover, a physical-chemical model considering the in-body/off-body ozone/squalene reactions, external convection along the skin surface, internal diffusion inside the skin, indoor surface uptake was proposed to more systematically predict the characteristics of reaction products in different phases (Zhang et al. 2021a). Generally, the prediction performance of mechanistic-based models depend significantly on the availability and accuracy of various key transport parameters in the model, and it is often a challenging problem to determine these parameters, especially for realistic indoor settings. Moreover, ideal assumptions are usually made to simplify complicated scenarios, which will further impact the prediction accuracy, especially when the indoor environmental conditions change over time.

A common problem in today's scientific community is that, the ability to collect and generate observational data far outstrips the ability to absorb and interpret it (Reichstein et al. 2019). Machine learning can extract relevant information and knowledge from various data streams in a data-driven manner, allowing it to comprehend the rules behind the things. Recently, machine learning and mathematical statistics have been widely applied in the environmental field (Wei et al. 2019). Machine learning approaches include classical algorithms such as perceptron and decision trees, ensemble methods with good robustness, and artificial neural networks (ANN). Although ANN has absolute advantages in the fields of visual and audio recognition, they may not perform well for standard regression prediction issues.

In the indoor field, the most often examined contaminant using machine learning is particulate matter (PM) (Wei et al. 2019). Park et al. (2018) adopted a feed-forward back-propagation network to predict PM_{10} concentration in Seoul subway stations. In addition, ensemble learning has recently been used to forecast indoor PM concentration (Yuchi et al. 2019; Xu et al. 2020; Li et al. 2021). The Japan Environment and Children's Study Programme Office used random forest regression (RFR) to conduct detailed studies of variables affecting indoor PM (Nishihama et al. 2021). There are also lots of studies on indoor CO_2 that use machine learning approaches. Khazaei et al. (2019) and Skön et al. (2012) used fully connected networks to evaluate CO_2 levels in dwellings. Taheri and Razban (2021) proposed an energy-saving ventilation strategy based on the prediction of indoor CO_2 through support vector regression (SVM) and other models. Kallio et al. (2021) studied the influence of different feature inputs and prediction periods on indoor CO_2 prediction. Indoor radon has been studied by kernel regression and Bayesian spatial quantile regression in Switzerland (Kropat et al. 2015). Indoor formaldehyde and VOC predictions using statistical models are still sparse, and are limited to a few methods (Wei et al. 2019). Chen et al. (2018) revealed that among several machine learning methods, SVM had the best prediction performance for CO_2 and TVOC in the classroom, whereas formaldehyde was difficult to predict. Zhang et al. (2021b, 2022) used the back propagation (BP) network and long short-term memory network (LSTM) to predict concentrations of some VOCs emitted from furniture in a controlled chamber under different conditions. In that study (Zhang et al. 2022), the machine learning approach to ozone-initiated VOCs only employed a single-feature LSTM model to predict 6-MHO and 4-OPA concentrations in a classroom. In fact, VOC emissions in a classroom are likely to be affected by multiple factors, such as number of occupants, ozone concentration, temperature and humidity. Since the number of occupants were not considered in Zhang et al.'s approach, the predictions deviated from the observations when occupancy changed greatly. Adding more relevant factor inputs will improve the prediction performance (Chen et al. 2018).

In addition, the mechanistic-based or statistical models mentioned above can only deliver point prediction, and cannot show the probability and fluctuation range of the predicted results. Interval models can provide an effective range in which pollutant output lies with a specified probability. Therefore, combining a point prediction model with an interval prediction model will provide more useful information for decision-makers. Interval prediction includes parameter estimation, kernel density estimation, and some other estimation methods (Zhang et al. 2014). There are currently few studies using interval prediction for indoor air pollutants, with a primary focus on PM in the atmosphere (Song et al. 2015; Xu et al. 2017).

In this study, we set out to use machine learning

approaches to achieve rapid and accurate predictions of indoor human-related VOCs. 6-MHO and 4-OPA concentrations due to the reactions of ozone with squalene in a university classroom were predicted by five machine learning approaches under different feature combinations, with the least squares support vector machine (LSSVM) achieving the best performance. The LSSVM was then combined with a kernel density estimation method, to construct a robust prediction system for indoor human-related VOCs, which could provide decision-makers with more useful information about indoor air quality.

2 Methodology

2.1 Introduction of some typical machine learning approaches

The indoor VOC concentration is often affected by multiple factors, and traditional machine learning models such as linear regression and regression tree are difficult to capture the concentration changes accurately. Compared with traditional models, ensemble models are flexible in solving complex non-linear regression problems (Sagi and Rokach 2018). Although ANN are powerful, they tend to be relatively complex, and require more experience to tune the model and more computing time. In this study, we select random forest regression (RFR), adaptive boosting (Adaboost), gradient boosting regression tree (GBRT) and extreme gradient boosting (XGboost), as representative ensemble machine learning models. Moreover, a prior study demonstrated that SVM had good performance for indoor TVOC prediction (Chen et al. 2018). Inspired by this work, we also select an improved SVM, the least squares SVM (LSSVM) approach for analysis, which inherits the advantages of SVM and can also enhance the speed and accuracy. Since the mathematical underpinning of each approach (or model or algorithm) is very complicated, we give only a brief introduction to the core concepts or formulas of each approach here.

(1) Random forest regression (RFR)

RFR was proposed by Breiman (2001), the creator of the bagging method. It samples the training dataset and builds multiple independent tree models through the bootstrap method. For the regression problem, the final output is the average of all the basic learner results. RFR has the advantages of simple principle, easy implementation, and low computational cost. Furthermore, using a tree model as the base learner makes it easier to interpret.

(2) Adaptive boosting (Adaboost)

As an adaptive ensemble learning algorithm, Adaboost can automatically increase the weight of mis-predicted samples in an iterative manner, so that the basic learner can be adjusted according to the prediction performance of the current model (Domingo and Osamu 2000). Thus, a set of basic learners with complementary performance can be obtained. Finally, an ensemble model with better performance will be constructed by means of a weighted average. Adaboost can freely choose basic learners, e.g., linear regression, perceptron. In this study, the regression tree is selected as the basic learner to construct the Adaboost ensemble model for prediction.

(3) Gradient boosting regression tree (GBRT)

GBRT is another typical algorithm of the boosting class, where, for a given training set (x_i, y_i), GBRT generates a new base learner $f_m(x)$ by fitting the current model residuals $r_{m,i} = y_i - f_{m-1}(x_i)$. The estimation model for GBRT can be expressed as (Friedman 2001):

$$f_M(x) = \sum_{m=1}^{M} R(x;\theta_m) \tag{1}$$

where, $R(x; \theta_m)$ represents the regression tree model; θ_m are the regression tree parameters; M is the number of trees.

(4) Extreme gradient boosting (XGboost)

The XGboost algorithm was proposed by Chen and Guestrin (2016). As an improved form of the GBRT algorithm, it still adopts the tree model as the base estimator, and uses the forward step algorithm to build the ensemble model. Since XGboost explicitly adds the tree model complexity as a regularization term to the optimization objective, the model's ability to resist over-fitting is greatly enhanced. In addition, the traditional GBRT only uses the first-order derivative information when optimizing the loss function, while XGboost performs the second-order Taylor expansion of the objective function and uses the first-order and second-order derivative information at the same time, which makes the model establishment more accurate. Its loss function for the t-th iteration is expressed as:

$$obj^{(t)} = \sum_{i=1}^{N}\left[l(y_i, \overline{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2}$$

where, $l(y_i, \overline{y}_i^{(t-1)})$ is the loss function of the $(t-1)$th iteration; f_t is the newly created tree model for this iteration; g_i, h_i are the first and second order gradient statistics of the loss function, respectively; $\Omega(f_t)$ is the tree model complexity function, expressed as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{3}$$

where, λ and γ are regularization coefficients; T is the number of leaves; ω_j is the weight of the leaves.

(5) Least square SVM (LSSVM)

SVM was pioneered by Cortes and Vapnik (1995). The core idea is to find support vectors to maximize the interval, and formalize the problem as a solvable convex quadratic programming problem. Further, SVM maps low-dimensional features to high-dimensional space by a kernel trick, which can effectively solve complex nonlinear problems and greatly reduce the computational difficulty. The LSSVM model can be obtained by replacing the inequality constraints of the SVM optimization problem with equality constraints and introducing the L2 regular term of the sample error e_i, as follows:

$$\begin{aligned} &\arg\min_{W,b} \frac{1}{2}\|\boldsymbol{W}\|^2 + \frac{\lambda}{2}\sum_{i=1}^{N} e_i^2 \\ &\text{subject to} \quad \boldsymbol{W}^{\mathrm{T}} \cdot \varphi(x_i) + b + e_i, \quad i = 1,2,...,N \end{aligned} \tag{4}$$

where, φ is the nonlinear mapping function, which maps the training data into a higher dimensional linear feature space; λ is the regularization coefficient; e_i is the sample error; $\boldsymbol{W}$ is weight vector; b is the bias.

To solve the LSSVM optimization problem, a Lagrange function is constructed as:

$$\begin{aligned} L(\boldsymbol{W},b,e,\alpha) = &\frac{1}{2}\|\boldsymbol{W}\|^2 + \frac{\lambda}{2}\sum_{i=1}^{N} e_i^2 \\ &-\sum_{I=1}^{N} \alpha_i \left\{ \boldsymbol{W}^{\mathrm{T}} \cdot \varphi(x_i) + b + e_i - y_i \right\} \end{aligned} \tag{5}$$

where, α_i are Lagrange multipliers.

A more specific mathematical formulation and explanation for LSSVM can be obtained in the literature (Wang and Hu 2005).

2.2 Interval prediction based on kernel density estimation

Kernel density estimation (KDE) is a nonparametric density estimation technique, for looking at data distribution by using only the given samples without any assumption of prior distribution (Yang et al. 2018b). Compared with the parametric method, the KDE method can describe the distribution of data more accurately and is more reliable. Therefore, this study uses the KDE method to establish an error distribution model of the predicted VOC concentrations, so as to realize the interval prediction.

For data series $\{x_1, x_2, ..., x_i\}$, the probability density function (PDF) calculated by KDE is:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

The cumulative distribution function (CDF) is expressed as:

$$F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z = (nh)^{-1}\int_{-\infty}^{x}\sum_{i=1}^{n} K\left(\frac{z - X_i}{h}\right)\mathrm{d}z \tag{7}$$

where, n is the length of offered datasets; $K(x)$ is the kernel function; h is the bandwidth for adjusting the width of the probability density curve. This study uses a Gaussian kernel function and the reference method for bandwidth selection, i.e., $h \approx 1.06\sigma n^{-0.2}$, σ is the standard deviation of offered samples.

2.3 Metrics for evaluating the model performance

There are various indicators that can be used to evaluate the performance of a point prediction model and an interval prediction model. To conduct a more comprehensive evaluation of the prediction performance of different models, two metrics are used, i.e., the mean absolute percentage error (MAPE), and the coefficient of determination (R^2), which are defined as:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{\mathrm{exp},i} - y_{\mathrm{pre},i}}{y_{\mathrm{exp},i}}\right| \times 100\% \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{\mathrm{exp},i} - y_{\mathrm{pre},i})^2}{\sum_{i=1}^{N}(y_{\mathrm{exp},i} - \overline{y})^2} \tag{9}$$

where, $y_{\mathrm{exp},i}$ is experimental data; $y_{\mathrm{pre},i}$ is predicted data; $\overline{y}$ is the average of the experimental data.

MAPE is a scale-independent metric, which is suitable for comparing the prediction accuracy between different pollutants. The value range is (0, +∞), and a MAPE less than 20% represents a decent prediction. R^2 reflects the variance between the predicted value and the real value, with values varying from 0 to 1. The closer R^2 is to 1, the better the model prediction performance is.

Another two metrics are used to assess the interval prediction performance, i.e., the IF coverage probability (IFCP), and the IF average width (IFAW), which are defined as:

$$\mathrm{IFCP} = \frac{1}{N}\sum_{i=1}^{N} c_i \tag{10}$$

$$\mathrm{IFAW} = \frac{1}{N}\sum_{i=1}^{N}(U_i - L_i) \tag{11}$$

where, c_i is the Boolean value, 1 means within the prediction interval, 0 means outside the prediction interval; U_i is the upper bound of the prediction interval; L_i is the lower bound of the the prediction interval. The interval

prediction model's performance improves as IFCP gets closer to 1.

3 Results and discussion

3.1 Parameter settings for different machine learning approaches

The experiments were performed over five weekdays in a densely-occupied realistic indoor setting, in this case a typical university classroom. Ambient ozone was introduced directly from the outside through a single-pass ventilation system. The number of occupants in the classroom was recorded manually. A proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS) was used to measure VOC concentrations. The experimental data for 6-MHO, 4-OPA, ozone and CO_2 concentrations, as well as the temperature, relative humidity, number of occupants, are used for this analysis. More detailed descriptions of the classroom's characteristics and experimental protocol can be found in previous publications (Tang et al. 2016; Yang et al. 2018a; Xiong et al. 2019).

The parameters of machine learning approaches are divided into: hyperparameters that need to be set artificially in advance (e.g., learning rate) and weight parameters that the algorithm automatically learns and adjusts based on training datasets. To make full use of the dataset to select the optimal hyperparameters, the k-fold cross validation method is often adopted. However, considering that the data of this study has the time-series characteristics, the use of k-fold cross validation will lead to the problem of using future data for training and past data for validation. Therefore, we use a combination of time-series cross validation and grid search to obtain the optimal hyperparameters. Time-series cross validation is a statistical validation technique used to evaluate the performance of models in machine learning, and grid search is a way of tuning parameters. Additional information about these functions and their implementations is available in Yasin et al. (2016). We performed parameter optimization for each feature combination, and Table 1 lists the optimal combination of hyperparameters for the five different machine learning approaches based on 6-MHO concentration. The experimental data from the first four days were used for training and optimization, and the data from the fifth day were used for evaluating the prediction performance.

The above five machine learning approaches use different combinations of features for prediction (see detail in the following Figure 2), including the VOC concentration (6-MHO), occupancy, ozone concentration, temperature and relative humidity. These features are normalized between −1 and 1 by using Z-score standardization. This study uses scikit-learn (Pedregosa et al. 2011) as a tool for implementing machine learning algorithms.

Table 1 Parameter settings of the different machine learning approaches

Machine learning approach	Name of parameter	Set value
RFR	Number of estimators	40
	Max depth	3
Adaboost	Base estimator	DecisionTreeRegressor
	Number of estimators	60
	Max depth	3
	Learning rate	0.05
GBRT	Number of estimators	80
	Max depth	4
	Learning rate	0.01
XGboost	Number of estimators	60
	Max depth	1
	Gamma	0.01
	Learning rate	0.1
	Booster	gbtree
LSSVM	Kernel	linear
	Penalty parameter C	1

3.2 Correlation analysis of the experimental data

Temperature, relative humidity, number of occupants, ozone concentration, CO_2 concentration, 6-MHO and 4-OPA concentration were all recorded in the classroom over the five-workday period. The correlation matrix between 6-MHO concentration and other related features is shown in the following heat map (Figure 1).

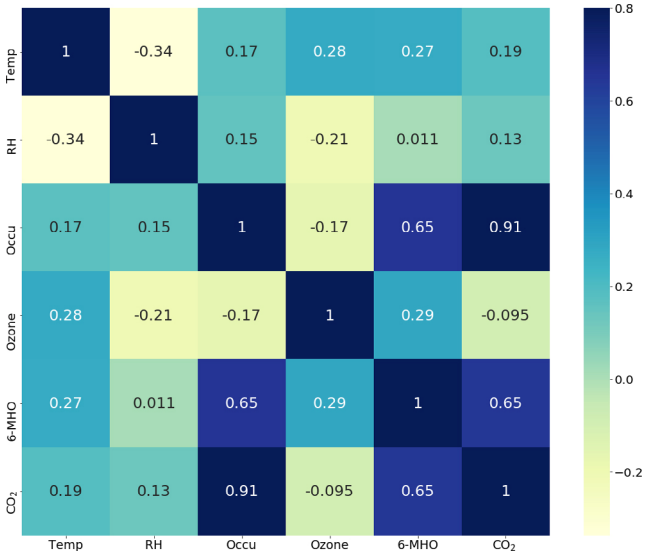

Fig. 1 Heat map representation of the correlation matrix

The correlation (η) between different features is calculated as (Taheri and Razban 2021):

$$\eta = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\mathrm{var}(x_i) \cdot \mathrm{var}(x_j)}} \quad (12)$$

where, cov(x_i, x_j) is the covariance of two features; var(x_i), var(x_j) are the variances of the two features, respectively.

If the η of two features is greater than 0.7 or less than −0.7, they can be considered to be the same feature, and one of them needs to be removed (Géron 2022). Generally, the human-related VOC concentration should be related to the number of occupants (occupancy), as well as the CO_2 concentration in the classroom. Our results indicate that the change of occupancy and CO_2 concentration is very similar, with η between them reaching 0.91 (shown in Figure 1). Therefore, these two features (occupancy, CO_2 concentration) is essentially one feature and should be combined. Since the main indoor VOCs in this study are produced by the reactions between squalene in human skin and ozone, we select the number of occupants (equivalent to CO_2 concentration) as a feature for analysis. In addition, when the machine learning algorithm incorporates the features most associated with the prediction target, the prediction performance of the established model will be improved (Skön et al. 2012). Figure 1 indicates that all the features positively correlate with 6-MHO concentration, with number of occupants being the most-correlated feature, with η of 0.65, followed by ozone concentration, temperature, and relative humidity, with η of 0.29, 0.27 and 0.011, respectively. For 4-OPA, similar results are obtained. Considering that the heat map is very similar for 6-MHO and 4-OPA, the heat map of 4-OPA is not given here. In the following sections, number of occupants is shown to be strongly correlated with the prediction of the target VOCs. Furthermore, due to the complexity and difference between different algorithms, the influence of adding the same feature is varied.

3.3 Prediction comparison for different machine learning approaches

In this section, we examine the impact of different feature combinations on the prediction performance of the five machine learning approaches when predicing 6-MHO concentration, so as to select the appropriate approach and feature combination. Since the performance of machine learning approaches is heavily dependent on the quantity and quality of the training data, models can often accurately imitate the rule of training data while performing poorly in prediction. In general, effective measures for overcoming model overfitting and improving their generalization ability, are to increase the amount of training data or to introduce additional beneficial features. In this study, the indoor 6-MHO and 4-OPA are produced by the reactions of ozone with squalene in human skin oils. Therefore, ozone levels and occupants will significantly affect the production rate, which may also be impacted by indoor temperature and humidity. Thus, the number of occupants, target VOC concentration, ozone concentration, temperature, and relative humidity are selected as useful features for this study. Then, according to the correlation coefficient obtained in Section 3.2, five feature combinations are obtained by adding these features successively from high to low (η with 6-MHO concentration).

Figure 2 shows the prediction performance of the five machine learning approaches for different feature combinations, which is evaluated by MAPE for the test data of 6-MHO concentration. This figure indicates that the performances of all approaches have been improved by adding the number of occupants, and the improvement in the performance of LSSVM and GBRT are very significant. When the other three features are added, however, the performance of the five approaches doesn't indicate the same tendency. In the case of LSSVM, adding the above four features can improve the results to a certain extent, and the degree of improvement is positively correlated with the correlation coefficient. However, the performance of GBRT gradually deteriorates after adding ozone concentration, temperature, and relative humidity features. When modelling with combinations of all five features, LSSVM has the best performance, with a MAPE as low as 8.9%, while the results of GBRT deteriorate to achieve the same as univariate prediction. The results of the remaining three approaches are similar, but all of them are better than using only a single

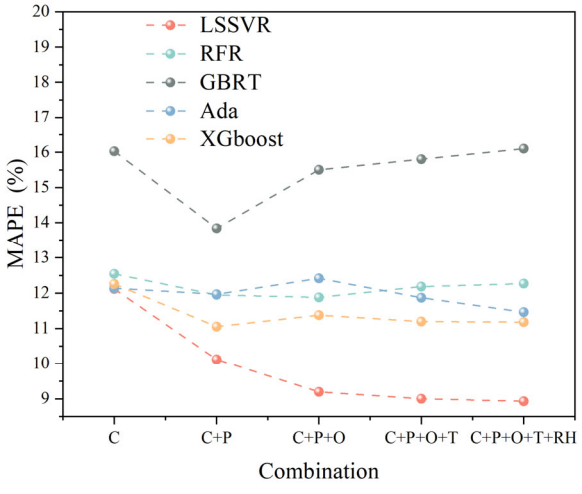

Fig. 2 Evaluation of the performance of five machine learning approaches under different feature combinations by MAPE (C: concentration, P: number of occupants, O: ozone, T: temperature, RH: relative humidity)

variable for prediction. Generally speaking, a certain feature has diverse effects for different machine learning approaches. The RFR is insensitive to the above features, while LSSVM can learn more useful information based on each feature.

To further assess the prediction performance, another metric, R^2, is also used. Figure 3 shows that the R^2 for most of the machine learning approaches follow similar trends. By adding the feature of number of occupants, the R^2 of all approaches increases by about 6%–14%. The highest R^2 is 0.886, achieved by the LSSVM using all features, while the lowest is 0.69, obtained by GBRT. Based on the above analysis with MAPE and R^2, we select the LSSVM approach using multi-feature parameters (number of occupants, ozone concentration, temperature and relative humidity) as input, to establish a prediction model to estimate the concentrations of human-related VOCs.

3.4 Prediction of 6-MHO and 4-OPA concentrations with LSSVM

According to the analysis in Section 3.3, we chose to use the multi-feature LSSVM approach to establish prediction models for 6-MHO and 4-OPA concentations in the classroom. Results indicate that the MAPEs between model predictions and experimental data for 6-MHO and 4-OPA are 8.93% and 4.98%, and the R^2 are 0.886 and 0.806, respectively, demonstrating high accuracy.

Figure 4 provides a visual comparision between the LSSVM model predictions and the experimental data for 6-MHO and 4-OPA concentrations on the fifth testing day. This figure indicates that the LSSVM approach can fairly accurately capture the VOC concentration profiles, with relatively small deviations. The time period where prediction is poor, is between 11:10 and 13:40. According to experimental records, the number of students in the classroom fluctuated greatly during this period, which resulted in a significant change in the VOC concentration. It is difficult for an LSSVM to accurately learn this sudden change. For the time periods when the number of students is relatively stable, the LSSVM works well in taking into account the change pattern of the predictor.

To examine the applicability of the LSSVM approach in other realistic indoor settings, we analyze the data in an occupied residence in a recent study (Zhang et al. 2021a). Figure 5 shows the comparison of model predictions with field measurements for 6-MHO and 4-OPA in a bedroom

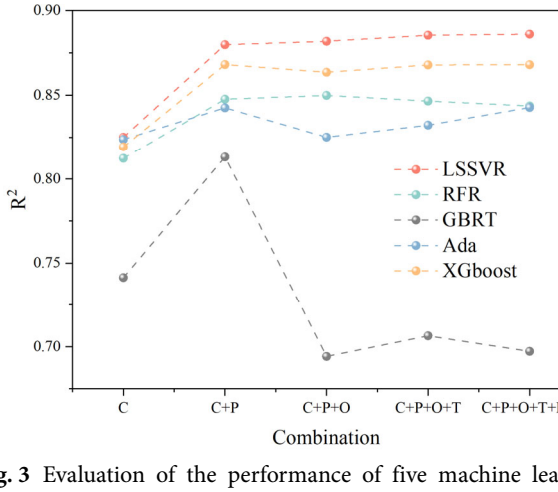

Fig. 3 Evaluation of the performance of five machine learning approaches under different feature combinations by R^2 (C: concentration, P: number of occupants, O: ozone, T: temperature, RH: relative humidity)

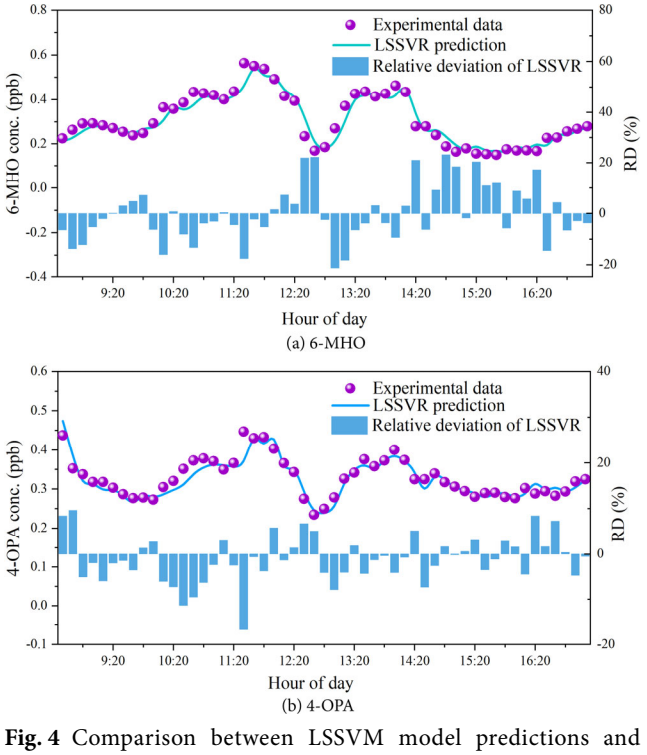

Fig. 4 Comparison between LSSVM model predictions and experimental data for (a) 6-MHO and (b) 4-OPA concentrations on the fifth testing day

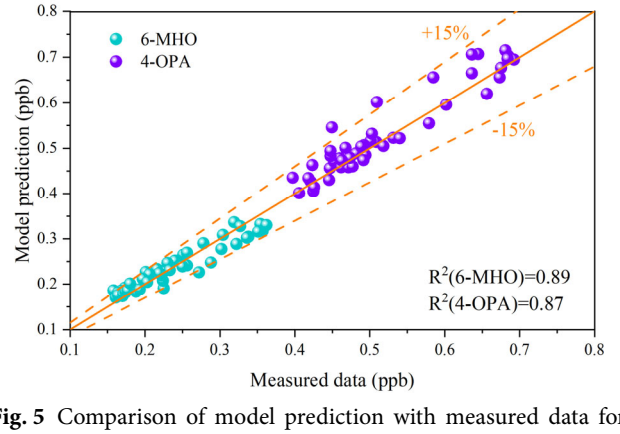

Fig. 5 Comparison of model prediction with measured data for 6-MHO and 4-OPA in a residence in literature for one day

of the residence. We can see that most of the data are located at the line $y = x$, implying good consistency between them. This analysis demonstrates the universality of the present model for different indoor environments.

3.5 Comparison of LSSVM with LSTM predictions

In a prior study, a long short-term memory (LSTM) approach was applied to predict the VOC concentrations in a classroom (Zhang et al. 2022). As a type of deep learning approach, LSTM incorporates a gating mechanism (forgetting gate, input gate, and output gate) within the neural units, which can cope with the vanishing gradient problem and capture the information between the input before and after. LSTM is currently widely used in the area of natural language processing. In Zhang et al.'s study, they just applied single-feature LSTM model for prediction, i.e., just the VOC concentration was involved. Zhang et al. (2022) used the VOC concentration at previous time to predict the VOC concentration at next time (called univariate prediction). This procedure has been widely applied in prior time-series predictions with machine learning (Chen et al. 2018; Park et al. 2018; Taheri and Razban 2021). Since the number of occupants was not considered, the predictions deviated from the observations when occupancy changed greatly. In this section, we make a comparison between the present multi-feature LSSVM approach and the prior single-feature LSTM approach. Figure 6 shows the comparison for the prediction of 6-MHO and 4-OPA concentrations. According to Zhang et al.'s study, the prediction performance of LSTM was poor around the first class in the morning due to the significant variation in the number of students. The prediction from the multi-feature LSSVM approach during this period is significantly better than that of LSTM. At 8:40, the MAPEs between the LSTM prediction and experimental data for 6-MHO and 4-OPA are 31.9% and 32.6%, respectively, whereas the prediction errors for the LSSVM approach are 13.7% and 9.6%, respectively. In addition, during the period of 8:40–9:30, the MAPEs of our multi-feature LSSVM approach for 6-MHO and 4-OPA are 6% and 4.3%, respectively, which are also much lower than those obtained with the LSTM approach. Calculations based on all the experimental data during the fifth testing day also produced similar results. Besides, on a desktop computer with the similar configuration, the LSTM takes about 60 seconds per prediction, while the multi-feature LSSVM only needs 0.05 seconds. This analysis implies that using a simple machine learning approach with appropriate feature combinations can also generate excellent or even better predictions with less computing time than deep learning approach.

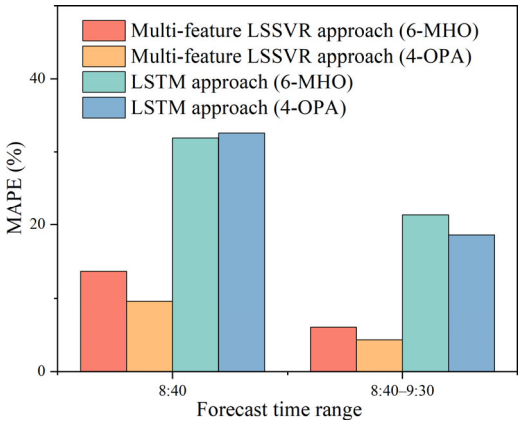

Fig. 6 Comparison between LSSVM and LSTM approaches for 6-MHO and 4-OPA predictions

3.6 Interval prediction based on KDE

Unlike deterministic point forecast, interval prediction can provide information about the predictive range, confidence level, and remaining uncertainties of future values. This is helpful for decision-makers to monitor and analyze indoor air quality, and thus merits investigation. According to the introduction in Section 2.2, here we examine the approach of LSSVM combined with KDE to establish an interval prediction model for 6-MHO and 4-OPA, and then obtain the interval prediction results for two confidence levels (90%, 80%). IFCP and IFAW are adopted to evaluate the predicted results, which are affected by the confidence level. The calculated results of the two metrics are given in Table 2. The constructed IFCP is considered to be theoretically valid if its coverage probability is greater than or equal to the corresponding nominal confidence level (Khosravi et al. 2013). According to the results in Table 2, the IFCP in all the experiments is greater than the corresponding nominal confidence level, demonstrating the validity of analysis with interval prediction.

Figure 7 shows the predicted intervals of 6-MHO and 4-OPA concentrations for the fifth testing day, based on the LSSVM and KDE combined approach, with 80% and 90% confidence levels, respectively. The prediction bandwidth widens as the confidence level increases, which is consistent with the theoretical principle. At both confidence levels, most

Table 2 The interval prediction results for different confidence levels

CI (%)	VOCs	IFCP (%)	IFAW
90	6-MHO	94.34	0.113
	4-OPA	94.34	0.071
80	6-MHO	83.01	0.080
	4-OPA	84.91	0.041

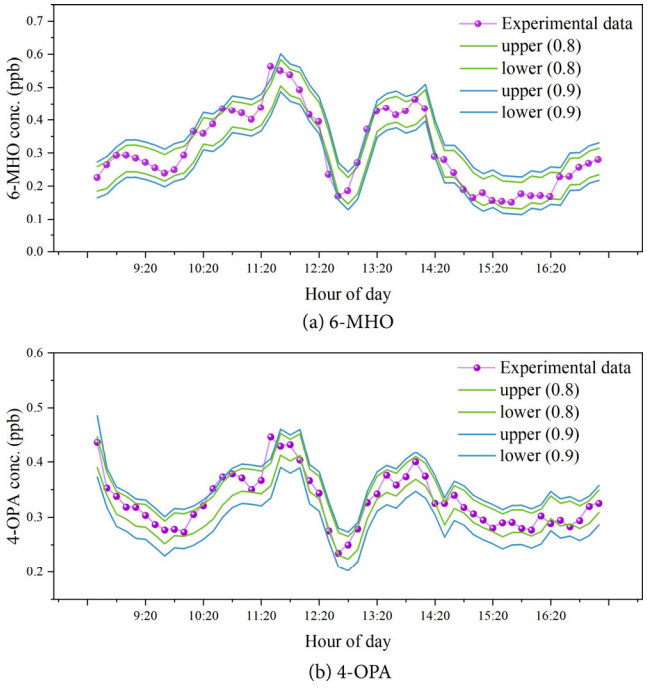

Fig. 7 Prediction intervals for (a) 6-MHO and (b) 4-OPA for 80% and 90% confidence levels

of the measured 6-MHO and 4-OPA data fall within the given prediction interval with moderate bandwidth. Unlike the normal distribution method, the KDE method can provide different upper and lower bounds to obtain a narrower bandwidth. This is a significant advantage of this method, which should be very helpful for future VOC prediction studies in realistic indoor settings.

3.7 Advantages and limitations

The advantages of this study can be summarized as follows:

(1) The prediction performance of traditional mechanistic-based models is heavily influenced by dozens of key parameters, and it is sometimes very challenging to measure these key parameters. When using a machine learning approach for prediction, there is no need to consider the actual physical mass transfer processes or chemical reaction processes, and also no need to carry out complicated and time-consuming parameter determination experiments, instead finding the governing laws through data driven, so the difficulty and cost of prediction are greatly reduced.

(2) Compared to previous studies, the multi-feature LSSVM approach presented here can capture the pattern of human-related VOCs more precisely.

(3) This is the first attempt to conduct interval prediction of indoor human-related VOCs, providing more useful information for decision-makers to monitor and analyze indoor air quality.

The limitations of this study include: (1) Although the prediction performance is improved compared with previous studies, a prediction discrepancy still exists when the number of occupants in realistic indoor settings vary drastically, and further investigation is needed. (2) The present prediction model is limited to a single machine learning approach, and subsequent research may attempt to average the results of multiple approaches using the voting method for robust predictions, or the stacking method for integration to improve model generalization.

4 Conclusions

We used a machine learning approach to rapidly and accurately estimate indoor human-related VOC concentration. We compared the prediction performance of five different approaches on 6-MHO concentration, and found that an LSSVM approach incorporating five features worked the best. We then used a multi-feature LSSVM approach to predict 4-OPA concentration in the university classroom, and obtained satisfactory results. Besides, the interval prediction model based on the kernel density estimation method has good performance, which makes up for the shortcoming that the deterministic point prediction model cannot provide uncertain information. This study contributes to the field by providing a method to accurately estimate indoor VOC exposure over a long period in realistic indoor settings. In addition to applying machine learning for time-series VOC concentration prediction, a promising direction of using machine learning could be to extract some key parameters needed for the physical or chemical models, or to discover some previously unknown controlling variables, which merits further and deep investigation.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 52178062), the Alfred P. Sloan Foundation (No. G-2016-7050), and the Opening Fund of State Key Laboratory of Green Building in Western China (LSKF202311). We thank Allen H. Goldstein, William W. Nazaroff, Pawel K. Misztal and Xiaochen Tang for the helpful comments and field campaign.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Author contribution statement

All authors contributed to the study conception and design. Data collection and analysis was performed by Jianyin

Xiong and Jialong Liu. The first draft of the manuscript was written by Jialong Liu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Abouleish MZ (2021). Indoor air quality and COVID-19. *Public Health*, 191: 1–2.

Amann A, Costello BD, Miekisch W, et al. (2014). The human volatilome: Volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *Journal of Breath Research*, 8: 034001.

Anderson SE, Franko J, Jackson LG, et al. (2012). Irritancy and allergic responses induced by exposure to the indoor air chemical 4-oxopentanal. *Toxicological Sciences*, 127: 371–381.

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32.

Bu Z, Dong C, Mmereki D, et al. (2021). Modeled exposure to phthalates via inhalation and dermal pathway in children's sleeping environment: A preliminary study and its implications. *Building Simulation*, 14: 1785–1794.

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.

Chen S, Mihara K, Wen J (2018). Time series prediction of CO_2, TVOC and HCHO based on machine learning at different sampling points. *Building and Environment*, 146: 238–246.

Cortes C, Vapnik V (1995). Support-vector networks. *Machine Learning*, 20: 273–297.

Cui P, Chen W, Wang J, et al. (2022). Numerical studies on issues of *Re*-independence for indoor airflow and pollutant dispersion within an isolated building. *Building Simulation*, 15: 1259–1276.

Domingo C, Watanabe O (2000). MadaBoost: A modification of AdaBoost. In: Proceedings of the 13th Annual Conference on Computational Learning Theory (COLT'00).

Friedman JH (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29: 1189–1232.

Fruekilde P, Hjorth J, Jensen NR, et al. (1998). Ozonolysis at vegetation surfaces a source of acetone, 4-oxopentanal, 6-methyl-5-hepten-2-one, and geranyl acetone in the troposphere. *Atmospheric Environment*, 32: 1893–1902.

Galanti T, Guidetti G, Mazzei E, et al. (2021). Work from home during the COVID-19 outbreak: The impact on employees' remote work productivity, engagement, and stress. *Journal of Occupational and Environmental Medicine*, 63: e426–e432.

Géron A (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, 3rd edn. Sebastopol, CA, USA: O'Reilly Media.

He Z, Xiong J, Kumagai K, et al. (2019). An improved mechanism-based model for predicting the long-term formaldehyde emissions from composite wood products with exposed edges and seams. *Environment International*, 132: 105086.

Hu Y, Xu L, Liang W (2023). A preliminary study on volatile organic compounds and odor in university dormitories: Situation, contribution, and correlation. *Building Simulation*, 16: 379–391.

Jarvis J, Seed MJ, Elton R, et al. (2005). Relationship between chemical structure and the occupational asthma hazard of low molecular weight organic compounds. *Occupational and Environmental Medicine*, 62: 243–250.

Kallio J, Tervonen J, Räsänen P, et al. (2021). Forecasting office indoor CO_2 concentration using machine learning with a one-year dataset. *Building and Environment*, 187: 107409.

Khazaei B, Shiehbeigi A, Ali Kani ARHM (2019). Modeling indoor air carbon dioxide concentration using artificial neural network. *International Journal of Environmental Science and Technology*, 16: 729–736.

Khosravi A, Nahavandi S, Creighton D (2013). Prediction intervals for short-term wind farm power generation forecasts. *IEEE Transactions on Sustainable Energy*, 4: 602–610.

Klepeis NE, Nelson WC, Ott WR, et al. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology*, 11: 231–252.

Kropat G, Bochud F, Jaboyedoff M, et al. (2015). Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. *Science of the Total Environment*, 505: 137–148.

Lakey PJ, Wisthaler A, Berkemeier T, et al. (2017). Chemical kinetics of multiphase reactions between ozone and human skin lipids: Implications for indoor air quality and health effects. *Indoor Air*, 27: 816–828.

Lakey PSJ, Morrison GC, Won Y, et al. (2019). The impact of clothing on ozone and squalene ozonolysis products in indoor environments. *Communications Chemistry*, 2: 1–8.

Landrigan PJ, Fuller R, Acosta NJR, et al. (2018). The Lancet Commission on pollution and health. *Lancet*, 391: 462–512.

Li Z, Tong X, Ho JMW, et al. (2021). A practical framework for predicting residential indoor $PM_{2.5}$ concentration using land-use regression and machine learning methods. *Chemosphere*, 265: 129140.

Little JC, Hodgson AT, Gadgil AJ (1994). Modeling emissions of volatile organic compounds from new carpets. *Atmospheric Environment*, 28: 227–234.

Liu Z, Ye W, Little JC (2013). Predicting emissions of volatile and semivolatile organic compounds from building materials: a review. *Building and Environment*, 64: 7–25.

NASEM (2022). Why Indoor Chemistry Matters. National Academies of Sciences, Engineering, and Medicine (NASEM). Washington, DC: The National Academies Press.

Nishihama Y, Jung CR, Nakayama SF, et al. (2021). Indoor air quality of 5, 000 households and its determinants. Part A: Particulate matter ($PM_{2.5}$ and $PM_{10\text{-}2.5}$) concentrations in the Japan Environment and Children's Study. *Environmental Research*, 198: 111196.

Park S, Kim M, Kim M, et al. (2018). Predicting PM_{10} concentration in Seoul metropolitan subway stations using artificial neural network (ANN). *Journal of Hazardous Materials*, 341: 75–82.

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.

Pham DM, Boussouira B, Moyal D, et al. (2015). Oxidization of squalene, a human skin lipid: a new and reliable marker of environmental pollution studies. *International Journal of Cosmetic Science*, 37: 357–365.

Reichstein M, Camps-Valls G, Stevens B, et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566: 195–204.

Sagi O, Rokach L (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, 8: e1249.

Salthammer T, Mentese S, Marutzky R (2010). Formaldehyde in the indoor environment. *Chemical Reviews*, 110: 2536–2572.

Skön J, Johansson M, Raatikainen M, et al. (2012). Modelling indoor air carbon dioxide (CO2) concentration using neural network. *International Journal of Chemical Engineering*, 6: 737–741.

Song Y, Qin S, Qu J, et al. (2015). The forecasting research of early warning systems for atmospheric pollutants: a case in Yangtze River Delta region. *Atmospheric Environment*, 118: 58–69.

Taheri S, Razban A (2021). Learning-based CO_2 concentration prediction: Application to indoor air quality control using demand-controlled ventilation. *Building and Environment*, 205: 108164.

Tang X, Misztal PK, Nazaroff WW, et al. (2015). Siloxanes are the most abundant volatile organic compound emitted from engineering students in a classroom. *Environmental Science & Technology Letters*, 2: 303–307.

Tang X, Misztal PK, Nazaroff WW, et al. (2016). Volatile organic compound emissions from humans indoors. *Environmental Science & Technology*, 50: 12686–12694.

Tian E, Yu Q, Gao Y, et al. (2021). Ultralow resistance two-stage electrostatically assisted air filtration by polydopamine coated PET coarse filter. *Small*, 17: e2102051.

Wang HF, Hu DJ (2005). Comparison of SVM and LS-SVM for regression. In: Proceedings of 2005 International Conference on Neural Networks and Brain, Beijing, China.

Wang H, Xiong J, Wei W (2022). Measurement methods and impact factors for the key parameters of VOC/SVOC emissions from materials in indoor and vehicular environments: A review. *Environment International*, 168: 107451.

Wei W, Ramalho O, Malingre L, et al. (2019). Machine learning and statistical models for predicting indoor air quality. *Indoor Air*, 29: 704–726.

Weschler CJ (2009). Changes in indoor pollutants since the 1950s. *Atmospheric Environment*, 43: 153–169.

WHO (2007). Indoor air pollution: National burden of disease estimates. World Health Organization.

Wisthaler A, Weschler CJ (2010). Reactions of ozone with human skin lipids: sources of carbonyls, dicarbonyls, and hydroxycarbonyls in indoor air. *Proceedings of the National Academy of Sciences of the United States of America*, 107: 6568–6575.

Wolkoff P, Larsen ST, Hammer M, et al. (2013). Human reference values for acute airway effects of five common ozone-initiated terpene reaction products in indoor air. *Toxicology Letters*, 216: 54–64.

Xiong J, Yao Y, Zhang Y (2011). C-history method: rapid measurement of the initial emittable concentration, diffusion and partition coefficients for formaldehyde and VOCs in building materials. *Environmental Science & Technology*, 45: 3584–3590.

Xiong J, He Z, Tang X, et al. (2019). Modeling the time-dependent concentrations of primary and secondary reaction products of ozone with squalene in a university classroom. *Environmental Science & Technology*, 53: 8262–8270.

Xu Y, Du P, Wang J (2017). Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: A case study in China. *Environmental Pollution*, 223: 435–448.

Xu C, Xu D, Liu Z, et al. (2020). Estimating hourly average indoor $PM_{2.5}$ using the random forest approach in two megacities, China. *Building and Environment*, 180: 107025.

Yang X, Chen Q, Zhang JS, et al. (2001). Numerical simulation of VOC emissions from dry materials. *Building and Environment*, 36: 1099–1107.

Yang T, Xiong J, Tang X, et al. (2018a). Predicting indoor emissions of cyclic volatile methylsiloxanes from the use of personal care products by university students. *Environmental Science & Technology*, 52: 14208–14215.

Yang X, Ma X, Kang N, et al. (2018b). Probability interval prediction of wind power based on KDE method with rough sets and weighted Markov chain. *IEEE Access*, 6: 51556–51565.

Yasin H, Caraka R, Hoyyi A, et al. (2016). Prediction of crude oil prices using support vector regression (SVR) with grid search–cross validation algorithm. *Global Journal of Pure and Applied Mathematics*, 12: 3009–3020.

Yuchi W, Gombojav E, Boldbaatar B, et al. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental Pollution*, 245: 746–753.

Zhang Y, Wang J, Wang X (2014). Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32: 255–270.

Zhang Y, Xiong J, Mo J, et al. (2016). Understanding and controlling airborne organic compounds in the indoor environment: mass transfer analysis and applications. *Indoor Air*, 26: 39–60.

Zhang M, Xiong J, Liu Y, et al. (2021a). Physical-chemical coupling model for characterizing the reaction of ozone with squalene in realistic indoor environments. *Environmental Science & Technology*, 55: 1690–1698.

Zhang R, Wang H, Tan Y, et al. (2021b). Using a machine learning approach to predict the emission characteristics of VOCs from furniture. *Building and Environment*, 196: 107786.

Zhang R, Tan Y, Wang Y, et al. (2022). Predicting the concentrations of VOCs in a controlled chamber and an occupied classroom via a deep learning approach. *Building and Environment*, 207: 108525.

Zhao L, Zhou H, Jin Y, et al. (2022). Experimental and numerical investigation of TVOC concentrations and ventilation dilution in enclosed train cabin. *Building Simulation*, 15: 831–844.

Zhou X, Liu Y, Liu J (2018). Alternately airtight/ventilated emission method: A universal experimental method for determining the VOC emission characteristic parameters of building materials. *Building and Environment*, 130: 179–189.