Beyond One Room: Comprehensive Predictive Analysis of CO2 in Indoor Air Quality

Ariel Isaac Posada Barrera
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
arielisaac.posada@upaep.edu.mx

Laura Margarita Rodríguez Peralta
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
lauramargarita.rodriguez01@upaep.mx

Éldman de Oliveira Nunes
Centro Universitário SENAI/CIMATEC
Bahía, Brasil
eldman.nunes@gmail.com

Paulo Nazareno Maia Sampaio
Universidad de Lima (ULima)
Lima, Peru
pmaia@ulima.edu.pe

Fabian Leonardo Cuesta Astudillo
Universidad Politécnica Salesiana,
Cuenca, Ecuador,
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
fabian.cuesta@upaep.edu.mx

***Abstract*— Indoor air quality is important for public health. This study was designed to develop predictive models that concentrate on indoor air quality, focusing on the CO2 concentrations. Implementing and training the Machine Learning Models—Regression Forest Model and Gradient-Boosted Tree Model—on a dataset of measurements in Mexico and other sources having pollutant levels, temperature, relative humidity, people density, and ventilation characteristics. The models used had many scenarios of relative humidity, temperature and pollutant levels, demonstrating the relation between the characteristics of the space, human activity and indoor CO2 concentration. The result was a model with acceptable accuracy, predicting CO2 levels in indoor spaces.**

Keywords— IoT, Big Data, Machine learning, Monitoring, Sick buildings syndrome.

I. Introduction

This investigation, set against the backdrop of Mexico, delves into the crucial task of predicting indoor air quality, focusing particularly on CO2 concentrations. Previous studies, such as [1], have underscored the vital importance of comprehending and ensuring optimal indoor air quality. They emphasized the need for vigilant monitoring and effective management of indoor pollutants to safeguard public health. Additionally, their research incorporated a range of EPA-endorsed variables, underscoring their relevance in machine learning analyses, especially in the context of respiratory diseases and Sick Building Syndrome (SBS).

As urban centers in Mexico continue to expand and indoor activities become integral to our daily lives, the significance of indoor air quality becomes increasingly evident. The manifestation of adverse indoor conditions, as detailed in [2] concerning Sick Building Syndrome (SBS), is further supported by [3], which emphasizes the essential role of thermal comfort within buildings.

Previous studies, such as [4], conducted since the last century, have demonstrated that indoor spaces contain higher concentrations of pollutants compared to outdoor environments. Certain indoor air quality assurance policies, like those detailed in [5], imply a specific density per cubic meter, setting an ideal social distance of 1.83 meters between individuals. Moreover, specific room height standards, as outlined in the Mexican global policy [6], establish room heights of 2.3-2.5 meters in the state of Puebla and 2.7 meters in Morelos. Studies [7] and [8] aim to implement proactive strategies to ensure indoor air quality, maintaining low pollutant levels while considering routine household and office activities, along with the behavior of environmental, ventilation, and human activity variables. These studies utilize the knowledge generated on the correlation between these variables and pollutant levels to enhance the effectiveness of these strategies.

This study is a continuation of previous work in [7], and [8] that involved the use of data collected from indoor locations in the Mexican States Puebla and Morelos.

In the study [7], working with the variables previously defined in [1], an analysis was conducted with a sample of

979-8-3503-7457-5/24/$31.00 ©2024 IEEE

values to observe if there was a correlation between environmental and human activity values, and the concentrations of pollutants in indoor environments. It was found that there is a variation in the correlations between the values when in a combination of relative humidity and recommended temperature (temperature between 18°C and 28°C, and relative humidity between 30% and 50%), providing a basis for forecasting pollutant levels.

Time later in study [8], the forecasting of the values was carried out using Big Data Spark technology, and the machine learning models of Random Forests and Gradient Boosted Tree Models. The forecast was made on a larger sample for the four pollutants: CO2, TVOC, PM2.5, and PM10, yielding high forecast results for PM2.5 and PM10, acceptable for TVOC, and with CO2 as one of the future tasks is to improve the forecasts having R-squared values of 0.530 for the Random Forest Model, and 0.717 for the Gradient-Boosted Tree Model.

To provide a comprehensive overview, several studies previously conducted on indoor air quality forecasting were summarized in Table I below. This table showcases the various techniques and variables employed in the field, with CO2 frequently emerging as a significant variable, underscoring its importance in air quality prediction.

TABLE I. RECENT STUDIES UTILIZING HISTORICAL DATA AS INPUTS, AND PREDICTING FUTURE AIR QUALITY INDEX

Study	***Approach***	***Predicted variables***
[7], [8]	Analysis of correlation with human activity, space, and environmental data	CO2, TVOC, PM 2.5, PM10
[9], [10], [11]	Prediction of indoor air quality in office environments using neural network models	CO2 concentrations
[12]	Multilayer Perceptron (MLP) predictions, with temperature, relative humidity, and dewpoint	CO2 concentrations
[13]	Ridge, decision trees, random forest, and multilayer perceptrons with temperature, relative humidity, air pressure and passive infrared sensors data	CO2 concentrations
[14]	Estimation and forecasting of indoor air quality	Ambient temperature and humidity
[15]	Real-time prediction of indoor air pollutants using indoor environmental data	Air quality prediction
[16]	Deep learning with time series	Temperature and CO2
[17]	Prediction of a spectrum of indoor air quality pollutants	CO and CO2
[18]	Prediction of indoor bioaerosol concentrations using advanced artificial intelligence methods	Air quality prediction
[19]	Indoor air quality prediction using a deep learning framework	Air quality prediction
[20]	Examination of the significance of environmental parameters in indoor air quality	Relative humidity and temperature
[21]	Exploration of various variables affecting indoor air quality	PM2.5, PM10, and CO
[22], [23], [24], [25]	Examination of the importance of various variables in indoor air quality	CO2, PM2.5, and TVOC

This paper advances the state of the art by developing models specifically designed for CO2 level forecasting in a single, extensively documented location, using a significantly larger dataset that incorporates new data types. This refined approach leverages only present-time data to predict CO2 levels, employing advanced Random Forest and Gradient Boosted Tree models based on the architecture from study [8]. Our aim is to enhance forecast precision utilizing a dataset enriched with temperature and humidity variations.

The objectives of the current study are as follows:

I. Expanding our understanding of indoor air quality through an extensive dataset that captures a wide range of environmental variables, human behaviors, and pollutants. This approach is intended to refine model training for better accuracy.

II. Developing a location-specific model that reduces prediction error and increases precision, laying the groundwork for the medium-term implementation of a real-time recommendation system for use in homes and offices. This system will ensure the maintenance of air quality based on historical data specific to each location.

This paper is organized in the following sections: In Section II, "Collected Data," we explore the methodologies used for data gathering and outline the variables under consideration. Section III, "Machine Learning," delves into data analysis, detailing the use of Random Forests and Gradient Boosted Tree models. Section IV, "Lessons Learned," summarizes the machine learning model results, and reflects on the observed relationships between human activities, environmental conditions, and CO2 levels. Section V, "Conclusions," summarizes the conclusions of the study in the prediction of CO2 with Machine Learning Models. Finally, Section VI, "Future Works," looks ahead to potential areas of further research.

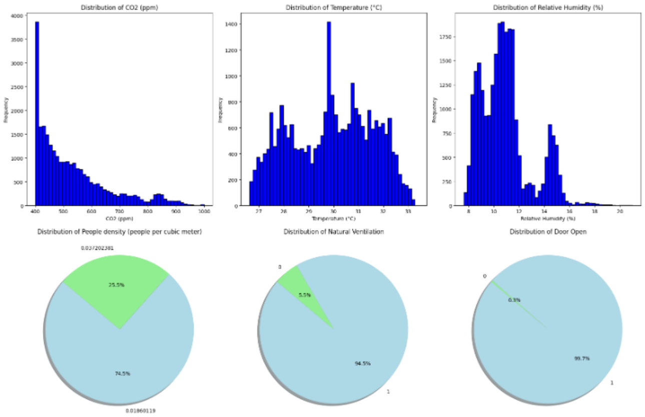
II. Collected Data

Data collection occurred in a 3.2x5x2.8 meter room in Morelos, Mexico, which a person used as a home office. Over 43 hours on working days, the study compiled a dataset of 25,888 continuous values, captured every 10 seconds using two primary sensor systems: the Databot (referenced in [26]) and the air quality monitor (referenced in [27]). To ensure data integrity, outputs from both sensors were compared. This room, conditioned for the experiment, was monitored for temperature, relative humidity, and people density, alongside detailed logs of door and window operations, reflecting open or closed states. Additionally, the measurements included detailed records of the entries and exits of individuals from the room, providing a comprehensive overview of the occupancy patterns and their impact on air quality. For the environmental variables, no outliers were removed, and no data cleaning was performed on the input variables to ensure that all values recorded during the experiment were considered.

The model considers only natural ventilation. Key insights from [28] highlight the environmental benefits of natural ventilation compared to mechanical ventilation, particularly in urban residential settings. The study demonstrates that natural ventilation is more effective in ensuring air quality and reducing energy consumption. Additionally, the study [29] provides a foundation for designing effective natural ventilation strategies, focusing on the precise characterization of ventilation components.

Figure 1 describes the distribution of the variables.

Fig. 1. Distributions of the obtained dataset



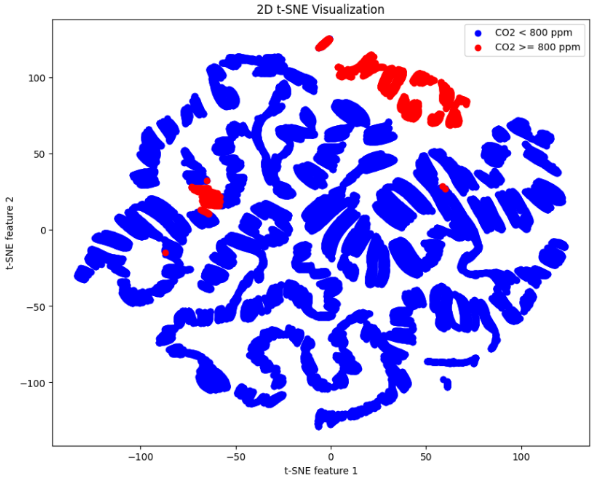
The data collected was utilized in machine learning algorithms to predict indoor pollutant levels.

III. Machine Learning

Before delving into data analysis, it was essential to determine whether there was a visually discernible influence concerning the relationship between CO2 and the input variables. The first step in our analysis involved creating a graphical representation of the dataset using t-SNE. As described in [30], t-SNE is adept at reducing the dimensionality of systems populated with multiple variables, providing a reduced-dimension visualization where certain intricate patterns or relationships might be latent.

Figure 2 shows the t-SNE visualization of the data, representing variables such as Temperature, Relative Humidity, People Density, Natural Ventilation, and Door Open status. In this figure, the CO2 levels, as the dependent variable, segregate the environment into two categories: high risk (with CO2 levels equal to or exceeding 800 ppm) and safe (with CO2 levels falling below 800 ppm). This segregation is crucial as the primary objective is to prevent the increase of CO2 levels beyond 800 ppm to ensure the health and well-being of individuals in indoor environments. The analysis further explores, based on the distribution, the feasibility of implementing complex Machine Learning models to achieve real-time diagnostics.

Fig. 2. Distributions of the obtained dataset



The t-SNE results revealed an absence of direct, observable relationships between the CO2 groupings and the dimensions of the input variables, indicating a challenge for straightforward classification algorithms due to the lack of clear demarcation among groups.

This complexity is further underscored by the Levene test (referenced in [31]), which assesses the homogeneity of variances across groups. The necessity of this test stems from its critical role in confirming the assumption of equal variances, a foundational requirement for the application of numerous statistical methods.

In our context, the Levene test was applied to two distinct groups: high risk (characterized by CO2 levels equal to or exceeding 800 ppm) and safe (where CO2 levels are below 800 ppm), with findings presented in Table II.

TABLE II. LEVENE TEST RESULTS

Variable	*Levene Statistic*	*P-value*
Temperature	328.772412	5.040745e-73
Relative Humidity	993.578114	4.968721e-214
People Density	126.279130	3.122203e-29
Natural Ventilation	3,668.279653	0.0
Door Open	979.757688	3.890805

Upon examining the Levene test results, it's evident that variables such as Temperature, People Density, and Door Open have P-values near zero, signaling significant variance differences between the two CO2 groups (high risk and safe). Likewise, Relative Humidity and Natural Ventilation show extremely small P-values, underscoring notable variance disparities.

The distinctions in variances across these variables imply that there's a substantial differentiation between the high risk and safe CO2 groups in terms of these environmental parameters.

This difference in variances suggests that there exists an inherent structure or pattern within the data that could be leveraged by a machine learning model to predict the CO2 levels.

After the preliminary analysis, the machine learning models were trained in the Big Data technology Apache Spark, as described in [32], being able to process big sets of data, applying distributed computing.

The machine learning models trained were Linear Regression, Random Forests Model and Gradient-Boosted Tree Model. These models were trained using the average CO2 level resulting from the measurement data with the specified characteristics as the target variable. The dataset was split into 60% for training and 40% for validation, with each model being trained using the same subsets.

Linear Regression analysis ([33]) was employed as a preliminary analytical method to examine the relationships between variables before considering more complex modeling techniques, based on the assumptions made in previous analyses that correlations exist.

Random Forests utilize a collection of decision trees, each developed using distinct data subsets. The objective behind this strategy is to bolster both the precision and resilience of the model by capitalizing on the variability among the trees, as outlined in [34].

On the other hand, Gradient-Boosted Tree Models function with a distinct mechanism. They progressively assemble a series of weak learners, generally decision trees, where each successive tree aims to rectify the inaccuracies of the one before it. This step-by-step refinement is pivotal in elevating the model's predictive capabilities, as explained in [35].

The core differences between Random Forests and Gradient-Boosted Tree Models lie in their respective methodologies and goals. While Random Forests aim to craft a varied ensemble of trees using methods like bootstrapping and feature selection, Gradient-Boosted Trees focus on sequential enhancement by addressing residual errors. Such differences in approach culminate in distinct advantages and potential uses for each model when predicting indoor air quality from the previously discussed dataset.

The performance of the models was assessed using several metrics. Mean Absolute Error ([36]) calculates the average of the absolute variances between forecasted and actual values, offering a direct and more resilient measure against outliers. Mean Squared Error ([36]) takes the average of the squared discrepancies between predicted and real values, thus accentuating bigger errors and exhibiting higher sensitivity to outliers. The RMSE (Root Mean Square Error) is the square root of MSE, mirroring the properties of emphasizing larger discrepancies and outlier sensitivity.

Furthermore, to gauge the efficacy of the machine learning models, the Coefficient of Determination, also called R-squared ([37]) was utilized. R-squared serves as a statistical metric denoting the percentage of variance in the dependent variable accountable by the independent variable(s) within a regression context. Its selection was influenced by its prowess in offering supplementary insights, particularly regarding how well the model conforms to the dataset. Typically, a greater R-squared value signals a superior fit, which in turn suggests a more dependable model.

Table III presents the performance metrics of the machine learning models. From the data displayed, it is evident that the model effectively predicts a significant portion of the outcomes.

TABLE III. MODEL METRICS FOR CO2 PREDICTION WITH ENVIRONMENTAL VARIABLES.

Metric	*Linear Regression Model*	*Random Forest Regression Model*	*Gradient-Boosted Tree Regression Model*
RMSE (ppm)	92.252	76.575	50.722
MSE (ppm^2)	8,510.361	5,848.480	2,572.696
MAE (ppm)	72.273	58.299	36.053
R-squared	0.471	0.637	0.840

In evaluating the performance of our machine learning models, both the Regression Forest Model and the Gradient-Boosted Tree Model demonstrate strong predictive capabilities, achieving better scores than the traditional Linear Regression approach. Their results highlight not just the precision of the models but also their reliability in forecasting CO2 concentrations.

The relatively low error margins of these models emphasize their robustness in various scenarios. Specifically, the Gradient-Boosted Tree Model, boasting an R-squared value of 0.840, showcases its aptitude as a highly effective tool. This high degree of fit suggests that it can be confidently deployed across a wider range of environments. Furthermore, this performance also underscores the importance and impact of the model's selected variables in determining CO2 levels. The interplay of these variables, as reflected in the model's outcomes, confirms their critical role in influencing indoor air quality.

IV. LESSONS LEARNED

The comprehensive analysis of this study has yielded profound insights into the complex dynamics between environmental factors and their impact on CO2 levels in indoor spaces. It has been observed that variables such as temperature, relative humidity, people density, and ventilation measures, along with the status of doors, play a significant role in influencing CO2 concentrations. The intricate nature of these relationships is highlighted by the fact that they do not lend themselves to simple classification algorithms for prediction. This complexity is further evidenced by the Levene test results, which indicate significant variances in these variables, suggesting a patterned structure within the data.

In response to these complexities, the deployment of advanced machine learning models like Random Forests and Gradient-Boosted Trees, particularly when trained on big data platforms such as Apache Spark, has been instrumental. These models have demonstrated a remarkable ability to discern and harness the nuanced patterns present in the data for more accurate CO2 level predictions. The robustness of these models is reflected in performance metrics such as RMSE, MSE, MAE, and R-squared, with the Gradient-Boosted Tree Model showing exceptional predictive strength, as evidenced by its high R-squared value. This level of predictive accuracy indicates that when these models are appropriately utilized, they can serve as powerful tools for forecasting CO2 levels across diverse indoor environments. The key to improving the precision of CO2 concentration forecasts—and consequently enhancing indoor air quality management—lies in understanding and leveraging the intricate relationships between these environmental variables.

V. CONCLUSSIONS

In conclusion, our study confirms the Gradient-Boosted Tree Model as a highly effective method for predicting CO2 levels in indoor environments. The model's precision, as indicated by its superior performance metrics, has practical implications for the proactive management of indoor air quality. Lower RMSE and MAE values in the Gradient-Boosted Tree Model compared to the Regression Forest Model demonstrate its greater accuracy in predicting actual CO2 levels, while the higher R-squared value suggests a better fit for the data.

This accuracy is critical for policymakers and building managers who rely on precise data to implement air quality regulations and improvements.

By applying such advanced machine learning techniques, we can inform decisions that lead to healthier indoor spaces. This is especially pertinent in urban areas where managing air quality is increasingly challenging. The metrics from our model provide a clear indication of its reliability and underscore the potential for machine learning to play a vital role in public health initiatives and the creation of safer living and working spaces.

VI. FUTURE WORKS

Building on the findings of this study, our forthcoming endeavors will concentrate on the development of a real-time system to pinpoint the risk of respiratory disease transmission in indoor environments in Mexico. This system will leverage the capabilities of Big Data and Internet of Things (IoT) technologies, incorporating a Kappa architecture that seamlessly combines the processing of streaming data in real-time with batch processing. This will enable continuous monitoring and in-depth analysis of indoor air quality parameters.

Furthermore, we aim to delve deeper into the realm of advanced analytics by implementing machine learning models that utilize image analysis through deep neural networks and Generative Adversarial Networks (GANs) as described in [38]. Additionally, experiments will also be conducted using Kolmogorov Arnold Networks, as described in [39].

Through these research initiatives, we are committed to improving indoor air quality and ensuring safer environments in Mexico.

DATA AVAILABILITY

The dataset is available upon request.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

ETHICAL CONSIDERATIONS

This study was conducted with the utmost respect for ethical standards. Measurements were taken in workspaces and living rooms under normal conditions during everyday activities, ensuring that subjects were not exposed to any risks. No experiments were conducted on animals or humans. Furthermore, the research activities did not produce any environmental pollution or harm. All procedures were designed to minimize any potential negative impacts and to prioritize the well-being and safety of all involved.

REFERENCES

[1] Domínguez Portillo, J., Rodríguez Peralta, L. M., Sampaio, P. N. M., & Nunes, E. d. O. (2023). Determining environmental indicators related to the propagation of contagious diseases and health issues: A Systematic Literature Review. In 2023 18th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). https://doi.org/10.23919/CISTI58278.2023.10212010

[2] Joshi, S. M. (Aug de 2008). The sick building syndrome. Indian J Occup Environ Med. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796751

[3] Kunkel S. Kontonasiou E. (2015). Indoor air quality, thermal comfort and daylight policies on the way to nZEB – status of selected MS and future policy recommendations. BPIE – Buildings Performance Institute Europe.

[4] Jones, A. P. (1999). Indoor air quality and health. In Atmospheric Environment (Vol. 33, Issue 28, pp. 4535–4564). Elsevier BV. https://doi.org/10.1016/s1352-2310(99)00272-1

[5] Feng Y., Marchal T., Sperry T., Yi H. (2020) Influence of wind and relative humidity on the social distancing effectiveness to prevent COVID-19 airborne transmission: a numerical study. J Aerosol Sci. 2020:105585

[6] Wolpert Kuri, J., Kotecki Golasinska, T. D., & Morales Ramírez, A. (2017). Código de Edificación de Vivienda (3ª edición). Secretaría de Desarrollo Agrario, Territorial y Urbano. https://www.gob.mx/inafed/documentos/codigo-de-edificacion-de-vivienda-3era-edicion

[7] Posada Barrera, A.I., Rodríguez Peralta, L.M., de Oliveira Nunes, É., Sampaio, P.N.M. (2024). Influence of Indoor Conditions on Sick Building Syndrome: A Data-Driven Investigation. In: Rocha, Á., Ferrás, C., Hochstetter Diez, J., Diéguez Rebolledo, M. (eds) Information Technology and Systems. ICITS 2024. Lecture Notes in Networks and Systems, vol 932. Springer, Cham. https://doi.org/10.1007/978-3-031-54235-0_5

[8] Posada Barrera, A. I., Rodríguez Peralta, L. M., de Oliveira Nunes, É., Maia Sampaio, P. N., & Cuesta Astudillo, F. L. (2023). Improving Public Health Policies with Indoor Air Quality Predictive Models. In 2023 International Conference on Computational Science and Computational Intelligence (CSCI). 2023 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE. https://doi.org/10.1109/csci62032.2023.00040

[9] Putra, J. C. P., Safrilah, & Ihsan, M. (2018). The prediction of indoor air quality in office room using artificial neural network. In AIP Conference Proceedings. https://doi.org/10.1063/1.5042896

[10] Skön, J., Johansson, M., Raatikainen, M., Leiviskä, K., & Kolehmainen, M. (2012). Modelling indoor air carbon dioxide (CO2) concentration using neural network. World Academy of Science, Engineering and Technology, International Science Index, 6, 737–741.

[11] Khazaei, B., Shiehbeigi, A., & Haji Molla Ali Kani, A. R. (2019). Modeling indoor air carbon dioxide concentration using artificial neural network. International Journal of Environmental Science and Technology, 16, 729–736

[12] Taheri, S., & Razban, A. (2021). Learning-based CO2 concentration prediction: Application to indoor air quality control using demand-controlled ventilation. In Building and Environment (Vol. 205, p. 108164). Elsevier BV. https://doi.org/10.1016/j.buildenv.2021.108164

[13] Kallio, J., Tervonen, J., Räsänen, P., Mäkynen, R., Koivusaari, J., & Peltola, J. (2021). Forecasting office indoor CO2 concentration using machine learning with a one-year dataset. In Building and Environment (Vol. 187, p. 107409). Elsevier BV. https://doi.org/10.1016/j.buildenv.2020.107409

[14] Sharma, P. K., Mondal, A., Jaiswal, S., Saha, M., Nandi, S., De, T., & Saha, S. (2021). IndoAirSense: A framework for indoor air quality estimation and forecasting. Atmospheric Pollution Research, 12(1), 10–22. https://doi.org/10.1016/j.apr.2020.07.027

[15] Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., & Mandin, C. (2019). Machine learning and statistical models for predicting indoor air quality. Indoor Air, 29(5), 704–726. https://doi.org/10.1111/ina.12580

[16] Chen, E. X., Han, X., Malkawi, A., Zhang, R., & Li, N. (2023). Adaptive model predictive control with ensembled multi-time scale deep-learning models for smart control of natural ventilation. In Building and Environment (Vol. 242, p. 110519). Elsevier BV. https://doi.org/10.1016/j.buildenv.2023.110519

[17] Baqer, N.S., Albahri, A.S., Mohammed, H.A. et al. (2022). Indoor air quality pollutants predicting approach using unified labelling process-based multi-criteria decision making and machine learning techniques. Telecommun Syst, 81, 591–613. https://doi.org/10.1007/s11235-022-00959-2

[18] Lee, J. Y. Y., Miao, Y., Chau, R. L. T., Hernandez, M., & Lee, P. K. H. (2023). Artificial intelligence-based prediction of indoor bioaerosol concentrations from indoor air quality sensor data. Environment International, 174(107900), 107900. https://doi.org/10.1016/j.envint.2023.107900

[19] Wu, Q., Geng, Y., Wang, X., Wang, D., Yoo, C., & Liu, H. (2023). A novel deep learning framework with variational auto-encoder for indoor air quality prediction. Frontiers of Environmental Science & Engineering, 18(1). https://doi.org/10.1007/s11783-024-1768-7

[20] Hoang, M. L., Carratù, M., Paciello, V., & Pietrosanto, A. (2021). Body Temperature-Indoor Condition Monitor and Activity Recognition by MEMS Accelerometer Based on IoT-Alert System for People in Quarantine Due to COVID-19. Sensors (Basel, Switzerland), 21(7), 2313. https://doi.org/10.3390/s21072313

[21] M. S. H. Sassi and L. C. Fourati, "Deep Learning and Augmented Reality for IoT-based Air Quality Monitoring and Prediction System," 2021 International Symposium on Networks, Computers and Communications (ISNCC), Dubai, United Arab Emirates, 2021, pp. 1-6, doi: 10.1109/ISNCC52172.2021.9615639

[22] L. J. M and S. R. S., "Role of Nano-Sensors towards CO2 Concentrations in an Indoor Classroom Environment to improve Occupational Health," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022, pp. 1-7, https://doi.org/10.1109/ICDCECE53908.2022.9792733

[23] Ho, Y.-H., Li, P.-E., Chen, L.-J., & Liu, Y.-L. (2020). Indoor air quality monitoring system for proactive control of respiratory infectious diseases. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems. SenSys '20: The 18th ACM Conference on Embedded Networked Sensor Systems. ACM. https://doi.org/10.1145/3384419.3430456

[24] Spandonidis, C., Tsantilas, S., Giannopoulos, F., Giordamlis, C., Zyrichidou, I., & Syropoulou, P. (2020). Design and Development of a New Cost-Effective Internet of Things Sensor Platform for Air Quality Measurements. In Journal of Engineering Science and Technology Review (Vol. 13, Issue 6, pp. 81–91). International Hellenic University. https://doi.org/10.25103/jestr.136.12

[25] Moursi, A. S., El-Fishawy, N., Djahel, S., & Shouman, M. A. (2021). An IoT enabled system for enhanced air quality monitoring and prediction on the edge. Complex & intelligent systems, 7(6), 2923–2947. https://doi.org/10.1007/s40747

[26] Databot. (2023). Home - databot™ - Real Data, Real Science, Real Fun! Retrieved from https://databot.us.com/

[27] Inkbird. (2023). Wi-Fi 8-in-1 Air Quality Monitor IAQM-128W. Retrieved from https://inkbird.com/products/wi-fi-air-monitor-iaqm-128w

[28] Hu, Y., Wu, Y., Wang, Q., Hang, J., Li, Q., Liang, J., Ling, H., & Zhang, X. (2021). Impact of Indoor-Outdoor Temperature Difference on Building Ventilation and Pollutant Dispersion within Urban Communities. Atmosphere. https://doi.org/10.3390/atmos13010028

[29] Jones, B., Cook, M., Fitzgerald, S., & Iddon, C. (2016). A review of ventilation opening area terminology. Energy and Buildings, 118, 249-258. https://doi.org/10.1016/J.ENBUILD.2016.02.053

[30] van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9(nov), 2579-2605.

[31] Nordstokke, David & Zumbo, Bruno. (2010). A New Nonparametric Levene Test for Equal Variances. Psicológica. 31. 401-430.

[32] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65.

[33] Yan, X. (2009). Linear Regression analysis: Theory and Computing. World Scientific.

[34] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[35] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[36] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean-absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79-82.

[37] German, G. (2021). Coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in evaluating regression analysis. PeerJ Computer Science, 7(e623), e623. https://doi.org/10.7717/peerj-cs.623

[38] Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. In International Journal of Information Management Data Insights (Vol. 1, Issue 1, p. 100004). Elsevier BV. https://doi.org/10.1016/j.jjimei.2020.100004

[39] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov-Arnold Networks (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2404.19756