International Journal of
Environmental Research and Public Health

MDPI

Article

Updating Indoor Air Quality (IAQ) Assessment Screening Levels with Machine Learning Models

Ling-Tim Wong, Kwok-Wai Mui * and Tsz-Wun Tsang

Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong; ling-tim.wong@polyu.edu.hk (L.-T.W.); hayley.tsang@polyu.edu.hk (T.-W.T.)
* Correspondence: horace.mui@polyu.edu.hk; Tel.: +852-2766-5835

Abstract: Indoor air quality (IAQ) standards have been evolving to improve the overall IAQ situation. To enhance the performances of IAQ screening models using surrogate parameters in identifying unsatisfactory IAQ, and to update the screening models such that they can apply to a new standard, a novel framework for the updating of screening levels, using machine learning methods, is proposed in this study. The classification models employed are Support Vector Machine (SVM) algorithm with different kernel functions (linear, polynomial, radial basis function (RBF) and sigmoid), k-Nearest Neighbors (kNN), Logistic Regression, Decision Tree (DT), Random Forest (RF) and Multilayer Perceptron Artificial Neural Network (MLP-ANN). With carefully selected model hyperparameters, the IAQ assessment made by the models achieved a mean test accuracy of 0.536–0.805 and a maximum test accuracy of 0.807–0.820, indicating that machine learning models are suitable for screening the unsatisfactory IAQ. Further to that, using the updated IAQ standard in Hong Kong as an example, the update of an IAQ screening model against a new IAQ standard was conducted by determining the relative impact ratio of the updated standard to the old standard. Relative impact ratios of 1.1–1.5 were estimated and the corresponding likelihood ratios in the updated scheme were found to be higher than expected due to the tightening of exposure levels in the updated scheme. The presented framework shows the feasibility of updating a machine learning IAQ model when a new standard is being adopted, which shall provide an ultimate method for IAQ assessment prediction that is compatible with all IAQ standards and exposure criteria.

Keywords: machine learning model; indoor air quality (IAQ) index; screening; assessment

Citation: Wong, L.-T.; Mui, K.-W.; Tsang, T.-W. Updating Indoor Air Quality (IAQ) Assessment Screening Levels with Machine Learning Models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 5724. https://doi.org/10.3390/ijerph19095724

Academic Editor: Andrew S. Hursthouse

Received: 22 March 2022
Accepted: 6 May 2022
Published: 8 May 2022

1. Introduction

Indoor air quality (IAQ) has gained enormous attention in the past decade due to the considerable amount of time we spend indoors nowadays [1,2]. To tackle the problem of poor IAQ, different countries have their own set of IAQ standards, with different measurement parameters and range of exposure limits. Representative parameters, such as carbon dioxide (CO_2) and respirable suspended particulates (RSP), are always on the list, while total volatile organic compounds (TVOC), carbon monoxide (CO), ozone (O_3), formaldehyde (HCHO), airborne bacteria count (ABC) may be included, depending on the application purpose of the standard [3–7]. The exposure limits are usually established based on health risk analysis, in which lifelong exposure to that level of pollutant shall not produce significant adverse effects on the public [8].

Alternatively, instead of complying strictly with the IAQ standard, the screening approach for assessing IAQ has become popular in recent years due to its simplicity and cheaper monitoring cost. With a large enough sample size, we can find out the "common" IAQ problems one type of premises often experiences, therefore, identifying the representative IAQ parameters that explain the majority of poor IAQ. The simplest way to reduce the cost of IAQ assessment is to just measure these representative parameters and see if they exceed the standard. One of the most notable examples is using CO_2

level as an indicator of acceptable IAQ to adjust the fresh air quantity [9]. However, this approach may overlook the possibility of having IAQ problems caused by other IAQ parameters; therefore, a surrogate approach was proposed to identify surrogate IAQ parameters that are not just representative but also statistically correlated with other IAQ parameters. An express assessment protocol using three or five IAQ parameters, developed by Hui et al. [10], successfully screened out more than 90% of offices with poor IAQ, which provided an alternative for IAQ pre-assessment without the need to conduct a full assessment (all nine parameters). This study gave insight into the ability of a limited number of parameters in identifying problematic IAQ. Further to that, Wong et al. [11] proposed using CO_2, RSP and TVOC as the surrogate indicators for evaluating IAQ in offices. The dependence and the correlations of the other nine parameters on the levels of the proposed surrogate indicators were found to be statistically significant. The result served as strong support that CO_2, RSP and TVOC could be good surrogate indicators for other IAQ parameters, in terms of representativeness, ease of measurement and the possibility of real-time monitoring [12]. Individually, CO_2, RSP and TVOC represent occupant load and ventilation rate, system filtration performance and indoor activities, and emissions from building materials and finishes, respectively, which serve as good indicators for the general IAQ of an environment with a ventilation system. To sum up, using surrogate indicators for IAQ evaluation can reduce the scale of measurement, as some high-risk premises are already being screened out preliminarily, therefore, reducing the resources required to identify problematic premises [10,11].

Based on the aforementioned efforts for simplifying IAQ assessment, an efficient and cost-effective IAQ screening protocol was proposed by Wong et al. [13] for identifying asymptomatic IAQ problems. IAQ index, the average fractional dose to exposure limits of the representative pollutants, was proposed and was used to diagnose unsatisfied IAQ in air-conditioned offices in the study by Mui et al. [14]. IAQ indices from 525 offices were evaluated using a five-level screening test with thresholds determined by likelihood ratios of unsatisfactory IAQ. A likelihood ratio larger than 1 indicates a high-risk sample having an excessive occurrence of unsatisfactory IAQ, whereas a smaller than 1 likelihood ratio identifies a low-risk sample. Given the pre-test probability of unsatisfactory IAQ and the regional failure percentage of the Hong Kong IAQ Certification Scheme, the post-test probability of offices with unsatisfactory IAQ can be estimated using the IAQ screening test. This screening test with representative IAQ parameters provides a much simpler and cost-effective alternative for IAQ assessment. If an environment "fails" in the screening test (i.e., any one of the three surrogate indicators exceeds the exposure limit), immediate remedies can be decided on to improve the IAQ. If not, based on the post-test probability given by the screening test, facility management can determine the threshold of the test and threshold of the remedy regarding the willingness to invest manpower and resources in improving the IAQ. Further test, a comprehensive one, will only be needed if the screening test result is in between the two thresholds [14].

It is noteworthy that this approach does not simply test some of the parameters against the standard, but rather uses these parameters to predict the probability of dissatisfying the standard based on correlation. Therefore, an assessment model developed based on the levels of surrogate parameters and probability of failing an IAQ standard is essential in IAQ screening practice. More improvements have been made to the IAQ index to further reduce the resources required for IAQ screening [15]; however, as powerful as it is in screening the IAQ of similar environments, prior knowledge of the IAQ of premises in the region is required [10], and the index may not be applicable to other kinds of space or against another set of IAQ standards.

In fact, throughout the development of IAQ policy, exposure limits have been updated from time to time, based on collective professional judgement and managerial decisions with a balance of social acceptance. The World Health Organization (WHO) has been making constant efforts to improve and refine the air quality standards, since the establishment of the air quality guidelines on selected pollutants in 2005 [16], which include

the REVIHAAP project to review the health impacts of air pollution [17], and the HRAPIE project to identify dose–response relationship for RSP, O_3 and nitrogen dioxide (NO_2) [18]. Results from these two projects supported the comprehensive review of the European Union air quality policy in 2013 and many follow-up consultations and discussion forums on the preparation for an updated guideline [19]. In September 2021, the WHO issued the new Global Air Quality Guideline that reduced levels of key air pollutants to address the accumulated pieces of evidence of health effects and significant risks associated with poor air quality [20]. In 2019, the IAQ standard in Hong Kong was updated with stricter exposure limits to meet the updated IAQ guidelines published by the World Health Organization. The update consisted of the removal of three comfort parameters, the inclusion of visual inspection of mould condition and more stringent limits for CO, RSP and radon (Rn). Considering that the IAQ index itself, the screening levels and the likelihood ratios were all developed using the old standard, it is essential to identify the effect of the new IAQ standard on the suitability and performance of the established screening methods and to provide a framework for "updating" the screening levels.

With exposure standards being updated regularly in practical situations without the quantitatively assessed probable impact of the tightening of levels, fine tuning the IAQ screening baseline is deemed necessary. However, given that past data were assessed using the old standard, the iterative process for baseline determination using newly collected data takes a long time and is not ideal for responding to the rapid change in the need for environmental control. This presents a problem if the standard is being updated. Can the existing IAQ assessment model based on a statistical analysis of old data be useful against the new standard?

In this study, we proposed using machine learning methods for the development of a surrogate IAQ assessment model, which may be a solution to the problem of an updated IAQ standard and avoid the iterative process for baseline determination. Machine learning is a state-of-the-art method for environmental prediction. It is commonly used in outdoor pollution predictions [21] and indoor energy simulations [22]. The awareness and application of machine learning modeling in IAQ emerged in the past decade. A comprehensive review of existing machine learning and statistical models for IAQ prediction, conducted by Wei et al. [23], suggested that the majority of existing research focuses on using machine learning algorithms to predict pollutant concentrations. The most popular statistical models applied to IAQ consist of artificial neural network (ANN), multiple linear regression (MLR), partial least squares (PLS), and random forest (RF). They focus on predicting the concentrations of airborne particles, including RSP, e.g., [24–26], CO_2, e.g., [27,28], NO_2, e.g., [29] and Rn, e.g., [30,31], in indoor environments using outdoor data. Recently, the forecasting of IAQ has become popular for the sake of improving public health and well-being, since precautionary actions can be acted on ahead of time [32]. Machine learning methods, such as linear and non-linear autoregressive models [33], are used to develop IAQ forecasting models using the historical profile of IAQ parameters. As continuous monitoring of IAQ is required as the basis of time-series machine learning models, it is common to forecast temperature, e.g., [34,35], relative humidity, e.g., [35,36], CO_2, e.g., [34–36] and CO, e.g., [36], as they can be easily monitored using low-cost sensors [23]. Forecasting the concentration of indoor aldehydes, volatile organic compounds (VOC), and semi-VOC using statistical models remains scarce [33], and an example of using the nonlinear threshold autoregressive (TAR) model and Chaos-dynamics-based model to forecast HCHO is presented in the study by Ouaret et al. [37]. All things considered, it is advisable to test and compare different statistical models for each specific case, as demonstrated by many studies that used machine learning methods for IAQ modelling [33].

Besides indoor air pollutant prediction and forecasting, there are other examples of applying machine learning methods in IAQ-related research that can be found in the literature. Zimmerman et al. [38] applied random forests (RFs) to improve low-cost sensor performance for more accurate IAQ monitoring. Leong et al. [39] used a support vector machine (SVM) for the prediction of the air pollution index (API) in Malaysia. Their study

demonstrated that the radial basis function (RBF) kernel function could accurately and effectively predict API. Sarkhosh et al. [40] used a decision tree (DT) model to identify the most influential parameters that contributed to the prevalence of Sick Building Syndrome (SBS) in office buildings. The high prevalence of SBS was found to be related to job satisfaction, ergonomic parameters, microbiological pollutants and 1-methyl-4-(1-methylethyl) benzene concentration.

While IAQ prediction and forecasting give us a better understanding of the IAQ situation we are experiencing, it is of equal importance to identify whether the level of IAQ is considered acceptable or not before any follow-up mitigation or precautionary strategies are taken; therefore, an IAQ assessment model is essential.

To our best knowledge, we have identified the following research gaps in the field:

- Using machine learning methods to assess whether the IAQ is acceptable or not with a given IAQ standard;
- Addressing the issues of updating/changing IAQ standards, which would affect the screening levels and results; and
- Predicting the updated screening baselines of IAQ with new standards.

Therefore, in this study, we discuss the possibility of using machine learning methods to "update" the screening levels, such that the IAQ screening method can still be applicable with a new standard. Using Hong Kong's case of an updated IAQ standard as an example, in this paper, we present a universal framework of using machine learning models in predicting the updated IAQ screening levels, which includes:

- Developing and evaluating the performance of machine learning IAQ assessment models with surrogate IAQ parameters;
- Quantifying the impact of an updated scheme (i.e., an IAQ standard) on the machine learning IAQ assessment model; and
- Evaluating the model flexibility in adapting an updated/another exposure standard.

Applicable to all IAQ standards and guidelines, this framework not only enables the implementation of a territory-wide IAQ screening program but also facilitates IAQ monitoring and improvements.

2. Materials and Methods

In the following section, the framework for updating the screening levels of IAQ assessment models is presented. To demonstrate the updating process, machine learning models for IAQ assessment based on the developed IAQ index algorithm and screening methodology were first developed using selected machine learning modelling methods. The performances of the models were evaluated, and with the average assessment results from the models, the relative impact ratios of the updated standard on the old standard were determined. The framework details the feasibility of developing machine learning IAQ assessment models, methods for model performance evaluation and the procedures for updating the screening levels with an updated standard.

2.1. Overview of the Data

IAQ assessment data collected from a cross-sectional IAQ survey of 525 air-conditioned offices in Hong Kong reported in a previous study was adopted to evaluate the performance of machine learning models [14]. The surveyed premises, which covered various grades, types and ages, included a wide range of open-plan offices from 10 m^2 to 300 m^2. The IAQ survey was conducted for the fulfilment of the Hong Kong IAQ Certification Scheme (the Scheme); therefore, the measurement protocol, sampling locations, period and equipment strictly followed the requirements stated in the Scheme. As such, 8 h continuous samplings were conducted during the office-occupied hours with a sampling density of 500 m^2. All the sampling points were selected by the IAQ professionals during the walkthrough inspection before the actual measurement.

Two IAQ assessment schemes, Schemes 1 and 2, are exhibited in Table 1. Scheme 1 was the old IAQ objective in the Hong Kong IAQ Certification Scheme and Scheme 2 was the updated one to update the requirement against the latest IAQ guidelines by the World Health Organization [41]. In the updated scheme, exposure limits of CO, Rn and RSP are tightened to provide better public health protection. As mentioned above, the IAQ index using likelihood ratio cannot adapt to an updated standard since it was developed based on the previous standard, so using machine learning algorithms to model the IAQ index and IAQ dissatisfaction can, therefore, be a universal solution to the existing barrier.

Table 1. 8 h exposure limits of satisfactory indoor air quality.

Parameter (Unit)	Scheme 1	Scheme 2
CO_2 (ppm)	1000	1000
CO (ppm)	8.7	6.1
RSP (μg m^{-3})	180	100
NO_2 (μg m^{-3})	150	150
O_3 (μg m^{-3})	120	120
HCHO (μg m^{-3})	100	100
TVOC (μg m^{-3})	600	600
Radon (Bq m^{-3})	200	167
Airborne bacteria (CFU m^{-3})	1000	1000

A statistical summary of the dataset extracted for this study, which consists of three independent yet closely correlated IAQ surrogate indicators concerning the IAQ index [14], namely CO_2, RSP and TVOC, is presented in Table 2. These three parameters were selected as the surrogate indicators among the remaining 9 pollutants in the Scheme, among which, RSP represents the filtering efficiency of the air-conditioning system, CO_2 represents the occupant load and ventilation rate, and TVOC indicates building emission [13]. The overall summary of the dataset is shown at the top of the table, with the range of CO_2 = 339–1497 ppm, RSP = 4–125 μg m^{-3}, TVOC = 0–3144 μg m^{-3} and the calculated IAQ index = 0.189–1.99. Using the two assessment schemes introduced in Table 1 above, this dataset was further classified into "Satisfactory IAQ" (i.e., if all of the 9 pollutant levels fulfil the assessment scheme) or "Unsatisfactory IAQ" (i.e., 1 or more of the 9 pollutant levels fail the assessment scheme). While the mean values of CO_2, RSP and TVOC in the "Satisfactory IAQ" group were significantly different from those in the "Unsatisfactory IAQ" group ($p < 0.05$, t-test), the sample (satisfactory or unsatisfactory) group means results from Schemes 1 and 2 were statistically the same ($p > 0.1$, t-test). Table 2 also exhibits the IAQ index θ, which is an IAQ indicator determined using Equation (1), with $j = 1, \ldots, 3$, $\Phi_j{}^*$ being the fractional dose of RSP, CO_2 and TVOC, Φ_j the exposure level of the assessed parameter over an exposure time, and $\Phi_{j,e}$ the reference exposure limit under Scheme 1 (RSP = 180 μg m^{-3}, CO_2 = 1000 ppm, TVOC = 600 μg m^{-3}) [15].

$$\theta = \frac{1}{3}\sum_{j=1}^{3}\Phi_j^*;\ \Phi_j^* = \frac{\Phi_j}{\Phi_{j,e}} \quad (1)$$

Table 2. Statistical summary of levels of indoor air quality surrogate parameters in 525 offices, (**a**) overall summary; (**b**) summary of the dataset being classified as "Satisfactory IAQ" regarding Schemes 1 and 2; (**c**) summary of the dataset being classified as "Unsatisfactory IAQ" regarding Schemes 1 and 2.

	CO_2 (ppm)	RSP ($\mu g\ m^{-3}$)	TVOC ($\mu g\ m^{-3}$)	IAQ Index
(a) Overall Summary				
mean	658	30	358	0.473
std dev	151	20	328	0.201
min	339	4	0	0.189
25%	556	15	140	0.333
50%	639	22	295	0.431
75%	746	38	466	0.558
max	1497	125	3144	1.99
(b) Satisfactory IAQ				
Scheme 1				
Count		358		
mean	634	28	242	0.397
std dev	126	20	152	0.111
min	339	4	0	0.189
25%	546	14	113	0.312
50%	624	20	209	0.381
75%	714	33	354	0.477
max	998	125	597	0.725
Scheme 2				
Count		352		
mean	634	27	240	0.394
std dev	126	18	152	0.110
min	339	4	0.0	0.189
25%	547	14	112	0.311
50%	623	20	208	0.378
75%	713	32	354	0.474
max	998	99	597	0.725
(c) Unsatisfactory IAQ				
Scheme 1				
Count		167		
mean	709	34	607	0.637
std dev	184	19	446	0.249
min	396	7	45	0.202
25%	384	19	346	0.488
50%	678	29	517	0.406
75%	807	44	738	0.737
max	1497	91	3144	1.991
Scheme 2				
Count		173		
mean	707	36	598	0.634
std dev	183	22	442	0.246
min	396	7	45.0	0.202
25%	583	19	338	0.487
50%	678	29	497	0.603
75%	804	46	715	0.725
max	1497	125	3144	1.991

2.2. Data Preprocessing

Figure 1 shows the pair plots of the IAQ parameters grouped by satisfactory and unsatisfactory IAQ assessed using Schemes 1 and 2. A linear data scaling to the range [0, 1] was applied for data normalization.

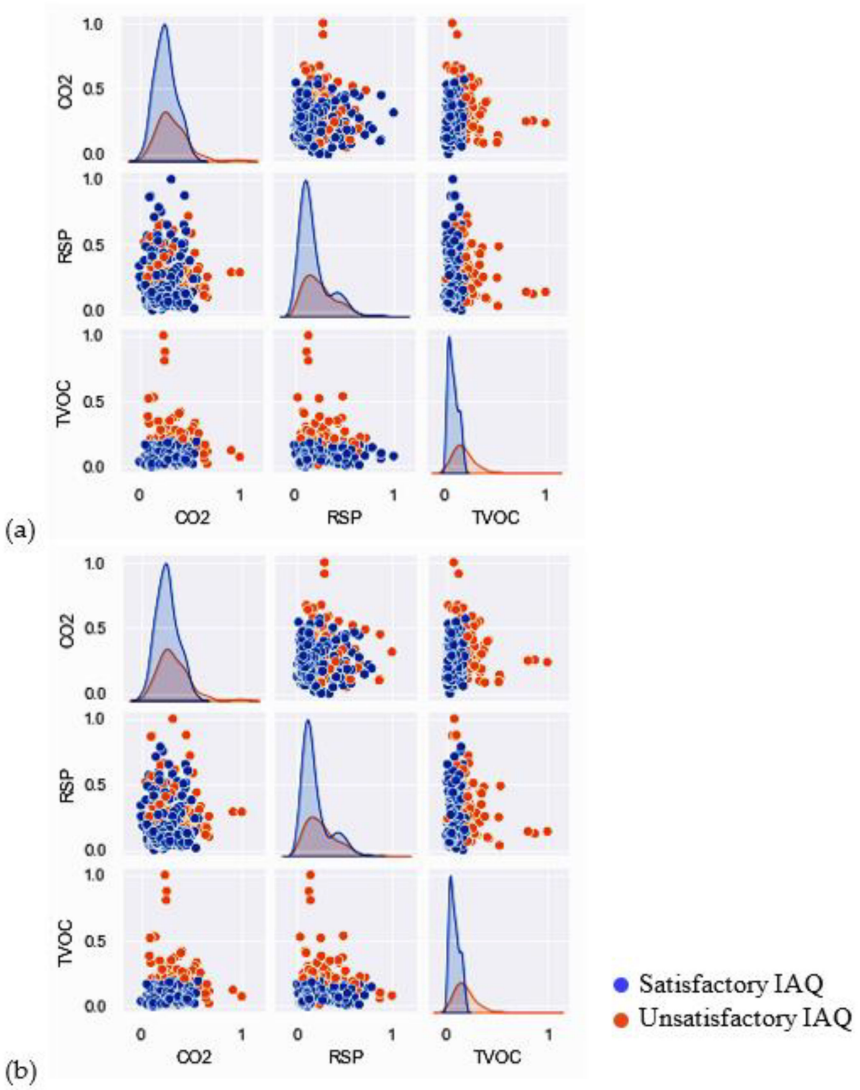

Figure 1. Pair plots of CO_2, RSP, and TVOC grouped by assessed indoor air quality (IAQ) against assessment (**a**) Scheme 1 (**b**) Scheme 2.

The training data and testing data were randomly selected at a distribution ratio of training data $(1 - r_d)$ and testing data (r_d), as shown in Equation (2), where $n_{d,t}$ and $n_{d,g}$ are the numbers of data points in the testing and training datasets, respectively.

$$r_d = \frac{n_{d,t}}{n_{d,g}} \tag{2}$$

Multifold cross-validation was employed for model validation. The training dataset was divided into 5 and 10 subsets of equal size and each subset was tested using the hyperparameters trained on the remaining subsets. The cross-validation accuracy was determined based on the percentage of correctly classified data. A grid search was then conducted to optimize the model hyperparameters, which were later used to retrain the model for evaluation.

The model accuracy AC, the probability of the model making a correct prediction [14], is usually compared with the baseline accuracy AC_{bl} in Equation (3) which indicates the certainty of the predictions made without the algorithm, where mode (N) is the mode of true result and N is the sample size.

$$AC_{bl} = \frac{\text{mode}\ (N)}{N} \tag{3}$$

The baseline accuracy values adopted are 0.682 and 0.670 for Schemes 1 and 2, respectively. A model with an accuracy below the baseline is considered to be unsatisfactory.

In this study, as shown in Figure 2, a total of 16 (=4 × 2 × 2) evaluation conditions were generated from 4 different combinations (r_d = 0.2, 0.3, 0.4, 0.5) of training and testing data, 2 multifold cross-validations (K = 5, 10) and 2 IAQ schemes (Schemes 1 and 2). Trained models (without grid-search-tuned model hyperparameters) and retrained models (with grid-search-tuned model hyperparameters) were then evaluated using the testing data of the 16 evaluation conditions, and finally, 32 sets of testing results were obtained for evaluating the performance of the 9 models for IAQ assessment.

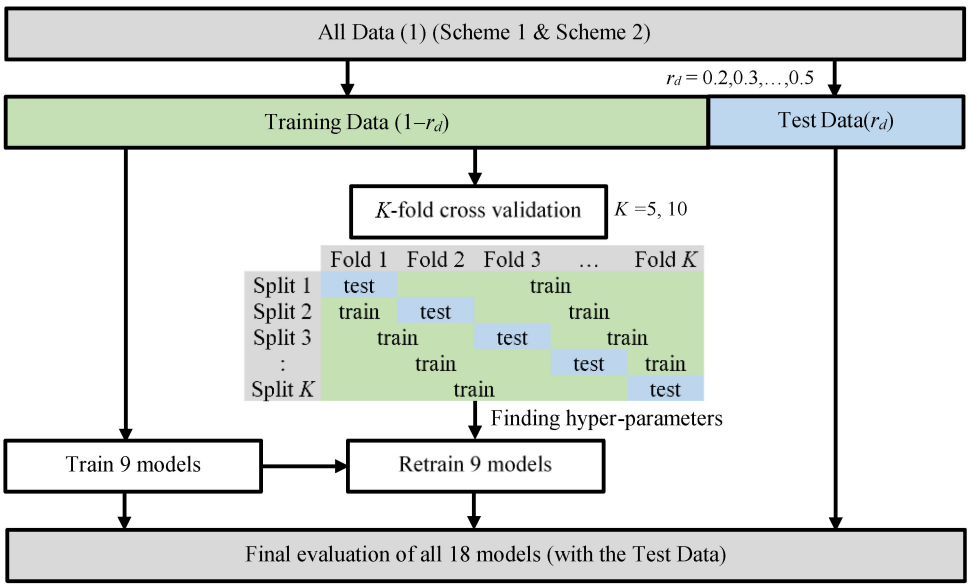

Figure 2. Data processing for model training and evaluation.

2.3. Models for Evaluation

Table 3 shows the classification models (classifiers) employed for developing the IAQ assessment model. The selected models included Support Vector Machine (SVM) with different kernel functions (i.e., linear, polynomial, radial basis function (RBF), and sigmoid), k-Nearest Neighbors (kNN), Logistic Regression, Decision Tree (DT), Random Forest (RF) and Multilayer Perceptron Artificial Neural Network (MLP-ANN). These algorithms are commonly used for developing IAQ prediction and forecasting models based on the literature review described in the introduction. In order to provide a universal framework for developing the IAQ assessment models and updating the screening levels, these popular models were adopted and their performances were evaluated. More details of each machine learning model and its hyperparameters can be found in Appendix A.

Table 3. Selected machine learning models and hyperparameters for the development of IAQ assessment models.

Models	Hyper-Parameters	Test Range	Validation Accuracy	Test Accuracy	Hyperparameters Used
SVM (linear)	r_d C	0.2–0.5 0.1–10,000	0.794–0.832	0.752–0.824	0.4 1.0
SVM (polynomial)	r_d C c_1 c_0	0.2–0.5 0.1–10,000 2, 3 0, 1	0.813–0.839	0.753–0.833	0.4 1000 3 1
SVM (rbf)	r_d C	0.2–0.5 0.1–10,000	0.806–0.831	0.762–0.824	0.4 1.0
SVM (sigmoid)	r_d C c_0	0.2–0.5 0.0001–2000 0–1	0.638–0.652	0.443–0.800	0.2 0.0001 0
kNN	r_d k W	0.2–0.5 2, 3, . . . , 11 1, $1/d_k$	0.785–0.809	0.762–0.824	0.4 10 1
Logistic regression	r_d C	0.2–0.5 0.001–20,000	0.790–0.825	0.753–0.810	0.4 1
Decision tree	r_d D n_s n_r Impurity	0.2–0.5 3, 4, . . . , 14 3, 4, . . . , 19 2, 3, . . . , 6 *GI*, *EI*	0.805–0.829	0.714–0.838	0.2 4 3 2 *EI*
Random forest	r_d n_f D n_s n_r Impurity	0.2–0.5 10, 60, 110 1, 2, . . . , 11 1, 2, . . . , 9 2, 3, . . . , 6 *GI* or *EI*	0.824–0.844	0.724–0.829	0.3 60 2 3 1 *GI*
MLP-ANN	r_d C Neurons Hidden layer Activation Iteration Learning rate	0.2–0.5 0.0001, 0.05, 1 100, 200 1, 3, 4, 6 Identity, logistic, *tanh*, relu LBFGS, SDG, Adam Constant, invscaling, adaptive	0.807–0.836	0.714–0.810	0.4 0.0001 200 3 relu LBFGS Constant

Table 3 also presents the test ranges of the hyperparameters, the cross-validation accuracy and the model accuracy with the testing datasets, and the corresponding hyperparameters that gave the best prediction accuracy in all tests. The development and the training of models were coded using the Python programming language described by Pedregosa et al. [42].

Regularization was applied to avoid overfitting by penalizing large coefficients [43]. It was intended to reduce the generalization error but not the training error. As a result, the application of regularization allowed a certain amount of misclassified data points in the training dataset [44]. To minimize the error between the true value y_i and the predicted

value $x\beta$, the cost function f shown in Equation (4) could be expressed with the L2 loss function $\sum_i \left(y_i - \sum_j x_{ij}\beta_j\right)^2$ and the regularization factor C [45].

$$f = \sum_i \left(y_i - \sum_j x_{ij}\beta_j\right)^2 + C\sum_j \beta_j^2 \quad (4)$$

3. Results and Discussion

Figure 3 illustrates the cross-validation accuracy of the SVM classifiers with linear, RBF, sigmoid and polynomial kernels. Consistent accuracy of $AC > 0.8$ was observed when the regularization factor C was $\geq$2 for the SVM with linear kernel, and for the whole test ranges of the SVM with RBF and polynomial kernels. However, the SVM with sigmoid kernel did not perform well for the training datasets, as compared with other kernels, with $AC \leq 0.65$, which dropped significantly for $C \geq 0.6$.

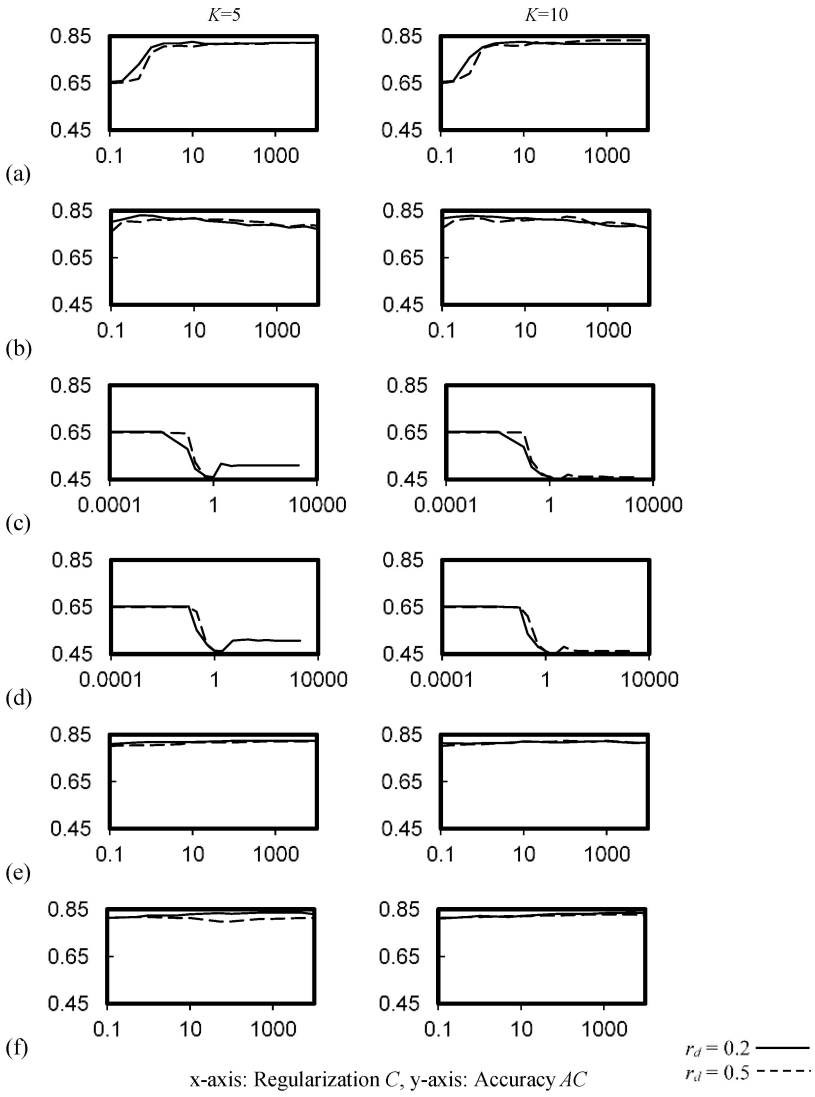

Figure 3. Cross-validation accuracy of the SVM classifier. (**a**) Linear kernel, (**b**) rbf kernel, (**c**) sigmoid kernel, c_0 = 0.01, (**d**) sigmoid kernel, c_0 = 0.5, (**e**) polynomial kernel, c_0 = 0, c_1 = 2, (**f**) polynomial kernel, c_0 = 1, c_1 = 3.

Figure 4 shows the cross-validation accuracy of the *k*NN classifier, which was consistent for k = 2–11. While the accuracy was more sensitive to the weight function applied, a larger k that compensated for the accuracy drop was observed in Figure 4a.

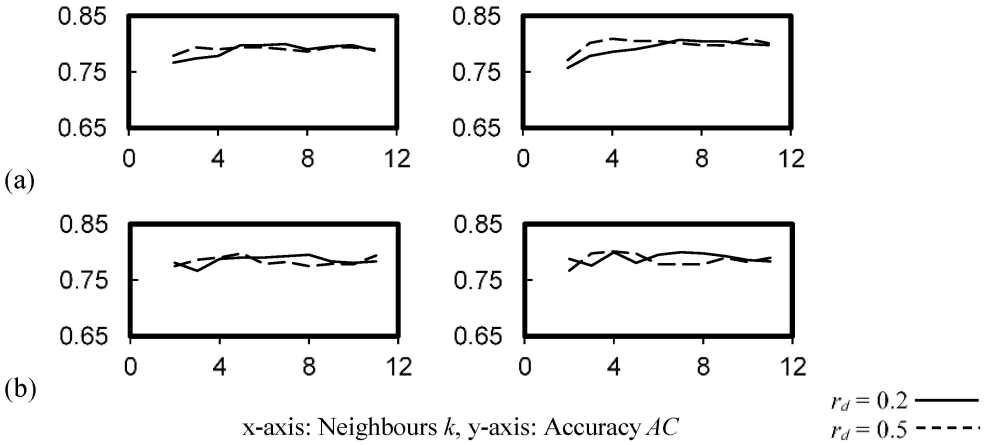

Figure 4. Cross-validation accuracy of the *k*NN classifier. (**a**) $W = 1/d_k$, (**b**) $W = 1$.

According to Figure 5, the logistic regression classifier improved the prediction accuracy for regularization factor $C > 2$. The choice of training dataset was found to be insignificant to the model accuracy.

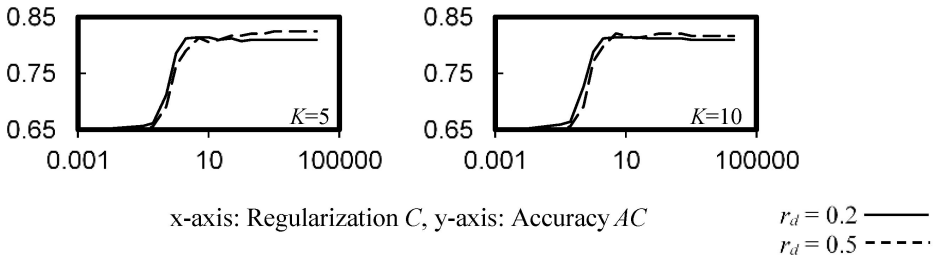

Figure 5. Cross-validation accuracy of the logistic classifier.

Figure 6 graphs the cross-validation accuracy of the decision tree classifier. Within the range of 0.75–0.8, the accuracy was sensitive to the size of the dataset, the impurity function, the minimum number of samples required to split an internal node n_s, and the minimum number of samples required to be at a leaf node n_r. It became less sensitive when the maximum depth value was greater than or equal to 10 (i.e., $D \geq 10$).

Figure 7 exhibits the cross-validation accuracy of the random forest classifier. The accuracy, which became less sensitive for $D \geq 2$, was improved, as compared with Figure 6. It can be seen that the number of trees n_f compensated for the accuracy drop due to $D \leq 5$.

A wide range of hyperparameters can be adopted for a MLP-ANN classifier. In this study, 100 and 200 neurons in the inner layers 1, 3, 4 and 6 were evaluated, with neuron arrangements of each layer in the ratios of (1), (1:8:1), (1:4:4:1) and (1:2:2:2:2:1). Figure 8 illustrates the cross-validation accuracy of the 60 configurations of the model hyperparameters for the inner-layer architecture (i.e., x-axis with legends 1–60, Table A1). A very sensitive accuracy ranging from <0.45 to about 0.8 was observed.

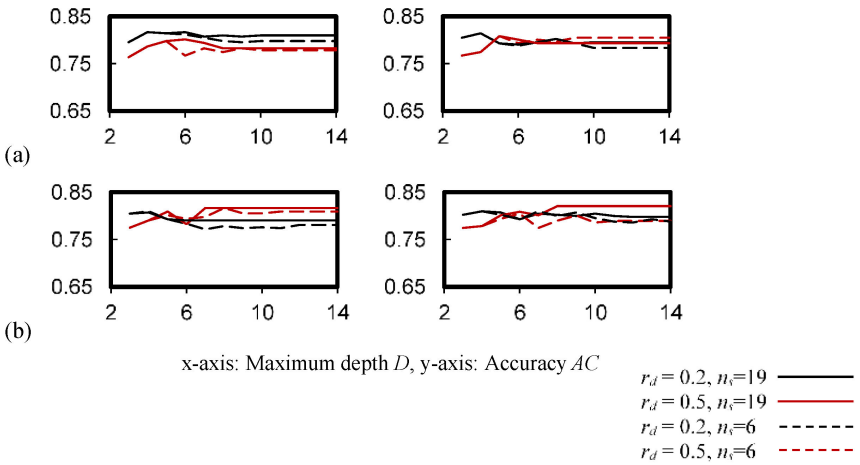

Figure 6. Cross-validation accuracy of the decision tree classifier. (**a**) Entropy impurity, n_r = 6 (**b**) Gini impurity, n_r = 2.

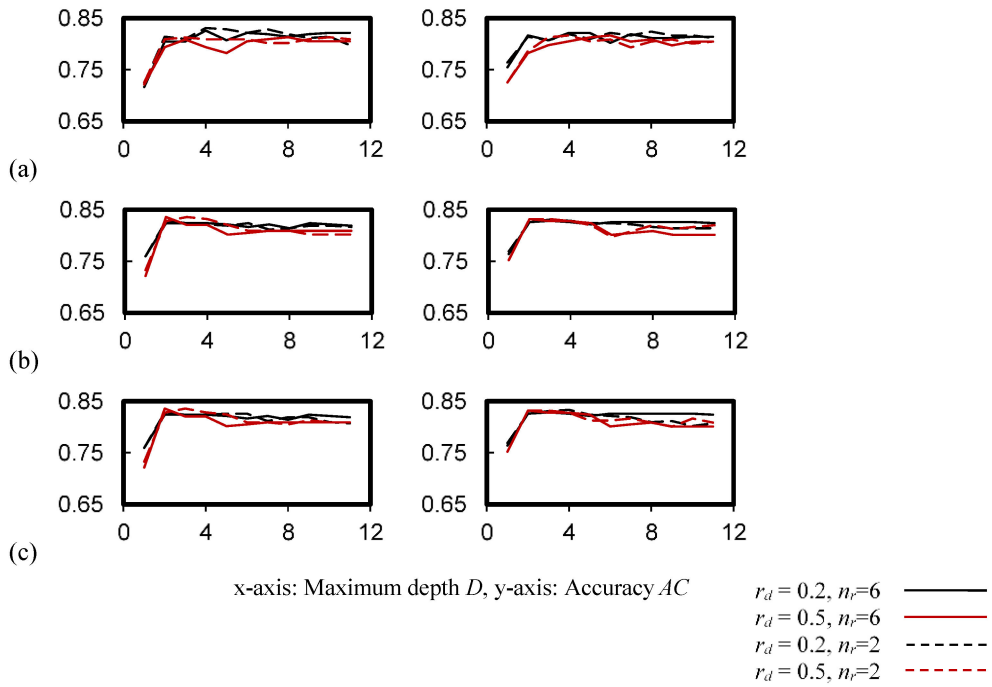

Figure 7. Cross-validation accuracy of the random forest classifier. (**a**) Entropy impurity, n_s = 9, n_f = 10 (**b**) Gini impurity, n_s = 9, n_f = 110, (**c**) Gini impurity, n_s = 2, n_f = 110.

It was challenging to set up a suitable MLP-ANN for an engineering application without prior selection of the model hyperparameters. Table 4 shows the test accuracy of the MLP-ANN classifier. The identity activation function made the best predictions with the highest (mean and median) test accuracy. Iteration schemes ADAM and L-BFGS, with constant learning rates only, returned more accurate predictions, as compared with SGD.

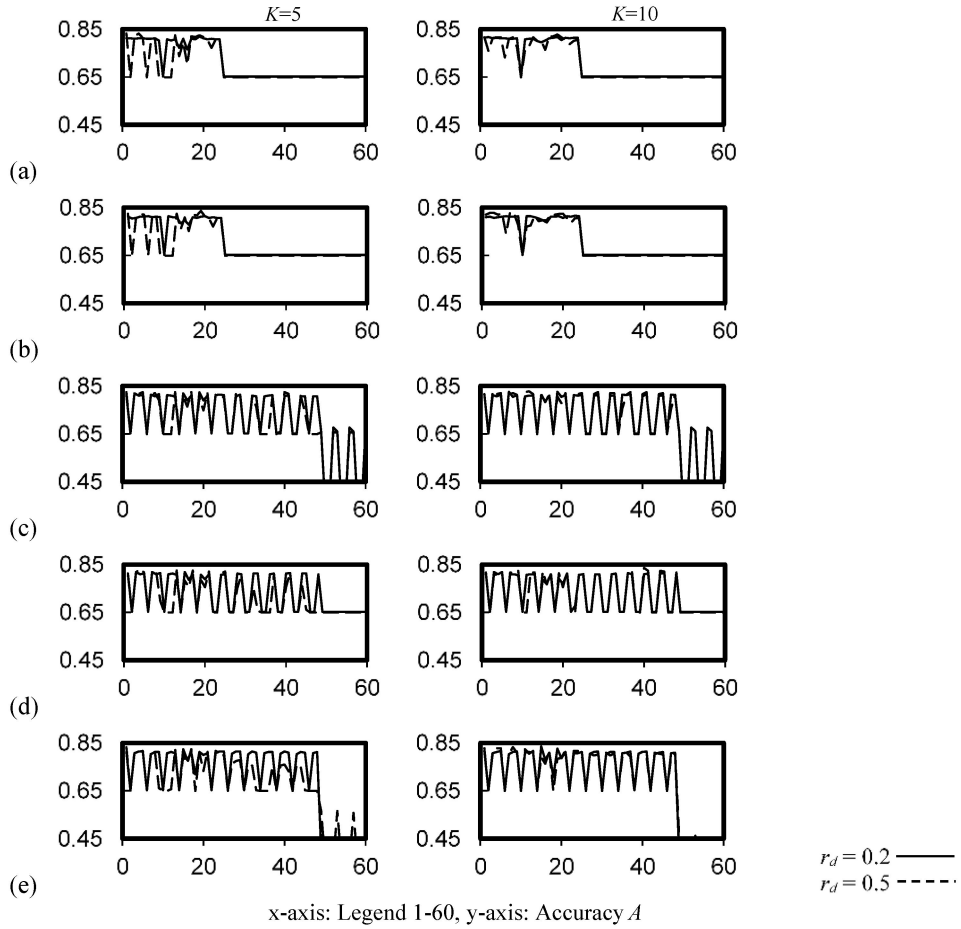

Figure 8. Cross-validation accuracy of the MLP-ANN classifier. (**a**) 100 neurons, 1 hidden layer, (**b**) 200 neurons, 1 hidden layer, (**c**) 100 neurons, 6 hidden layers (**d**) 200 neurons, 6 hidden layers, (**e**) 100 neurons, 3 hidden layers.

Table 4. Test accuracy of the MLP-ANN classifier (5-fold and 10-fold).

Hyper-Parameters			Test Accuracy			
Activation	**Iteration**	**Learning Rate**	**Mean**	**Median**	**Min**	**Max**
identity	All		0.740	0.795	0.336	0.836
logistic			0.636	0.646	0.348	0.828
tanh			0.728	0.783	0.348	0.836
relu			0.701	0.743	0.348	0.836
all	ADAM	Constant	0.765	0.801	0.638	0.832
	LBFGS	Constant	0.767	0.802	0.638	0.836
	SGD	Adaptive	0.712	0.648	0.638	0.828
		Constant	0.712	0.648	0.638	0.836
		invscaling	0.550	0.646	0.336	0.676
identity	ADAM	constant	0.801	0.806	0.641	0.832
	LBFGS	constant	0.805	0.806	0.778	0.824
	SGD	adaptive	0.758	0.793	0.638	0.828
		constant	0.758	0.791	0.638	0.836
		invscaling	0.579	0.646	0.336	0.668

Table 4. *Cont.*

Hyper-Parameters			Test Accuracy			
Activation	**Iteration**	**Learning Rate**	**Mean**	**Median**	**Min**	**Max**
logistic	ADAM	constant	0.667	0.646	0.638	0.828
	LBFGS	constant	0.683	0.646	0.638	0.820
	SGD	adaptive	0.646	0.646	0.638	0.652
		constant	0.646	0.646	0.638	0.652
		invscaling	0.536	0.646	0.348	0.652
relu	ADAM	constant	0.794	0.804	0.638	0.832
	LBFGS	constant	0.797	0.804	0.638	0.836
	SGD	adaptive	0.689	0.646	0.638	0.823
		constant	0.689	0.646	0.638	0.826
		invscaling	0.536	0.646	0.348	0.652
tanh	ADAM	constant	0.799	0.805	0.641	0.832
	LBFGS	constant	0.782	0.772	0.702	0.824
	SGD	adaptive	0.754	0.786	0.638	0.826
		constant	0.755	0.786	0.638	0.836
		invscaling	0.548	0.646	0.348	0.676

To sum up, all of the IAQ assessment models developed achieved the maximum test accuracy, in a narrow range of 0.807–0.820, with the mean test accuracy ranging from 0.536 to 0.805. Table 5 presents the best-performed models in the 32 tests (16 each for the trained and retrained models). The results showed that the SVM with polynomial kernel gave the highest test accuracy and next-best predictions in the trained and retrained model tests. Moreover, models with decision tree and random forest classifiers gained 4 and 3 counts (out of 16), respectively, in the trained model test, whereas the SVM with linear kernel gained 8 counts (i.e., the best prediction performance) in the retrained model test. These classifiers can be good choices for accurate IAQ assessment model development.

Table 5. The most accurate classifiers in 32 comparison tests.

Classifier	Trained Model		Retrained Model		Trained & Retrained Models	
	Count (N = 16)	**Test Accuracy**	**Count (N = 16)**	**Test Accuracy**	**Count (N = 16)**	**Test Accuracy**
SVM (linear)	0		8	0.811	8	0.811
SVM (polynomial)	6	0.820	6	0.816	12	0.818
SVM (rbf)	0		2	0.814	2	0.814
SVM (sigmoid)	0		0		0	
*k*NN	2	0.807	0		2	0.807
Logistic regression	0		0		0	
Decision tree	4	0.814	0		4	0.814
Random forest	3	0.819	0		3	0.819
MLP-ANN	1	0.810	0		1	0.810

4. Model Prediction of IAQ Assessment with IAQ Index Updates

The IAQ index was developed previously as a screening strategy to screen out premises with problematic IAQ based on assessment Scheme 1. Given that the assessment scheme has been updated to Scheme 2, this section evaluates the relative impact of the index due to the updated values of baselines in the two schemes.

The relative impact on the IAQ index for IAQ assessment with Schemes 1 and 2 was evaluated using three uniformly distributed ranges: CO_2 = 400–1400 ppm, RSP = 1–120 μg m^{-3},

and TVOC = 0–1500 μg m^{-3}. The selected ranges of surrogate pollutants generally cover the observable range in the office IAQ database. Determined by Monte Carlo sampling techniques, the three IAQ parameters in the above ranges were used to calculate the corresponding IAQ index and to predict the IAQ satisfaction/dissatisfaction using the trained and retrained classifiers.

Figure 9 shows the percentage of predicted satisfactory and unsatisfactory IAQ for the range of IAQ indices under Schemes 1 and 2. The IAQ satisfaction was assessed by the best performing trained and retrained IAQ classification models (with model accuracy shown in brackets). Classifications were performed with models with classifiers of a decision tree, a random forest, SVM with polynomial kernel and RBF kernel for Scheme 1, and models with classifiers of *k*NN, MLP-ANN, SVM with linear kernel and polynomial kernel for Scheme 2. The figure shows that the predictions of unsatisfactory IAQ made by these models generally agree with each other, with a deviation up to ±5% from the average prediction of satisfactory IAQ with Scheme 2.

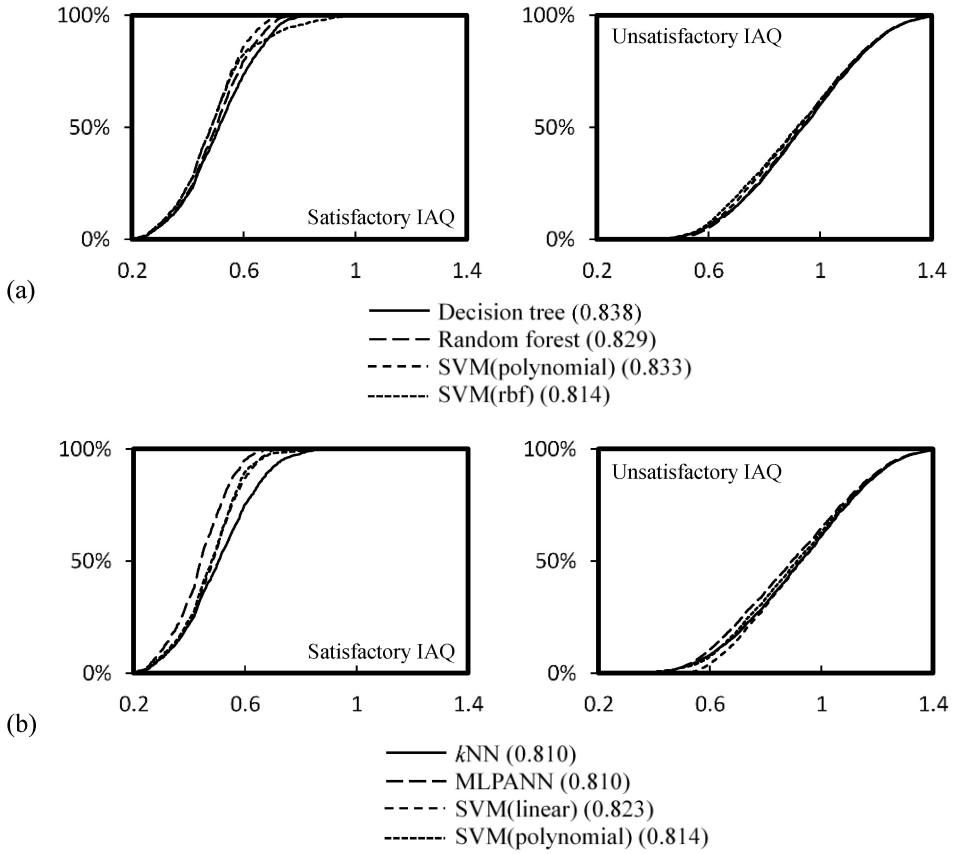

Figure 9. Predicted IAQ satisfaction and dissatisfaction with an IAQ index with assessment criteria, (**a**) Scheme 1, (**b**) Scheme 2.

The IAQ index in Figure 9 does not map any particular office distribution function and, thus, a relative approach was adopted to study the relative impact of Scheme 2 on Scheme 1, in terms of assessment likelihood, using the dataset summarized in Table 2. The relative impact ratio $r_{2,1}$ is determined by Equation (5), where x_u and x_s are the distribution functions of the IAQ index for unsatisfactory and satisfactory IAQ respectively.

$$r_{2,1} = \frac{LR_2}{LR_1};\ LR = \frac{\int_{x_1}^{x_2} f(x_u)dx}{\int_{x_1}^{x_2} f(x_s)dx} \quad (5)$$

Table 6 outlines a proposed likelihood ratio LR_1 for air-conditioned offices with unsatisfactory IAQ using Scheme 1, as reported in an earlier study [29]. The estimation of $r_{1,2}$ was made based on the average predictions from all models shown in Figure 9. Normality of the IAQ index was assumed ($p > 0.05$, w/s test). Based on the relative impact values determined for the IAQ index ranges <0.32, 0.32–0.42, 0.43–0.53, 0.54–0.64, ≥0.65, the corresponding values of LR_2 were computed (by $LR_2 = r_{2,1}\ LR_1$) and summarized in Table 6. The corresponding likelihood ratios in Scheme 2 were found to be higher due to the tightening of assessment criteria in the updated scheme.

Table 6. IAQ index of air-conditioned offices in Hong Kong.

IAQ Index θ	Likelihood Ratio (Scheme 1) LR_1	Relative Impact $r_{2,1}$	Likelihood Ratio (Scheme 2) LR_2
<0.32	0.1	1.4	0.1
0.32–0.42	0.4	1.2	0.5
0.43–0.53	0.8	1.1	0.9
0.54–0.64	1.7	1.3	2.2
≥0.65	25	1.5	38

5. Conclusions

One of the ongoing IAQ development tasks is to constantly improve IAQ objectives so that they are updated, relevant and attainable. Territory-wide IAQ screening should be implemented immediately, and later, periodically, to understand the overall IAQ situation and to maintain an up-to-date IAQ profile. Given so many IAQ standards with a wide range of exposure limits established by various governments, a universal framework for IAQ assessment modelling, which applies to all standards, is of urgent need.

In this study, a new strategy for unsatisfactory IAQ prediction using machine learning models of three surrogate IAQ indicators in the IAQ index was proposed. The results showed that all selected machine learning models performed well, achieving a maximum test accuracy of 0.807–0.820. Among the selected models, SVM with linear kernel and polynomial kernel, decision tree classifier and random forest classifier gave an IAQ classification with higher accuracy. To further demonstrate the use of IAQ index with different exposure limits in IAQ assessment, machine learning models of IAQ index using two different baselines (Schemes 1 and 2) were presented. The predictions of IAQ made by all selected models generally agreed with each other, with a ±5% deviation observed in the prediction of satisfactory IAQ under Scheme 2. The likelihood ratio of the IAQ index in Scheme 2 also increased with the tightening criteria for assessing exposure levels.

As demonstrated, machine learning models for IAQ index give promising prediction accuracy in identifying unsatisfactory IAQ, and that shall provide an ultimate strategy for IAQ screening and assessment, even under various IAQ standards and exposure criteria.

Author Contributions: Conceptualization, L.-T.W. and K.-W.M.; methodology, L.-T.W.; formal analysis, L.-T.W.; writing—original draft preparation, L.-T.W., K.-W.M. and T.-W.T.; writing—review and editing, L.-T.W., K.-W.M. and T.-W.T.; supervision, L.-T.W. and K.-W.M.; project administration, K.-W.M.; funding acquisition, K.-W.M. and L.-T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly supported by a grant from the Collaborative Research Fund (CRF) COVID-19 and Novel Infectious Disease (NID) Research Exercise, Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. PolyU P0033675/C5108-20G, HKPU P0033675/E-RB0P, PolyU 15217221 P0037773/Q-86B, PolyU 152088/17E P0005278/Q-59V) and the Research Institute for Smart Energy (RISE) Matching Fund (Project no. P0038532).

Data Availability Statement: Data available on request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Nomenclature

IAQ index and updates	
j	surrogate parameter
Φ_j^*	fractional dose
Φ_j	exposure level
$\Phi_{j,e}$	reference exposure limit
θ	IAQ index
r	relative impact ratio
x_u/x_s	distribution functions for unsatisfactory/satisfactory IAQ index
LR	likelihood ratio
Data processing	*Data processing*
X	data vector
$r_d/1-r_d$	test data/training data
$n_{d,t}/n_{d,g}$	number of data points in the test/training set
AC	model accuracy
AC_{bl}	baseline accuracy
TP/TN	true positive/negative
FP/FN	false positive/negative
N	sample size
K	number of folds
Units for IAQ parameters	
ppm	parts per million
µg m^{-3}	microgram per cubic meter
Bq m^{-3}	becquerels per cubic meter
CFU m^{-3}	colony-forming units per cubic meter
Regularization	
f	cost function
y_i	true value
$x\beta$	predicted value
C	regularization factor
n	number of dimensions
Decision tree/random forest	
p_j^2	probability of j
j	class
D	tree's maximum depth
n_s/n_r	minimum number of samplesrequired to split an internal node/be at a leaf node
n_f	number of trees
Support Vector Machines	
α, β	constants
x_i	inputs
y_i	output class
M	margin half-width
ε_i	slack variables
c_0, c_1	hyperparameters for $K(x_i,x_j)$
$K(x_i,x_j)$	kernel function
γ	kernel coefficient

k-Nearest Neighbors	
k	constant
$d(x_i,y_i)$	Euclidean distance
$\hat{y}$	predictions
W	weight function
$d_k{}^{-1}$	neighbour distance
MLP-ANN	
R	dataset
m/o	dimension for input/output
J	local gradient of function f
β	parameter
y	independent variables
δ	increment
Logistic regression	
x_0	sigmoid's midpoint of x
x	inputs
k	logistic growth rate
w	coefficient vector

Appendix A.

Appendix A.1. Support Vector Machine (SVM)

The support vector machine (SVM) algorithm identifies the optimal hyperplane in an n-dimensional space that distinctly separates the data points to be classified into two classes (in this study, satisfaction or dissatisfaction). The algorithm maximizes the margin between these two classes. The linear classifier can be expressed by Equation (A1), where α and β are constants, x is the input vector of inputs x_i [46,47], and y_i is the output class.

$$f(x) = \beta_0 + \sum_i \alpha_i \langle x_i, x \rangle;\ f(y_i) = \begin{cases} 0 & f(x_i) < 0 \\ 1 & f(x_i) > 0 \end{cases} \quad \text{(A1)}$$

To maximize the margin half-width M of the strip that separates the data points into the two classes, slack variables ε_i are specified for the soft margins, such that observations (training data) on the wrong side are allowed. It is a trade-off between misclassification of the training samples and simplicity of the decision surface suitable for a general model.

In Equation (A2), C is the regularization factor that is optimized for the number of samples [42]. For a large value of C, the optimizer chooses a smaller-margin hyperplane if that hyperplane can classify all the training points correctly. Conversely, a small value of C causes the optimizer to look for a larger-margin separating hyperplane. The application of regularization improves the numerical stability and the universality errors for predicting unseen data.

$$\sum_i \varepsilon_i \le C;\ y_i(\beta_0 + \beta_1 x_{i1} + \ldots) \ge M(1 - \varepsilon_i),\ \varepsilon_i \ge 0 \quad \text{(A2)}$$

Four types of kernel functions $K(x_i,x_j)$ in SVM were investigated in this study. They were linear, polynomial, radial basis function (RBF) and sigmoid kernel functions, expressed below in Equations (A3)–(A6), where c_0 and c_1 are the hyperparameters for the functions [48], and γ is the kernel coefficient, which defines how much influence a single training sample has. A large γ increases the area of influence of the support vectors but reduces the regularization for overfitting prevention, whereas a small γ constrains the model to capture the complexity of the data. The behavior of the model is very sensitive to the value of γ.

$$K(x_i,\ x_j) = \varphi(x_i)^T \varphi(x_j) = \langle x_i, x_j \rangle \quad \text{(A3)}$$

$$K(x_i,\ x_j) = \left[c_0 + \gamma \langle x_i, x_j \rangle\right]^{c_1} \quad \text{(A4)}$$

$$K(x_i,\ x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \quad \text{(A5)}$$

$$K(x_i,\ x_j)\ =\ tanh(c_0 + \gamma\langle x_i, x_j\rangle) \tag{A6}$$

Appendix A.2. k-Nearest Neighbors (kNN)

The *k*-nearest neighbors (*k*NN) algorithm is a non-parametric classification approach that classifies a point based on the majority class of the *k*-neighbors closest to the point. The average response of the *k*-closest points to x is given by Equation (A7).

$$f(x)\ =\ \frac{1}{k}\sum_{i\ =\ 1\ldots k} y_i \tag{A7}$$

The Euclidean distance $d(x_i,y_i)$, expressed in Equation (A8), is usually adopted for calculating the distance [49].

$$d(x_i, y_i)\ =\ \sqrt{\sum_{i\ =\ 1\ldots k}(x_i - y_i)^2} \tag{A8}$$

The neighbors closer to a query point have a greater influence than the neighbors that are farther away. Therefore, the predictions $\hat{y}$ can be made with a non-negative weight function to the neighbor distance $W{\sim}{d_k}^{-1}$, as shown in Equation (A9).

$$\hat{y}\ =\ \sum_{i\ =\ 1\ldots n} W(x_i,\ x_j)x_i \tag{A9}$$

Appendix A.3. Logistic Regression

A logistic regression algorithm is a linear classification model. The probabilities of the outcomes of a single trial are modelled using the logistic function exhibited in Equation (A10), where x_0 is the x value of the sigmoid's midpoint, and k is the logistic growth rate [50].

$$f(x)\ =\ \frac{1}{1 + exp[-k(x - x_0)]} \tag{A10}$$

The decision function is expressed in Equation (A11), where w is a coefficient vector.

$$f(x)\ =\ min_{w,c}\frac{1}{2}w^T w + C\sum_{i\ =\ 1\ldots n} log\left(exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right) \tag{A11}$$

Appendix A.4. Decision Tree (DT) and Random Forest (RF)

A decision tree (DT) is a non-parametric learning algorithm that partitions the data into subsets for classification [40]. The goal is to create the smallest possible tree (training model) that can predict the value of a target variable by learning simple decision rules. A tree can be seen as a piecewise constant approximation. The binary partitioning process continues until no further splits can be made so that the tree nodes are pure. The node purity can be measured by Gini impurity (GI) or by the information entropy (EI). GI measures the frequency at which any element of the dataset is mislabeled when it is randomly labeled. EI measures the disorder of the features with the target. A tree node is determined by minimizing the chosen index so that all the contained elements in the node are of one unique class. The GI and EI can be expressed by Equations (A12) and (A13), where ${p_j}^2$ is the probability of class j.

$$GI\ =\ 1 - \sum_j p_j^2 \tag{A12}$$

$$EI\ =\ -\sum_j p_j log_2 p_j \tag{A13}$$

Regularization can be done by confining the tree size, the tree's maximum depth D, the minimum number of samples required to split an internal node n_s, and the minimum number of samples required to be at a leaf node n_r.

A random forest (RF) is a meta-estimator that fits several decision tree classifiers to various subsamples of the dataset. It is also known as a random decision forest (RDF) that uses the mode of the classification to improve the predictive accuracy and control the problem of over-fitting [51]. The number of trees in the forest is a hyperparameter to be tuned, in addition to those hyperparameters for a decision tree.

Appendix A.5. Multilayer Perceptron Artificial Neural Network (MLP-ANN)

A multilayer perceptron artificial neural network (MLP-ANN) is a supervised learning algorithm that learns a function $f()$: $R^m \rightarrow R^o$ by training a dataset R with m-dimensional input and o-dimensional output. It can also learn a nonlinear function approximated for predicting the output. As ANNs do not have predefined assumptions, they have a low sensitivity to error term assumptions and high tolerance to noise. Therefore, an MLP-ANN can be used to examine the relationships in complex nonlinear datasets in the same way as conventional statistical techniques, but without many of the parametric restrictions about the nature of the data relationships [29]. The algorithm is described by Equation (A14), where J is the local gradient of function f concerning parameters β, y is independent variables and δ is the increment.

$$\left(J^T J + \lambda diag\left(J^T J\right)\right)\delta \;=\; J^T[y - f(B)] \tag{A14}$$

The hyperparameters are adjusted for model performance. Hidden layer arrangement includes the number of hidden layers and the number of neurons in each hidden layer. The activation function of a neuron defines the output of that neuron given an input. Four activation functions (identity, logistic, *tanh* and rectified linear unit (ReLU)) used in this study are given in Equations (A15)–(A18).

$$f(x) \;=\; x \tag{A15}$$

$$f(x) \;=\; \frac{1}{1 + exp(-x)} \tag{A16}$$

$$tanh(x) \;=\; \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} \tag{A17}$$

$$f(x) \;=\; \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{A18}$$

Moreover, iterative methods adopted for training the neural networks (weight optimization) can be specified. The L-BFGS type quasi-Newton method calculates the second derivative of the objective function and that leads to a more efficient descent direction [52]. Stochastic gradient descent (SGD), by using an estimate calculated from a randomly selected subset of the data rather than the entire dataset, optimizes an objective function with differentiable smoothness properties [53]. Adaptive moment estimation (Adam) is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [54].

Learning rate determines the weight updates. The default value for the constant learning rate is 0.001 for all iterative methods. Optional weights are available for the stochastic gradient descent solver. An "invscaling" weight gradually decreases the learning rate at each time step using an inverse scaling exponent to the time step, while an "adaptive" weight keeps the learning rate constant, as long as the training loss keeps decreasing. Dividing the current learning rate by 5 is generally adopted for the adaptive weight.

Appendix B.

Table A1. Configuration sets of the model hyperparameters for the inner layer architecture for the MLP-ANN classifier.

Legend	Activation	*C*	Learning Rate	Solver	Legend	Activation	*C*	Learning Rate	Solver
1	identity	0.0001	constant	Adam	31	relu	0.05	adaptive	SDG
2	logistic				32	*tanh*			
3	relu				33	identity	1		
4	*tanh*				34	logistic			
5	identity	0.05			35	relu			
6	logistic				36	*tanh*			
7	relu				37	identity	0.0001		
8	*tanh*				38	logistic			
9	identity	1			39	relu			
10	logistic				40	*tanh*			
11	relu				41	identity	0.05	constant	
12	*tanh*				42	logistic			
13	identity	0.0001		LBFGS	43	relu			
14	logistic				44	*tanh*			
15	relu				45	identity	1		
16	*tanh*				46	logistic			
17	identity	0.05			47	relu			
18	logistic				48	*tanh*			
19	relu				49	identity	0.0001		
20	*tanh*				50	logistic			
21	identity	1			51	relu			
22	logistic				52	*tanh*			
23	relu				53	identity	0.05	invscaling	
24	*tanh*				54	logistic			
25	identity	0.0001	adaptive	SDG	55	relu			
26	logistic				56	*tanh*			
27	relu				57	identity	1		
28	*tanh*				58	logistic			
29	identity	0.05			59	relu			
30	logistic				60	*tanh*			

References

1. Klepeis, N.E.; Nelson, W.C.; Ott, W.R.; Robinson, J.P.; Tsang, A.M.; Switzer, P.; Behar, J.V.; Hern, S.C.; Engelmann, W.H. The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. *J. Expo. Sci. Environ. Epidemiol.* **2011**, *11*, 231–252. [CrossRef] [PubMed]
2. Burroughs, H.E.; Hansen, S.J. *Managing Indoor Air Quality*; Fairmont Press: Lilburn, GA, USA, 2001.
3. Brown, S.K. *Indoor Air Quality. Australia: State of the Environment Technical Paper Series (Atmosphere)*; Department of the Environment, Sport and Territories: Canberra, Australia, 1997.
4. Husman, T.M. The Health Protection Act, national guidelines for indoor air quality and development of the national indoor air programs in Finland. *Environ. Health Perspect.* **1999**, *107* (Suppl. S3), 515–517. [CrossRef] [PubMed]
5. Azuma, K.; Uchiyama, I.; Ikeda, K. The regulations for indoor air pollution in Japan: A public health perspective. *J. Risk Res.* **2008**, *11*, 301–314. [CrossRef]
6. Aurola, R.; Valikyla, T. (Eds.) *Guidelines for Healthy Housing*; Ministry of Social Affairs and Health: Pori, Finland, 1997. (In Finnish)
7. Ad-hoc-Arbeitsgruppe IRK-AGLMB. Guideline values for indoor air: General Scheme. *Bundesgesundheitsblatt* **1996**, *39*, 422–426. (In German)
8. Meyers, R.A. *Encyclopedia of Physical Science and Technology*; Academic Press: San Diego, CA, USA, 2002.
9. Schell, M.; Int-Hout, D. Demand Control Ventilation Using CO_2. *ASHRAE J.* **2001**, *43*, 18–29.
10. Hui, P.S.; Wong, L.T.; Mui, K.W. Feasibility study of an Express Assessment Protocol for the indoor air quality of air-conditioned offices. *Indoor Built Environ.* **2006**, *15*, 373–378. [CrossRef]
11. Wong, L.T.; Mui, K.W.; Hui, P.S. A statistical model for characterizing common air pollutants in air-conditioned offices. *Atmos. Environ.* **2006**, *40*, 4246–4257. [CrossRef]
12. Indoor Air Quality Management Group. *Practice Note for Managing Air Quality in Air-Conditioned Public Transport. Facilities*; Environmental Protection Department: Hong Kong, China, 2003.

13. Wong, L.T.; Mui, K.W.; Hui, P.S. Screening for indoor air quality of air-conditioned offices. *Indoor Built Environ.* **2007**, *16*, 438–443. [CrossRef]
14. Mui, K.W.; Hui, P.S.; Wong, L.T. Diagnostics of unsatisfactory indoor air quality in air-conditional workplaces. *Indoor Built Environ.* **2011**, *20*, 313–320. [CrossRef]
15. Wong, L.T.; Mui, K.W.; Tsang, T.W. Evaluation of indoor air quality screening strategies: A step-wise approach for IAQ screening. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1240. [CrossRef]
16. WHO Regional Office for Europe. *Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2006.
17. WHO Regional Office for Europe. *Review of Evidence on Health Aspects of Air Pollution—REVIHAAP Project: Final Technical Report*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2013.
18. WHO Regional Office for Europe. *Health Risks of Air Pollution in Europe—HRAPIE Project. Recommendations for Concentration–Response Functions for Cost–Benefit Analysis of Particulate Matter, Ozone and Nitrogen Dioxide*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2013.
19. WHO Regional Office for Europe. *Evolution of WHO Air Quality Guidelines: Past, Present and Future*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2017.
20. WHO. *WHO Global Air Quality Guidelines. Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; World Health Organization: Geneva, Switzerland, 2021.
21. Rybarczyk, Y.; Zalakeviciute, R. Machine learning approaches for outdoor air quality modelling: A systematic review. *Appl. Sci.* **2018**, *8*, 2570. [CrossRef]
22. Seyedzadeh, S.; Rahimian, F.; Glesk, I.; Roper, M. Machine learning for estimation of building energy consumption and performance: A review. *Vis. Eng.* **2018**, *6*, 5. [CrossRef]
23. Wei, W.; Ramalho, O.; Malingre, L.; Sivanantham, S.; Little, J.C.; Mandin, C. Machine learning and statistical models for predicting indoor air quality. *Indoor Air* **2019**, *29*, 704–726. [CrossRef] [PubMed]
24. Elbayoumi, M.; Ramli, N.A.; Fitri Md Yusof, N.F. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM2.5–10 and PM2.5 concentrations in naturally ventilated schools. *Atmos. Pollut. Res.* **2015**, *6*, 1013–1023. [CrossRef]
25. Yuchi, W.; Gombojav, E.; Boldbaatar, B.; Galsuren, J.; Enkhmaa, S.; Beejin, B.; Naidan, G.; Ochir, C.; Legtseg, B.; Byambaa, T.; et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ. Pollut.* **2019**, *245*, 746–753. [CrossRef]
26. Park, S.; Kim, M.; Kim, M.; Namgung, H.G.; Kim, K.T.; Cho, K.H.; Kwon, S.B. Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). *J. Hazard. Mater.* **2018**, *341*, 75–82. [CrossRef]
27. Skön, J.; Johansson, M.; Raatikainen, M.; Leiviskä, K.; Kolehmainen, M. Modelling indoor air carbon dioxide (CO_2) concentration using neural network. *World Acad. Sci. Eng. Technol. Int. Sci. Index.* **2012**, *6*, 737–741.
28. Khazaei, B.; Shiehbeigi, A.; Haji Molla Ali Kani, A.R. Modeling indoor air carbon dioxide concentration using artificial neural network. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 729–736. [CrossRef]
29. Challoner, A.; Pilla, F.; Gill, L. Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings. *Int. J. Environ. Res. Public Health* **2015**, *12*, 15233–15253. [CrossRef]
30. Kropat, G.; Bochud, F.; Jaboyedoff, M.; Laedermann, J.P.; Murith, C.; Palacios, M. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. *J. Environ. Radioact.* **2015**, *147*, 51–62. [CrossRef]
31. Kropat, G.; Bochud, F.; Jaboyedoff, M.; Laedermann, J.P.; Murith, C.; Gruson, M.P.; Baechler, S. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. *Sci. Total Environ.* **2015**, *505*, 137–148. [CrossRef] [PubMed]
32. Ahn, J.; Shin, D.; Kim, K.; Yang, J. Indoor air quality analysis using deep learning with sensor data. *Sensors* **2017**, *17*, 2476. [CrossRef] [PubMed]
33. Saini, J.; Dutta, M.; Marques, G. Indoor air quality prediction systems for smart environments: A systematic review. *J. Ambient Intell. Smart Environ.* **2020**, *12*, 433–453. [CrossRef]
34. Montgomery, D.C.; Jennings, C.L.; Kulahci, M. *Introduction to Time Series Analysis and Forecasting*; John Wiley & Sons: New York, NY, USA, 2008.
35. Yu, T.C.; Lin, C.C. An intelligent wireless sensing and control system to improve indoor air quality: Monitoring, prediction, and preaction. *Int. J. Distrib. Sens. Netw.* **2015**, *11*, 140978. [CrossRef]
36. Han, Z.; Gao, R.X.; Fan, Z. Occupancy and indoor environment quality sensing for smart buildings. In Proceedings of the 2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings, Congress Graz, Graz, Austria, 13–16 May 2012; IEEE: Piscataway, NJ, USA, 2012.
37. Ouaret, R.; Ionescu, A.; Petrehus, V.; Candau, Y.; Ramalho, O. Spectral band decomposition combined with nonlinear models: Application to indoor formaldehyde concentration forecasting. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 985–997. [CrossRef]
38. Zimmerman, N.; Presto, A.A.; Kumar, P.N.; Gu, J.; Hauryliuk, A.; Robinson, E.S.; Robinson, A.L.; Subramanian, R. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmos. Meas. Tech.* **2018**, *11*, 291–313. [CrossRef]

39. Leong, W.C.; Kelani, R.O.; Ahmad, Z. Prediction of air pollution index (API) using support vector machine (SVM). *J. Environ. Chem. Eng.* **2020**, *8*, 103208. [CrossRef]
40. Sarkhosh, M.; Najafpoor, A.A.; Alidadi, H.; Shamsara, J.; Amiri, H.; Andrea, T.; Kariminejad, F. Indoor Air Quality associations with sick building syndrome: An application of decision tree technology. *Build. Environ.* **2021**, *188*, 107446. [CrossRef]
41. Indoor Air Quality Management Group. *A Guide on Indoor Air Quality Certification Scheme for Offices and Public Places*; Hong Kong Environmental Protection Department, Government of the Hong Kong Special Administrative Region: Hong Kong, China, 2019.
42. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
43. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
44. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233–234. [CrossRef] [PubMed]
45. Pecha, M.; Horák, D. Analyzing l1-loss and l2-loss Support Vector Machines Implemented in PERMON Toolbox. In *AETA 2018—Recent Advances in Electrical Engineering and Related Sciences: Theory and Application*; Zelinka, I., Brandstetter, P., Trong Dao, T., Hoang Duy, V., Kim, S., Eds.; Springer: Cham, Switzerland, 2020; pp. 13–23.
46. Adak, M.F.; Ercan, S. Identification of Indoor Harmful Gas to Human Respiratory System using Support Vector Machines. In Proceedings of the 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 1–13 October 2019; IEEE: Piscataway, NJ, USA, 2019.
47. Zhang, L.; Tian, F.; Nie, H.; Dang, L.; Li, G.; Ye, Q.; Kadri, C. Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine. *Sens. Actuators B Chem.* **2012**, *174*, 114–125. [CrossRef]
48. Intan, P.K. Comparison of Kernel Function on Support Vector Machine in Classification of Childbirth. *J. Mat. Mantik.* **2019**, *5*, 90–99. [CrossRef]
49. Imandoust, S.B.; Bolandraftar, M. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *Int. J. Eng.* **2013**, *3*, 605–610.
50. Schein, A.I.; Ungar, L.H. Active learning for logistic regression: An evaluation. *Mach. Learn.* **2007**, *68*, 235–265. [CrossRef]
51. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; IEEE: Piscataway, NJ, USA, 1995.
52. Bollapragada, R.; Nocedal, J.; Mudigere, D.; Shi, H.J.; Tang, P.T.P. A progressive batching L-BFGS method for machine learning. In Proceedings of the International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018.
53. Bottou, L. Stochastic gradient learning in neural networks. In Proceedings of the Neuro-Nımes, Nimes, France, 12–16 November 1990; EC2: Nanterre, France, 1991.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.