Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9404807, 16 pages
https://doi.org/10.1155/2022/9404807

Research Article

Machine Learning-Based CO_2 Prediction for Office Room: A Pilot Study

Nishant Raj Kapoor,[1,2] Ashok Kumar,[1,2] Anuj Kumar,[2,3] Aman Kumar,[2,4] Mazin Abed Mohammed,[5] Krishna Kumar,[6] Seifedine Kadry,[7] and Sangsoon Lim[8]

[1] *Architecture and Planning Department, CSIR-Central Building Research Institute, Roorkee 247667, India*
[2] *Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*
[3] *Building Energy Efficiency Division, CSIR-Central Building Research Institute, Roorkee 247667, India*
[4] *Structural Engineering Department, CSIR-Central Building Research Institute, Roorkee 247667, India*
[5] *College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Iraq*
[6] *Department of Hydro and Renewable Energy, Indian Institute of Technology, Roorkee 247667, India*
[7] *Department of Applied Data Science, Noroff University College, Kristiansand, Norway*
[8] *Department of Computer Engineering, Sungkyul University, Anyang 14097, Republic of Korea*

Correspondence should be addressed to Sangsoon Lim; slim@sungkyul.ac.kr

Received 25 January 2022; Revised 20 February 2022; Accepted 22 February 2022; Published 7 March 2022

Academic Editor: Isma F Siddiqui

Air pollution is increasing profusely in Indian cities as well as throughout the world, and it poses a major threat to climate as well as the health of all living things. Air pollution is the reason behind degraded indoor air quality (IAQ) in urban buildings. Carbon dioxide (CO_2) is the main contributor to indoor pollution as humans themselves are one of the generating sources of this pollutant. The testing and monitoring of CO_2 consume cost and time and require smart sensors. Thus, to solve these limitations, machine learning (ML) has been used to predict the concentration of CO_2 inside an office room. This study is based on the data collected through real-time measurements of indoor CO_2, number of occupants, area per person, outdoor temperature, outer wind speed, relative humidity, and air quality index used as input parameters. In this study, ten algorithms, namely, artificial neural network (ANN), support vector machine (SVM), decision tree (DT), Gaussian process regression (GPR), linear regression (LR), ensemble learning (EL), optimized GPR, optimized EL, optimized DT, and optimized SVM, were used to predict the concentration of CO_2. It has been found that the optimized GPR model performs better than other selected models in terms of prediction accuracy. The result of this study indicated that the optimized GPR model can predict the concentration of CO_2 with the highest prediction accuracy having R, RMSE, MAE, NS, and a20-index values of 0.98874, 4.20068 ppm, 3.35098 ppm, 0.9817, and 1, respectively. This study can be utilized by the designers, researchers, healthcare professionals, and smart city developers to analyse the indoor air quality for designing air ventilation systems and monitoring CO_2 level inside the buildings.

1. Introduction

Human health, performance, satisfaction, and productivity inside built environments are affected primarily by indoor environment quality (IEQ). Among major IEQ parameters like thermal comfort, acoustic comfort, and visual comfort, the indoor air quality (IAQ) is directly linked to sick building syndrome (SBS) which affects occupants' comfort and health negatively [1–4]. Generally, the most effective and usual method to improve IAQ is to bring in enough fresh air from outside [5]. Previous research, however, has revealed that the majority of the focus on air quality had been on the outdoor air quality and its impact on human health [6–8]. This can also be seen clearly in the history of the Clean Air Act, which began with the Air Pollution Control Act of 1955 and then tailed by the 1963 and 1970 Clean

Air Acts, the Air Quality Act of 1967, and the amendments in 1977 and 1990, all of which focused on pollution control from outdoor sources [9]. The major pollution in urban areas and big cities is primarily due to transportation. Massive investment to develop transport networks and infrastructure along with a growing economy is inevitable in developing countries. With the rapid growth of travel demand along with better finance and investment options, vehicular air pollution is emerging and dominating other pollution sources in cities. Transportation is the prime sector contributing to the air pollution in urban agglomerations followed by the industry and the agriculture sector. This growth leads to more emissions of pollutants in the air. This polluted air has mild to severe adverse effects on all living beings depending upon the duration of exposure, the concentration of the pollutants in the air, and the health status of the living one. While most people know the effects of air pollution, less are aware that the quality of their indoor air may be worse than that of their outside air. Around the world, an increasing proportion of the population works in office buildings. Existing research on office building IAQ has concentrated on particular concerns such as photocopier and printer emissions as well as some other indoor sources. Workplaces and offices are generally situated near busy roads and marketplaces for economic reasons. The outer air pollutant enters the building and enhances the concentration of indoor air pollutants (IAPs). IAPs have recently been acknowledged as having an equivalent impact on human health as outside air pollution [4]. People spend most of their waking time inside different types of buildings and full sleeping time in residential buildings majorly. IAPs can be classified into 3 categories [10] (i) gases, (ii) biological contaminants, and (iii) particulate matter. Adequate outdoor fresh air is necessary to ensure excellent IAQ. If outdoor air quality is not good, then it is difficult to maintain good IAQ in naturally ventilated buildings. Throughout the globe, investigations on volatile organic compounds (VOCs), aldehydes, ammonia, particulate matter (PM), and other pollutants in office buildings were conducted. BASE, IAQ-AUDIT, HOPE, AIRMEX, and OFFICAIR are some of the major milestone projects in the development of the existing knowledge on IAQ in the office-built environments [11]. Some of these studies also included energy efficiency, occupant performance, and satisfaction as additional important parameters along with IAQ. Apart from these, several studies have been undertaken to forecast the ventilation performance of buildings and occupant's perceptions [12–14]. Low ventilation rates in houses are linked to asthma and allergic symptoms [15]. Inadequate ventilation will result in sick building syndrome (SBS), and excessive outdoor air will result in increased energy demand for buildings to maintain thermal comfort indoors. If the nearby roads are busy, then it is not easy to prevent degradation of IAQ when using natural ventilation to ventilate the stale air out having a high concentration of unwanted gases like CO_2 and other resuspended particulate matters due to worker's activity. Outer CO_2 concentration and conditions affect inner concentrations of CO_2 along with inner sources like humans and other indoor anthropogenic activities. It is also not feasible to open windows and doors in office buildings situated in or near any noisy area as this results in hampering the concentration and performance of the worker (due to reduced acoustic comfort) inside the building. Increased IAPs lead towards health issues in building occupants. SBS is a phenomenon in which inhabitants of a structure may have a feeling of discomfort along with a variety of health symptoms that cannot be ascribed to a single cause or sickness and which generally improve once they left that particular building and space [16]. Apart from SBS, some occupants are affected by building-related illness (BRI) which affects for a longer duration than SBS. According to a report in 2010 by the WHO [17], IAPs are the main reason behind 2.7% of all the diseases globally. Additionally, a report in 2018 presented by the WHO [18] revealed that 3.8 million people die every year due to diseases that can be attributed to poor indoor environments. The usage of motorised vehicles such as heavy-duty and light-duty vehicles contributes to ambient air pollution caused by traffic activities. Carbon compounds, hydrocarbons, nitrogen oxides, sulphur oxides, particulate matter (PM) with a diameter less than 2.5 μm ($PM_{2.5}$) as well as diameter less than 10 μm (PM_{10}), and ultrafine particles (UFP) are only a few of the pollutants released by these vehicles [19]. After infiltration or by natural ventilation and wind, these pollutants enter inside the built environment and enhance IAP concentration. Additionally, pollution from indoor sources also nudges the poor IAQ conditions. Worldwide studies have been done to address ventilation strategies to reduce CO_2 and other pollutants using artificial intelligence (AI). Fuzzy logic (FL), artificial intelligence (AI), and genetic algorithms (GA) are commonly used in intelligent control modelling for enhancing IAQ. Vanus et al. [20] predicted CO_2 levels inside smart homes considering relative humidity and temperature as input using the decision tree regression method. Pantazaras et al. [21] look at the possibility of developing predictive models that are suited to certain regions in order to forecast future CO_2 concentrations in their study. Kallio et al. [22] predicted CO_2 in the office environment. Their study investigated the suitability of four ML approaches for simulating the future CO_2 concentration in the indoor office environment: multilayer perceptron, random forest, decision tree, and ridge regression. The authors explore that the decision tree model was equally accurate as of the computationally more difficult random forest model. Khazaei et al. [23] used ML to predict the concentration of CO_2 in indoor offices. On the basis of the mean-square-error approach, the authors determined that the most accurate model was the four-steps-ahead prediction model, which had an average difference of less than 17 ppm from the actual CO_2 content in the room. Skön et al. [24] modelled CO_2 concentration in apartment buildings using artificial neural networks. They considered temperature and relative humidity as input parameters. Taheri and Razban [25] developed a dynamic indoor CO_2 model to predict CO_2 levels. The data set includes temperature, relative humidity, dew point, and CO_2. The authors compared six learning algorithms including multilayer perceptron (MLP), logistic regression (LR), gradient boosting (GB), random forest (RF), AdaBoost, and support vector machine

(SVM). The MLP surpasses other algorithms in terms of accuracy and can accurately forecast CO_2 behavior. Mohammadshirazi et al. [26] tested four different ML methods, rolling average, random forest, gradient boosting, and long short-term memory for the prediction of indoor concentration levels of carbon dioxide, total volatile organic compound, formaldehyde, PM_{10}, $PM_{2.5}$, PM_1, ozone, and nitrogen dioxide. The study concluded that the best approach for forecasting indoor pollutants was consistently long short-term memory, while the optimum combinations of input factors varied depending on the pollutant of interest. The predicted results show that the LSTM training and validation MSEs from interpolating data varied from 0.001 to 0.007 and 0.001 to 0.003. ML models were used by Lillstrang et al. [27] to forecast the quality of indoor air in smart campuses. Predicting energy loads and inferring space occupancy status are critical activities that increase building energy efficiency and user comfort. The findings can be used to assess and improve the quality of sensor-based indoor data used in machine learning models, to determine whether a data set is representative enough to build a model that is robust under changing building conditions, and to determine the appropriate number of sensors per space when constructing an indoor wireless sensor network. The proposed work uses artificial neural networks (ANN) and other machine learning methods to forecast CO_2 level inside an office building. As the severity of CO_2 concentration affects the human health. Every machine leaning method has some pros and cons. The performance of individual machine learning models depends on the type and number of data sets. Various literatures are available on predicting the CO_2 level inside buildings; however, an accurate mathematical model to determine the quantity of CO_2 emission is difficult, as the input parameters are complex in nature. Therefore, monitoring and prediction of CO_2 concentration inside buildings are an important aspect. The objective of this study is to address the research gaps identified from the selective literature review using the ANN and other ML methods to predict CO_2 concentration inside the office building. Also, the performance comparison of different machine learning models used for predicting the CO_2 level inside the building has been presented.

The main contribution in this study is listed below:

(i) Identified the critical parameters used as input for prediction the CO_2 concentration

(ii) The identified critical parameters used as input for predicting the CO_2 concentration are number of occupants, area per person, outdoor temperature, outer wind speed, relative humidity, and air quality index

(iii) Collected real-time CO_2 concentration data from the office building

(iv) Modelled six machine learning algorithms, namely, artificial neural network (ANN), support vector machine (SVM), decision tree (DT), Gaussian process regression (GPR), linear regression (LR), and ensemble learning (EL), and four optimized machine learning algorithms GPR, EL, DT, and SVM for predicting the CO_2 concentration inside the office building

The work in the research article is divided into five sections: Section 2 provides the details of data generation, data normalization, and performance indices which are used to evaluate the prediction of ML models. Section 3 describes all the machine learning (ML) approaches. The results and discussion part are presented in Section 4, and the conclusion of this study is presented in Section 5.

2. Materials and Methods

To predict the CO_2 concentration inside the office room due to internal emission, exterior transportation movement and industry emissions are studied in this article. The total number of 169 data sets was used to construct the prediction models which include the input variables such as temperature, relative humidity, air quality index, wind speed, occupancy, area per person, and one output variable, that is, carbon dioxide. The concentration of carbon dioxide inside the room mainly affected the occupancy inside the room as humans themselves are the emitting source. The concentration of CO_2 inside any building also depends upon the exterior environment surrounding the building. Buildings near industrial areas and busy roads are mostly seen affected by the pollutants.

2.1. Data Generation. The data used in this work was generated in the lab of CSIR. The area of the office room was approximately 24 m^2. The office is situated on the ground floor and contains one window on the north-faced wall having a width and height of 2.5 m and 1.5 m, respectively. The office room contains two doors; the dimensions of door 1 and door 2 are 1.2 m × 2.9 m and 0.9 m × 2.0 m, respectively. The 3D diagram of the office room with two doors and one window with the arrangement of furniture is shown in Figure 1. The collected seven parameters are indoor carbon dioxide (CO_2), number of occupants (O), area per person (A), outdoor temperature (T_o), outer wind speed (W_S), relative humidity (RH), and air quality index (AQI). The data were collected six times a day. The timing of the data collection is shown in Figure 2, where the watch represents the office hours 9 AM to 6 PM. One reading was recorded every day after one hour later than office hours at 7 PM.

The maximum observed CO_2 level inside the office room was 572 ppm, while the minimum value observed was 445 ppm. The other statistical analysis of the collected data such as minimum value, maximum value, mean, standard deviation, kurtosis, and skewness for input and output database is shown in Table 1. Figure 3 shows the distribution of all data in terms of contributing parameters on CO_2.

2.2. Evaluation Criteria. For evaluating the accuracy of ML models, four commonly used performance indices such as correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), mean square error (MSE), mean absolute percentage error (MAPE), Nash-Sutcliffe (NS) efficiency index, and a20-indices were used. Equations (1)–(7)

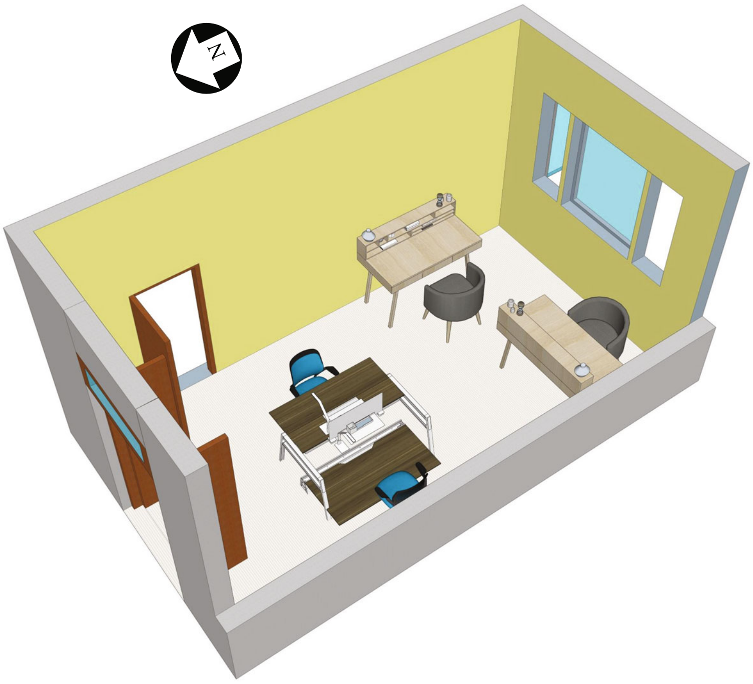


Figure 1: 3D diagram of an office room.

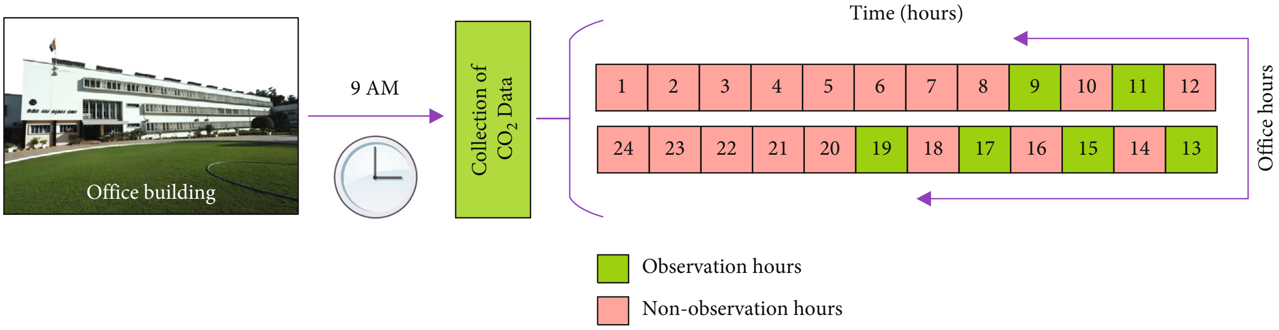
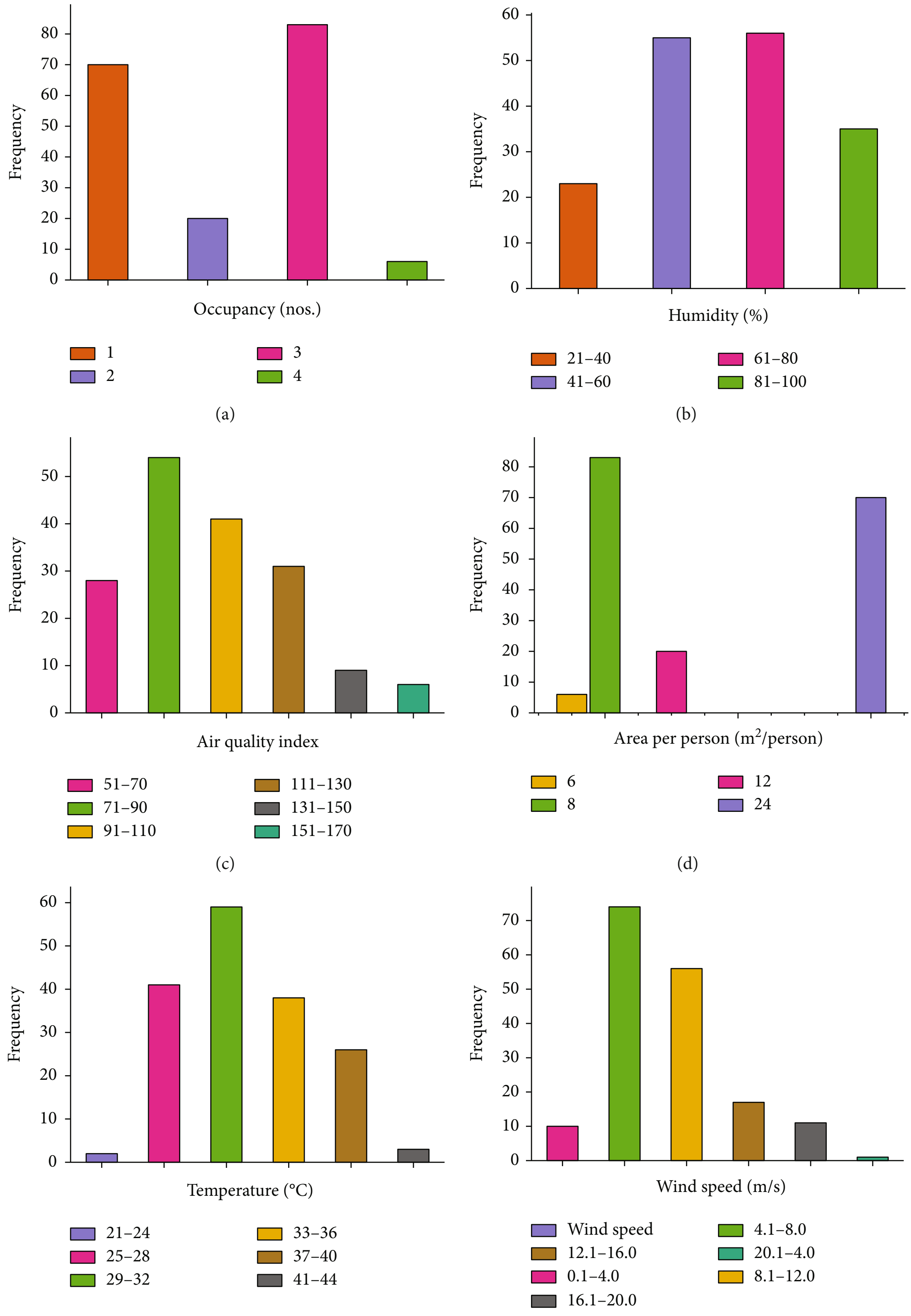


Figure 2: Data collection details (hrs).

Table 1: Statistical data for every major parameter in the study.

Parameter	Symbol	Unit	Min	Max	Mean	Std.	Kurtosis	Skewness	Type
Occupants	O	Nos.	1	4	2.1479	1.0157	1.3274	-0.0938	Input
Office area	A_O	m^2	6	24	14.7929	7.8322	1.1397	0.3022	Input
Air quality	AQI	—	51	155	94.8639	24.7369	2.6583	0.5255	Input
Air temp.	T	˚C	22	42	31.6923	4.0473	2.3903	0.2013	Input
Rel. humidity	RH	%	21	100	63.0178	19.1552	2.1424	-0.0733	Input
Wind speed	W_S	m/s	1.3	23.1	8.7183	3.9735	3.3857	0.8041	Input
Indoor CO_2	CO_2	Ppm	445	572	509.7515	27.0277	2.2640	0.0202	Output



(a)

(b)

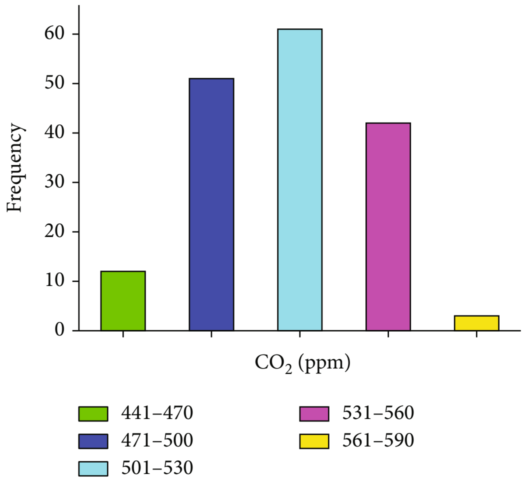
(c)

(d)

(e)

(f)

Figure 3: Continued.



(g)

Figure 3: Frequency distribution of collected data.

[28] represent the performance indices considered in this study.

$$R = \frac{\sum\left(A_s - \bar{A}\right) \times \left(P_s - \bar{P}\right)}{\sqrt{\sum\left(A_s - \bar{A}\right)^2 \times \sum\left(P_s - \bar{P}\right)^2}}, \quad (1)$$

$$\mathrm{MAE} = \frac{1}{T}\sum_{s=1}^{T}\left|A_s - P_s\right|, \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{s=1}^{T}\left(A_s - P_s\right)^2}{T}}, \quad (3)$$

$$\mathrm{MSE} = \frac{1}{T}\sum_{s=1}^{T}\left(A_s - P_s\right)^2, \quad (4)$$

$$\mathrm{MAPE} = \frac{1}{T}\sum_{s=1}^{T}\left|\frac{A_s - P_s}{A_s}\right| \times 100, \quad (5)$$

$$\mathrm{NS} = 1 - \frac{\sum_{i=1}^{T}\left(A_s - P_s\right)^2}{\sum_{i=1}^{T}\left(A_s - \overline{P_s}\right)^2}, \quad (6)$$

$$\text{a20-index} = \frac{m20}{N}, \quad (7)$$

where T is the number of samples in the data set, A_s is the measured values, P_s is the predicted values, and $\overline{P_s}$ is the mean of the predicted values. m20 is the number of samples with value rates measured/predicted values (range between 0.8 and 1.2).

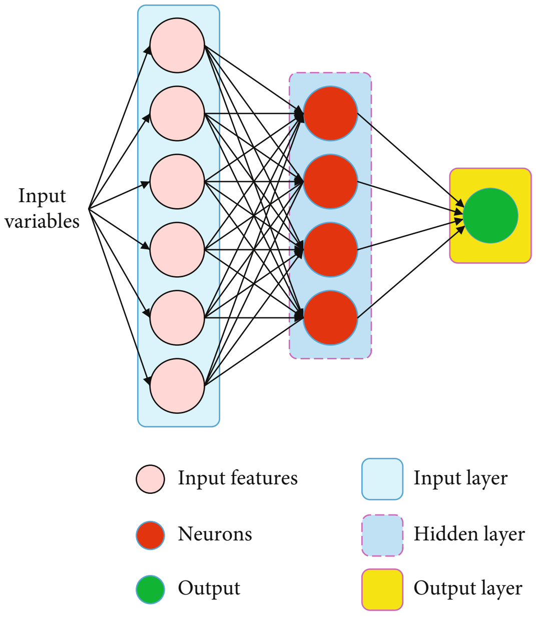


Figure 4: Basic structure of ANN.

2.3. Normalization of Selected Data. Data normalization was performed to decrease undesirable feature scaling effects and increase computational stability. In this work, data was normalized in the range of 0 and 1 using equation (8) [29].

$$Y^* = \frac{\left(y_{I,i} - y_{I,\min}\right)}{\left(y_{I,\max} - y_{I,\min}\right)}, \quad (8)$$

where y_I, is the measured value (given value of CO_2) of the I^{th} input ($I = 1, 2, 3, 4, 5, 6$) in the i^{th} database ($i = 1, 2, \cdots, 169$). $y_{I,\max}$ and $y_{I,\min}$ are the maximum and minimum values in the I^{th} input, respectively.

3. Machine Learning Algorithms

In the literature, there are various studies available that measured the concentration of CO_2 inside different types of

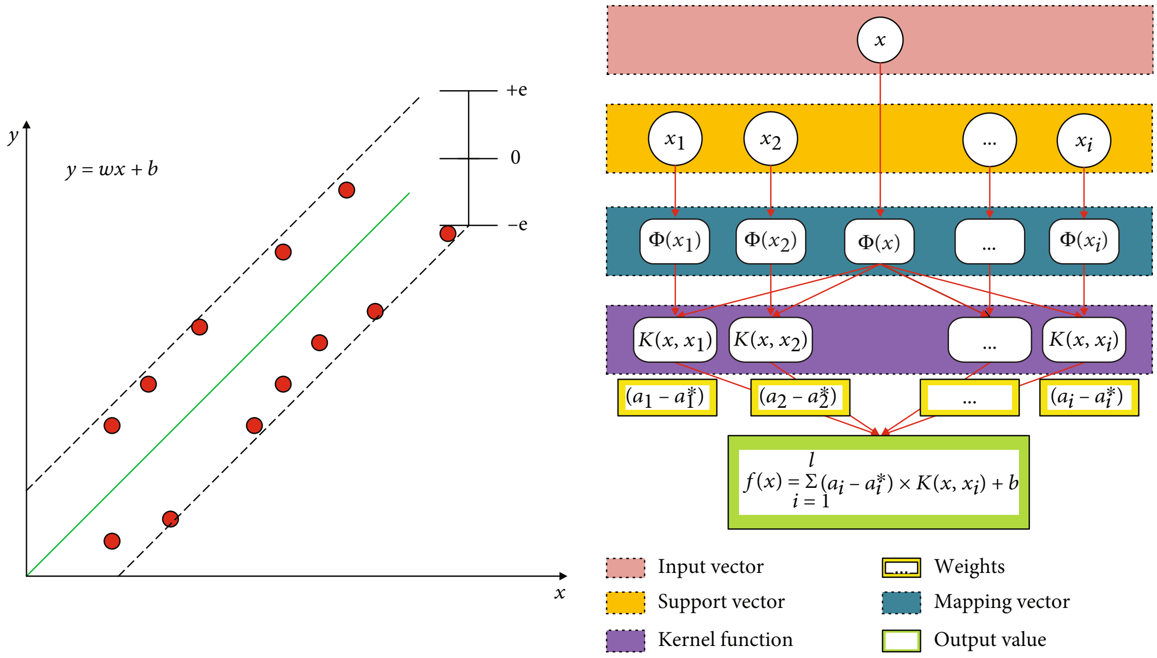


Figure 5: Graphical representation of SVM regression [31].

buildings using ML algorithms. But very limited studies are available to predict the effect of outside pollution on the inside concentration of CO_2. In this study, ML algorithms are used to predict the concentration of CO_2 in the office building and suggest the best-predicted model. Gaussian process regression (GPR), support vector machine (SVM), artificial neural network (ANN), decision tree (DT), ensemble learning (EL), and linear regression (LR) are used to predict the concentration of CO_2. In addition to that, optimized algorithms such as GPR, SVM, EL, and DT are also used. Each model's predictions were analysed and compared to achieve the best accurate model.

3.1. Artificial Neural Network (ANN). In the late 19^{th} and early 20^{th} centuries, the groundwork for the area of ANN was done. This mostly comprised of psychology, neurophysiology, and physics multidisciplinary work. This early study focused on general learning, vision, conditioning, and other ideas, rather than particular mathematical models of neuron activity. The field of neural networks has been revitalized as a result of these new advances. Many studies have been published in the previous two decades, and many different forms of ANNs have been studied. ANN models were first employed in the ecological field in the early 1990s, but they became increasingly popular in the late 1990s.

Around 10^{10} neurons, or computing components, make up the human brain, which interacts via a connecting network. ANNs are parallel distributed computer networks that share certain fundamental features with biological neural systems. Neurons ($X = [x_1 ; x_2 ; \cdots ; x_n]$) receive a variety of signals as input. Every input is given a relative weight ($W = [w_1 ; w_2 ; \cdots ; w_n]$), which influences its impact. The strength of the input signal is determined by weights, which are adjustable coefficients inside the network. The summation block, which approximately corresponds to the actual cell body, generates the neuron output signal (NET), which algebraically sums all of the weighted inputs. The basic structure of ANN is presented in Figure 4.

Several types of ANNs have been produced over the last 10-15 years; however, two primary groups may be distinguished based on how the learning process is carried out: "In 'supervised learning,' a 'teacher' 'tells' the ANN how well it performs or what the right behaviour would have been throughout the learning phase." The ANN independently examines the features of the data set and learns to reflect these properties in its output in "unsupervised" learning. The relevant information regarding ANN is mentioned in the literature [30].

3.2. Support Vector Machine (SVM). SVM is a moderately new concept in the field of environmental science. In comparison to other disciplines, researchers employing remote sensing in environmental and ecological applications adopted SVMs initially, possibly due to the prompt growth of data-intensive technologies and the accompanying gap in the development of analytical tools. The application of SVMs in the environmental research domain has increased in recent years. SVMs are used in the detection of pollution, mapping of contaminated areas and disease distributions, and air quality estimates. Indeed, whenever there is high-dimensional data and a related lack of understanding about the underlying distribution, SVMs offer tremendous potential to resolve the ensuing data processing issues. The graphical representation of SVM regression is shown in Figure 5.

Support vector classification is based on a specific form of statistical learning machine, with Vapnik's supporting theory. Except for the assumption that the data are identically distributed and independent, support vector classification makes no assumptions about the underlying population's distribution. Furthermore, rather than

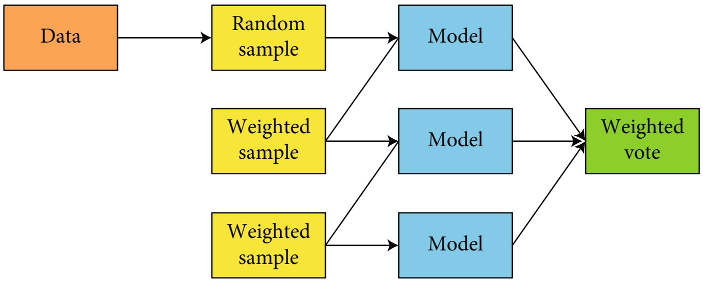


Figure 6: Boosted regression tree.

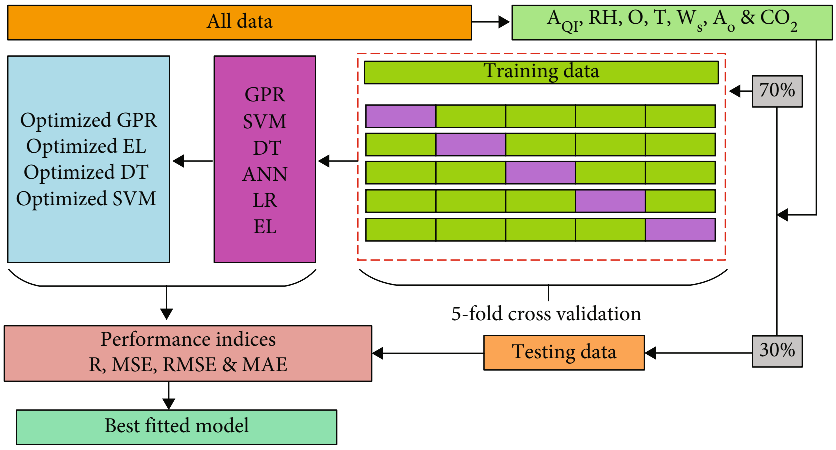


Figure 7: Framework of developing a CO_2 prediction model for office room.

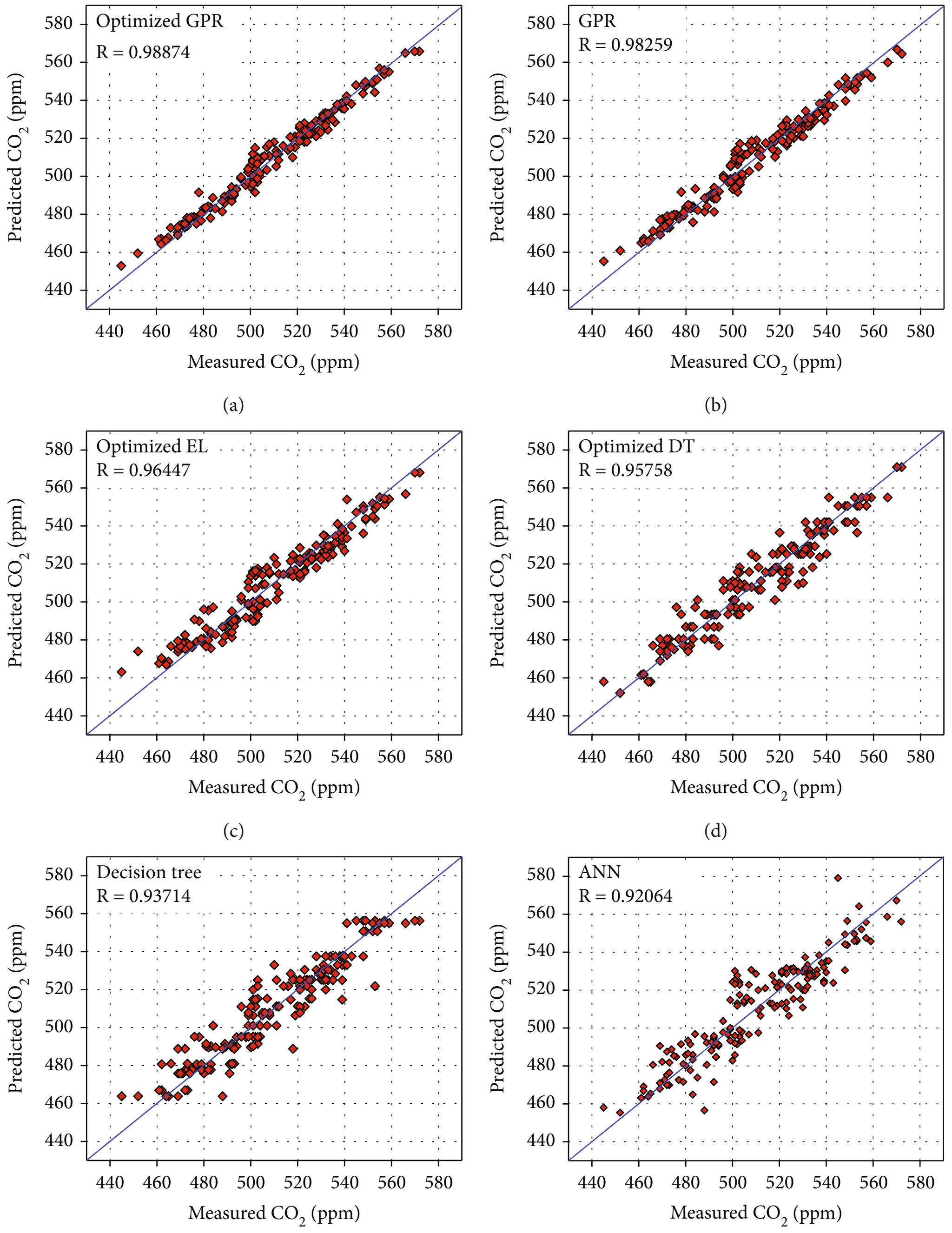
Table 2: Comparison of various models on the basis of statistical parameters.

S.N.	Methods	Statistical parameters for CO_2							Standard deviation
		R	R^2	MSE	RMSE	MAE	MAPE	NS	
1	Optimized GPR	0.98874	0.97761	17.64568	4.20068	3.35098	0.4325	0.9817	25.5432
2	GPR	0.98259	0.96548	27.08724	5.20454	4.25433	0.8393	0.9627	25.1313
3	Optimized EL	0.96447	0.93020	53.23267	7.29607	5.88308	1.1669	0.9267	24.4882
4	Optimized DT	0.95758	0.91696	60.30003	7.76531	6.00405	1.8131	0.9169	25.8812
5	DT	0.93714	0.87823	88.42094	9.40324	7.14055	1.4101	0.8782	25.3288
6	ANN	0.92064	0.84758	111.5761	10.56296	8.24138	1.7404	0.8314	25.7913
7	EL	0.89592	0.80267	156.4927	12.50970	9.97965	1.9695	0.7845	20.5966
8	LR	0.89566	0.80221	143.6322	11.98467	9.78463	1.9341	0.8022	24.2076
9	Optimized SVM	0.89349	0.79832	147.8662	12.16002	9.64578	1.9098	0.7964	24.5402
10	SVM	0.89044	0.79288	153.0833	12.37268	9.90023	1.9642	0.7892	24.1408

estimating error via asymptotic convergence to normalcy, SVMs use theorems limiting the real risk in terms of the empirical risk. As a result, even with tiny sample sizes, reliable estimates of the prediction error may be obtained without making any distributional assumptions. The ideal machine strikes a compromise between consistency in the training set and future data set generalization. Furthermore, SVMs allow us to avoid the degraded computing efficiency that is common in high-dimensional problems. Support vector classification is a suitable choice for the typically noisy, high-dimensional, and chaotic data encountered in environmental research because of these key features. More details of SVM can be found in [32].

3.3. Decision Tree (DT). One approach to demonstrate the links between samples in classification is to display them visually in the form of a "phylogenic tree." The tree-like structure represents the feature space in a hierarchical manner and helps to create links between the data: The characteristics identify the leaves of the tree, and all branches join together at the base, just as they do in a real tree. A DT, unlike a genuine tree, is generally portrayed as growing from top to bottom. Beginning at the root and working through the branches to a leaf that identifies the class, the membership of unknown data can be determined.

A DT is also a technique of encoding a series of choices produced by applying a set of classification rules in a



(a)

(b)

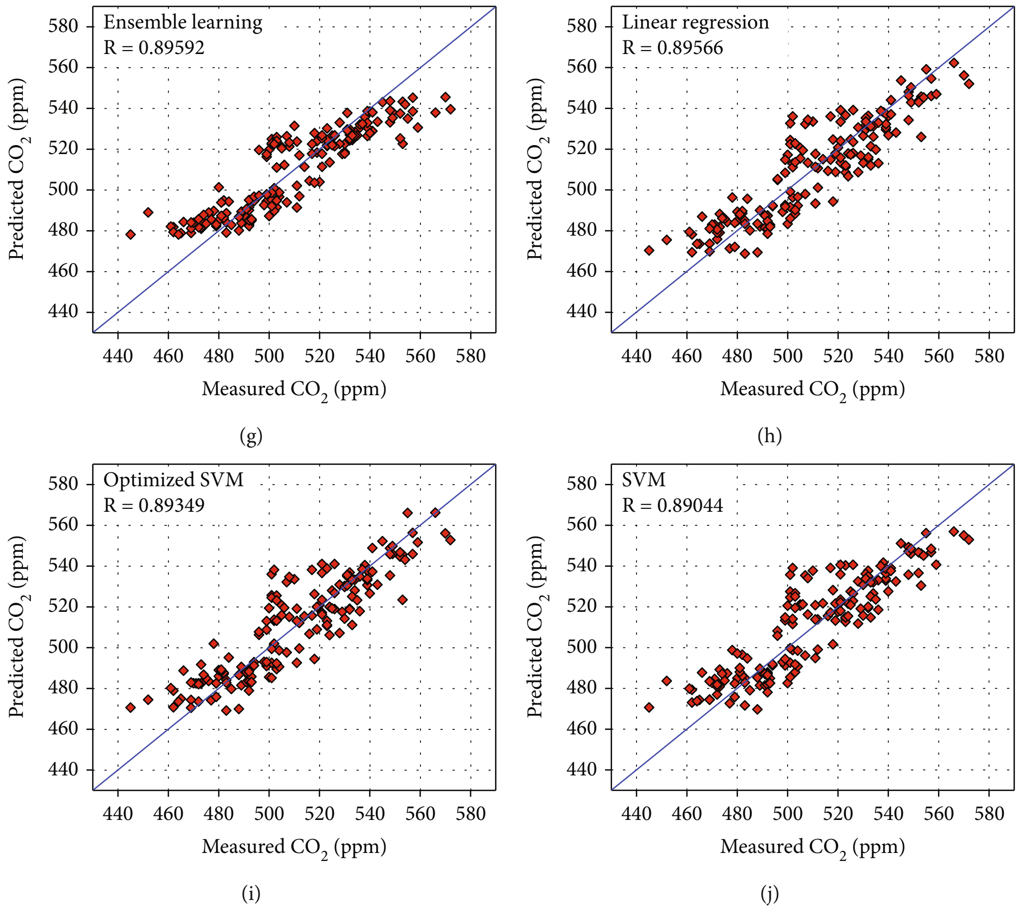
(c)

(d)

(e)

(f)

FIGURE 8: Continued.



(g) (h)

(i) (j)

Figure 8: Comparison of concentration of CO_2 prediction with various algorithms: (a) optimized GPR; (b) GPR; (c) optimized EL; (d) optimized DT; (e) decision tree; (f) ANN; (g) EL; (h) LR; (i) optimized SVM; (j) SVM.

sequential order to distinguish data. This method of classification has the benefit that, at least for small sets of rules, a graphical explanation of the set of rules is typically simple to comprehend. The discovery of such principles by methodical examination of the behavior of a set of known instances induces or creates a decision tree. More details of DT can be found in [33].

3.4. Gaussian Process Regression (GPR). GPR is the nonparametric, Bayesian method to regression and is used frequently in the field of ML. GPR offers numerous advantages, comprising the capacity to deal with tiny data sets along with providing uncertainty assessments on predictions. "Gaussian process regression is nonparametric (i.e., not constrained by a functional form), rather than computing the probability distribution of parameters of a single function, GPR computes the probability distribution of all admissible functions that fit the data. However, in order to calculate the posterior using the training data and compute the predicted posterior distribution on our points of interest, a prior specification (on the function space) is required." More information related to GPR can be found in [34]. Various studies on GPR are available in the literature for the forecasting of various parameters. Gaussian process regression (GPR) models have been widely used in ML applications because of their representation flexibility and inherent uncertainty measures over predictions.

3.5. Linear Regression (LR). In statistics and ML, LR is one of the most well-known and well-understood techniques. Many variables (or measures) are gathered for each individual or unit investigated in many scientific researches. Regression analysis is a statistical technique for predicting the value of one (or more) variables from a set of others. Basic or simple LR, multiple LR, and multivariate LR are the three forms of linear regression. The basic linear regression model is a model that is linear in these parameters. This model, often known as a straight-line model, is fitted using the least-squares method. When we want to model the link between one answer variable and more than one regression variable, we utilize multiple regression analysis. When we have more than one response variable and want to model the link between these variables and a collection of regression variables, we use multivariate multiple regression analysis. For more information, the reader should refer to a statistics textbook and [35].

3.6. Ensemble Learning (EL). EL is the process of intentionally generating and combining numerous models, such as classifiers or experts, to tackle a specific computational intelligence issue. Ensemble learning is generally used to increase the performance of a model (classification, prediction,

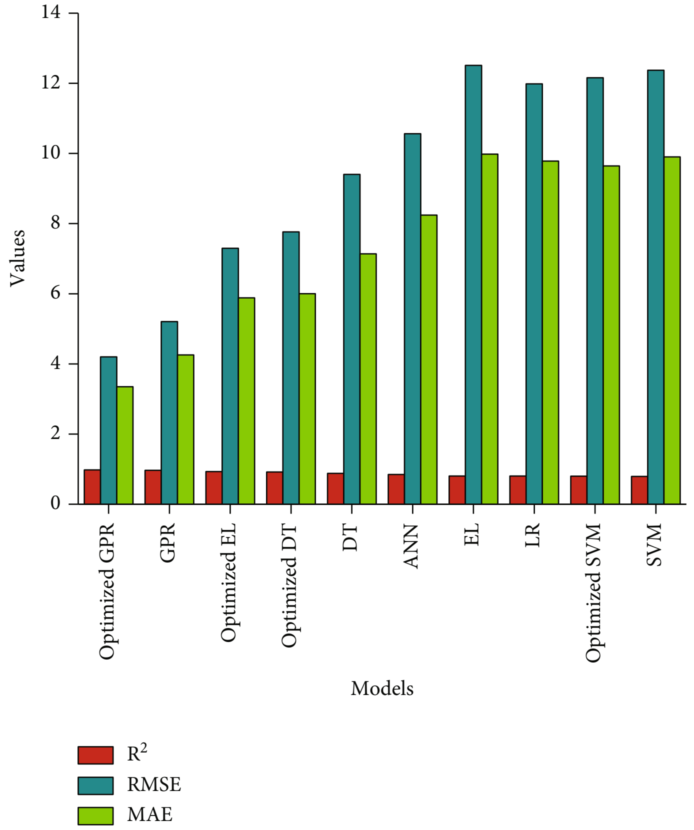


Figure 9: Statistical indicator for all ML models.

function approximation, etc.) or to minimize the chance of an unintentional selection of a poor one. Other uses of ensemble learning include providing a confidence level to the model's conclusion, nonstationary learning, data fusion, selecting optimal features, error correction, and incremental learning. The best-fitted model in the EL algorithm was boosted tree as shown in Figure 6 [36].

4. Results and Discussion

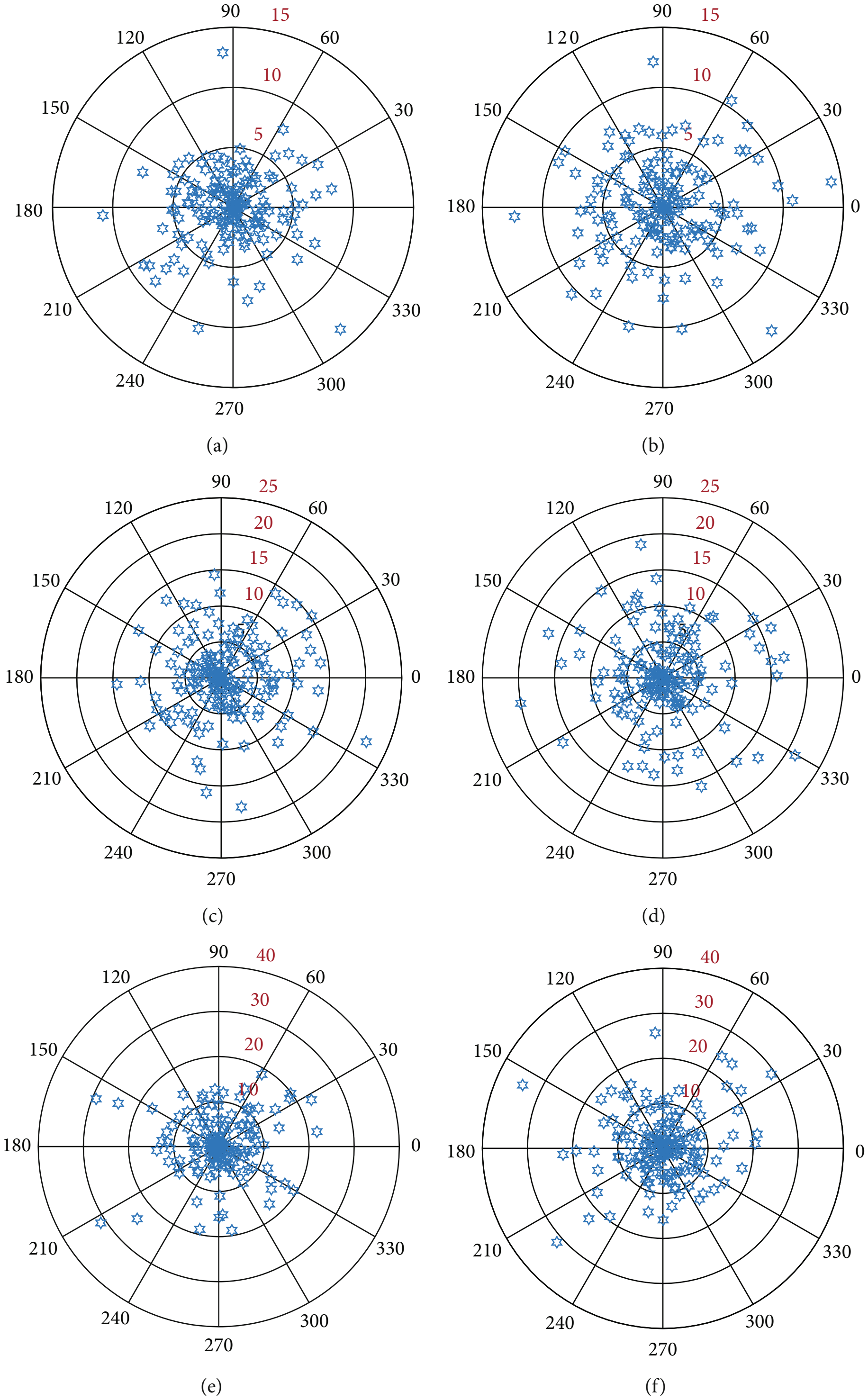
4.1. Implementation of Machine Learning Algorithms to Predict CO_2. Based on the training process of ML algorithms, the data were split into two ratios. To avoid the overfitting phenomenon, the distribution ratio of the two sets is adjusted to 7 : 3, 70% (130 samples) of the data used in the training process and the other 30% (56 samples) of the data utilized as testing data as shown in Figure 7. To validate the results of ML algorithms, the 5-fold cross-validation method is used. In 5-fold cross-validation, the data is further divided into 5 subsets. Then, each subset would be chosen in order for the validation process, the remaining 4 subsets being utilized for training inside the training stage.

4.2. Results of Machine Learning Models. The concentration of CO_2 inside the office room was predicted using various ML algorithms. The predicted values were compared with the actual results and estimated the errors based on performance indices.

4.2.1. ANN Model. A single hidden layer was investigated, with the number of neurons increased from 5 to 20, and the optimal model was discovered by trial and error. The performance parameters of the ANN model are presented in Table 2. As observed from the table, ANN attained almost 84.7% ($R^2 = 0.84758$) accuracy for the whole data set. In terms of MAE (8.24) and RMSE (10.56), the optimal feed-forward ANN structure with six inputs delivered the best training outcome. Figure 8(f) shows a comparison of the experimental and predicted values of the ANN model.

4.2.2. GPR Model. As we all know, the S (width of rbf) and ε (Gaussian noise) are two crucial factors that can be found by a trial-and-error procedure. $S = 0.40$ and $\varepsilon = 0.07$ are the final optimum values that are considered to design the optimum GPR model. In the training stage, the GPR model predicts the concentration of CO_2 practically perfectly, whereas, in the testing stage, there is a small difference. Figure 8(b) shows a comparison of the experimental and predicted values of the GPR model. The prediction accuracy of the GPR model with R^2, RMSE, and MAE values is 0.96548, 5.20, and 4.25, respectively, for the whole data set.

4.2.3. SVM Model. In the SVM model, the quadratic SVM is the best-fitted model and shows good accuracy. The values of the box constraint, epsilon, and kernel scale are 0.3, 0.03, and 1, respectively, for the best-fitted model. The comparison between predicted and experimental results is shown in Figure 8(j). The



(a)

(b)

(c)

(d)

(e)

(f)

Figure 10: Continued.

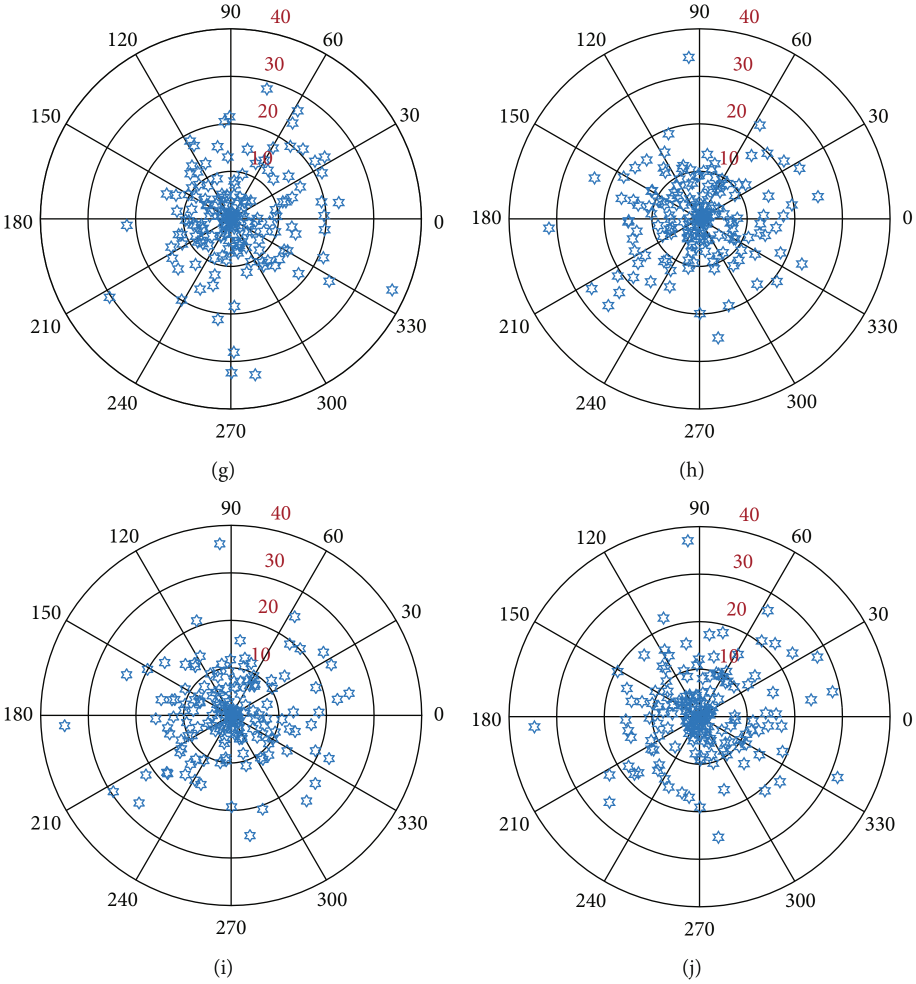


Figure 10: Representations of absolute error (CO_2 ppm): (a) optimized GPR; (b) GPR; (c) optimized EL; (d) optimized DT; (e) decision tree; (f) ANN; (g) EL; (h) LR; (i) optimized SVM; (j) SVM.

prediction accuracy of the SVM model is 79.29% with RMSE and MAE values being 12.51 and 9.98, respectively.

4.2.4. EL Model. In ensemble tree learning, the boosted tree is the best suitable model with minimum leaf size = 12, number of learners = 40, and learning rate = 0.2. The prediction accuracy of the EL model is 80.27% with RMSE and MAE values being 12.37 and 9.90, respectively. Figure 8(g) shows a comparison of the experimental and predicted values of the GPR model.

4.2.5. LR Model. In the linear regression model, the robust linear shows good results among linear, interaction linear, and stepwise linear models. The comparison between predicted and experimental results is shown in Figure 8(h). The prediction accuracy of the EL model is 80.22% with RMSE and MAE values being 11.98 and 9.78, respectively.

4.2.6. DT Model. In the decision tree model, the fine tree model shows the perfected fitted model than a medium and coarse tree. The minimum leaf size = 5 and maximum surrogates per node = 10 were considered as the best-suited hyperparameters. The prediction accuracy of the DT model is 88.42% with RMSE and MAE values being 9.40 and 7.14, respectively. The comparison between predicted and experimental results is shown in Figure 8(e).

The comparison between R^2, RMSE, and MAE indicators of different algorithms is presented in Figure 8. The optimized GPR model R^2 value is the highest among all models, the R^2 of optimized GPR is 1.24% greater than GPR, 4.84% greater than optimized EL, 6.20% more than optimized DT, 10.16% more than DT, 13.30% more than ANN, 17.89% more than EL, 17.94% more than LR, 18.34% more than optimized SVM, and 18.89% more than SVM. Similarly, both the performance indices RMSE and MAE of the optimized GPR model are the lowest among all the opted techniques and can be seen in the pictorial representation in Figure 8. On comparison, the RMSE values of GPR, optimized EL, optimized DT, DT, ANN, EL, LR, optimized SVM, and SVM models are 23.90%, 73.69%, 84.86%, 123.85%, 151.46%, 197.80%, 185.30%, 189.48%, and 194.54% higher than the optimized GPR model, respectively. Similarly comparing MAE values of GPR, optimized EL, optimized DT, DT, ANN, EL, LR, optimized SVM, and SVM models are 21.23%, 59.52%, 62.36%, 89.08%, 114.95%, 155.81%, 151.23%, 147.96%, and 153.94% higher than the optimized GPR model, respectively.

The best-predicted model is optimized GPR with the R-value of 0.98874, R^2 of 0.977607, MSE of 17.64568, and MAE of 3.350982 and having a standard deviation of 25.5432 as tabulated in Table 2. The worst prediction was from the SVM model with the R-value of 0.89044, R^2 of 0.792883, MSE of 153.0833, RMSE of 2.37268, and MAE of 9.900226 and having a standard deviation of 24.1408. The intermediate models analysed were GPR, optimized ensemble, optimized DT, DT, ANN, EL, LR, and optimized SVM with corresponding R-value of 0.98259, 0.96447, 0.95758, 0.93714, 92064, 0.89592, 0.89566, and 0.89349. The comparison between experimental and predicted optimized GPR, optimized EL, optimized DT, and optimized SVM is shown in Figures 8(a), 8(c), 8(d), and 8(i), respectively. The plotted graphs between measured CO_2 values and predicted output of indoor CO_2 by all the above-mentioned methods are presented below in Figure 8. The a20-index of all the ML models was having a value of 1. The performance of different ML models is shown in Figure 9.

In the optimized GPR model, the 96% data lies in the range of 10 ppm as shown in Figure 10(a). In the GPR model, the 94.67% data lies in the range of 10 ppm as shown in Figure 10(b). Similarly, in the optimized EL, optimized DT, and DT models, the data lies in the range of 10 ppm which is 82.84%, 79.29%, and 73.37%, respectively, as shown in Figures 10(c)–10(e). In the ANN model, the 70.79% data lies in the range of 10 ppm as shown in Figure 10(f). In EL, LR, optimized SVM, and SVM models, the data lies in the range of 10 ppm which is greater than 60% as presented in Figure 10.

5. Conclusion

In this study, the concentration of CO_2 inside an office room is evaluated using ANN, GPR, DT, EL, SVM, and LR algorithms along with optimized GPR, EL, DT, and SVM. A total of 169 real-time data sets were collected and used for predicting the CO_2 level, containing temperature, wind speed, air quality index, relative humidity, occupancy, and area per person as input parameters. To obtain the accurate result, all the data was scaled and normalized between 0 and 1, and 70% of the data is used for training and 30% for testing with 5-fold cross-validation process to validate the results. It has been found that the optimized GPR is quite accurate having R, RMSE, MAE, NS, and a20-index values of 0.98874, 4.20068 ppm, 3.35098 ppm, 0.9817, and 1, respectively.

This proposed prediction model is only valid for similar input data having similar statistical properties. The proposed study can help the researchers and professionals to predict the CO_2 concentration inside the office building and its effect on individual health. In future work, efficient machine learning models with large data sets can be used to predict the concentrations of various parameters like $PM_{2.5}$, PM_{10}, NO_x, SO_x, and CO_2.

Notations

A:	Area per person
AI:	Artificial intelligence
AIRMEX:	Air Monitoring and Exposure Assessment
ANN:	Artificial neural network
AQI:	Air quality index
A_S:	Original values
BASE:	Building Assessment Survey and Evaluation
BRI:	Building-related illness
CSIR:	Council of Scientific & Industrial Research
DT:	Decision tree
EL:	Ensemble learning
FL:	Fuzzy logic
GA:	Genetic algorithms
GPR:	Gaussian process regression
HOPE:	Health Optimization Protocol for Energy-Efficient Buildings
IAP:	Indoor air pollutants
IAQ:	Indoor air quality
IEQ:	Indoor environment quality
LR:	Linear regression
MAE:	Mean absolute error
MAPE:	Mean absolute percentage error
ML:	Machine learning
MSE:	Mean square error
NS:	Nash-Sutcliffe
O:	Number of occupants
PM:	Particulate matter
P_s:	Predicted values
R:	Correlation coefficient
RH:	Relative humidity
RMSE:	Root mean square error
SBS:	Sick building syndrome
SVM:	Support vector machine
T:	Number of samples
T_o:	Outdoor temperature
UFP:	Ultrafine particles
VOC:	Volatile organic compounds
WHO:	World Health Organization
W_S:	Wind speed
y_I:	Measured value.

Data Availability

The data cannot be shared at this time due to being part of future studies.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1063319).

References

[1] R. Kosonen and F. Tan, "The effect of perceived indoor air quality on productivity loss," *Energy and Buildings*, vol. 36, no. 10, pp. 981–986, 2004.

[2] N. R. Kapoor, A. Kumar, T. Alam, A. Kumar, K. S. Kulkarni, and P. Blecich, “A review on indoor environment quality of Indian school classrooms,” *Sustainability*, vol. 13, no. 21, article 11855, 2021.

[3] S. Mentese, N. A. Mirici, T. Elbir et al., “A long-term multiparametric monitoring study: indoor air quality (IAQ) and the sources of the pollutants, prevalence of sick building syndrome (SBS) symptoms, and respiratory health indicators,” *Atmospheric Pollution Research*, vol. 11, no. 12, pp. 2270–2281, 2020.

[4] N. R. Kapoor, A. Kumar, C. S. Meena et al., “A systematic review on indoor environmental quality in naturally ventilated school classrooms: a way forward,” *Advances in Civil Engineering*, vol. 2021, Article ID 8851685, 19 pages, 2021.

[5] N. Agarwal, C. S. Meena, B. P. Raj et al., “Indoor air quality improvement in COVID-19 pandemic: review,” *Sustainable Cities and Society*, vol. 70, p. 102942, 2021.

[6] J. C. M. Pires, S. I. V. Sousa, M. C. Pereira, M. C. M. Alvim-Ferraz, and F. G. Martins, “Management of air quality monitoring using principal component and cluster analysis–part I: SO_2 and PM_{10},” *Atmospheric Environment*, vol. 42, no. 6, pp. 1249–1260, 2008.

[7] N. Raj, A. Kumar, A. Kumar, and S. Goyal, “Indoor environmental quality: impact on productivity, comfort, and health of Indian occupants,” in *Proceedings of the Abstract Proceedings of International Conference on Building Energy Demand Reduction in Global South (BUILDER’19)*, pp. 1–9, New Delhi, India, December 2019, https://nzeb.in/event/builder19/.

[8] S. I. V. Sousa, F. G. Martins, M. C. Pereira et al., “Influence of atmospheric ozone, PM_{10} and meteorological factors on the concentration of airborne pollen and fungal spores,” *Atmospheric Environment*, vol. 42, no. 32, pp. 7452–7464, 2008.

[9] K. E. Paleologos, M. Y. E. Selim, and A.-M. O. Mohamed, “Chapter 8 - Indoor air quality: pollutants, health effects, and regulations,” in *Pollution Assessment for Sustainable Practices in Applied Sciences and Engineering*, A.-M. O. Mohamed, E. K. Paleologos, and F. M. Howari, Eds., pp. 405–489, Butterworth-Heinemann, 2021.

[10] L. Mølhave, “Organic compounds as indicators of air pollution,” *Indoor Air*, vol. 13, no. 6, pp. 12–19, 2003.

[11] C. Mandin, M. Trantallidi, A. Cattaneo et al., “Assessment of indoor air quality in office buildings across Europe - the OFFICAIR study,” *Science of the Total Environment*, vol. 579, pp. 169–178, 2017.

[12] P. F. Linden, “The fluid mechanics of natural ventilation,” *Annual Review of Fluid Mechanics*, vol. 31, no. 1, pp. 201–238, 1999.

[13] N. R. Kapoor and J. P. Tegar, “Human comfort indicators pertaining to indoor environmental quality parameters of residential buildings in Bhopal,” *International Research Journal of Engineering and Technology*, vol. 5, 2018.

[14] S. Kato, S. Murakami, A. Mochida, S.-i. Akabayashi, and Y. Tominaga, “Velocity-pressure field of cross ventilation with open windows analyzed by wind tunnel and numerical simulation,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 44, no. 1-3, pp. 2575–2586, 1992.

[15] C.-G. Bornehag, J. Sundell, L. Hägerhed-Engman, and T. Sigsgaard, “Association between ventilation rates in 390 Swedish homes and allergic symptoms in children,” *Indoor Air*, vol. 15, no. 4, pp. 275–280, 2005.

[16] P. S. Burge, “Sick building syndrome,” *Occupational and Environmental Medicine*, vol. 61, no. 2, pp. 185–190, 2004.

[17] A. Alwan, *Global status report on noncommunicable diseases 2010*, World Health Organization, 2011.

[18] World Health Organization, *Noncommunicable diseases country profiles 2018*, World Health Organization, 2018.

[19] Z. Tong, Y. Chen, A. Malkawi, G. Adamkiewicz, and J. D. Spengler, “Quantifying the impact of traffic-related air pollution on the indoor air quality of a naturally ventilated building,” *Environment International*, vol. 89-90, pp. 138–146, 2016.

[20] J. Vanus, R. Martinek, P. Bilik, J. Zídek, P. Dohnalek, and P. Gajdos, “New method for accurate prediction of CO_2 in the Smart Home,” in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–5, Taipei, Taiwan, 2016.

[21] A. Pantazaras, S. E. Lee, M. Santamouris, and J. Yang, “Predicting the CO_2 levels in buildings using deterministic and identified models,” *Energy and Buildings*, vol. 127, pp. 774–785, 2016.

[22] J. Kallio, J. Tervonen, P. Räsänen, R. Mäkynen, J. Koivusaari, and J. Peltola, “Forecasting office indoor CO_2 concentration using machine learning with a one-year dataset,” *Building and Environment*, vol. 187, p. 107409, 2021.

[23] B. Khazaei, A. Shiehbeigi, and A. R. H. M. A. Kani, “Modeling indoor air carbon dioxide concentration using artificial neural network,” *International journal of Environmental Science and Technology*, vol. 16, no. 2, pp. 729–736, 2019.

[24] J. P. Skön, M. Johansson, M. Raatikainen, K. Leiviskä, and M. Kolehmainen, “Modelling indoor air carbon dioxide (CO_2) concentration using neural network,” *Methods*, vol. 14, no. 15, p. 16, 2012, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1047.2472&rep=rep1&type=pdf.

[25] S. Taheri and A. Razban, “Learning-based CO_2 concentration prediction: application to indoor air quality control using demand-controlled ventilation,” *Building and Environment*, vol. 205, p. 108164, 2021.

[26] A. Mohammadshirazi, V. A. Kalkhorani, J. Humes et al., “Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: a case study,” *Building and Environment*, vol. 213, p. 108833, 2022.

[27] M. Lillstrang, M. Harju, G. del Campo, G. Calderon, J. Röning, and S. Tamminen, “Implications of properties and quality of indoor sensor data for building machine learning applications: two case studies in smart campuses,” *Building and Environment*, vol. 207, p. 108529, 2022.

[28] A. Kumar, H. C. Arora, N. R. Kapoor et al., “Compressive strength prediction of lightweight concrete: machine learning models,” *Sustainability*, vol. 14, no. 4, p. 2404, 2022.

[29] A. Kumar, H. C. Arora, M. A. Mohammed, K. Kumar, and J. Nedoma, “An optimized neuro-bee algorithm approach to predict the FRP-concrete bond strength of RC beams,” *IEEE Access*, vol. 10, pp. 3790–3806, 2022.

[30] S. Lek and Y. S. Park, “Artificial neural networks,” in *Encyclopedia of Ecology, Five-Volume Set*, pp. 237–245, Elsevier Inc, 2008.

[31] K. Kumar and R. P. Saini, “Development of correlation to predict the efficiency of a hydro machine under different operating conditions,” *Sustainable Energy Technologies and Assessments*, vol. 50, p. 101859, 2022.

[32] M. D. Wilson, “Support vector machines,” in *Encyclopedia of Ecology*, pp. 3431–3437, Academic Press, 2008.

[33] S. D. Brown and A. J. Myles, *Decision Tree Modeling in Classification*, Elsevier, 2009.

[34] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning (Vol. 2)*, MIT press Cambridge, MA, 2006.

[35] S. Ganesh, *Multivariate Linear Regression*, Elsevier Ltd, 2010.

[36] A. Kumar, H. C. Arora, K. Kumar et al., “Prediction of FRCM-concrete bond strength with machine learning approach,” *Sustainability*, vol. 14, no. 2, p. 845, 2022.