# Detecting CO2 anomalies using machine learning: case study of a library

Ayoub Hannad, Alain Godon, Franck Mercier, Charline Dematteo, Hassan Chehouani, Marie-Lise Pannier

# Detecting $CO_2$ anomalies using machine learning: case study of a library

**Ayoub Hannad*[1,2], Alain Godon[3], Franck Mercier[3], Charline Dematteo[3,4], Hassan Chehouani[5], Marie-Lise Pannier[2]**

**[1] Mohammed VI Polytechnic University,Ben Guerir, 43150 660, Hay Moulay Rachid, Maroc**
**[2] Univ Angers, LARIS, SFR MATHSTIC, F-49000 Angers, France 62 avenue Notre Dame du Lac, 49000 Angers**
**[3] Univ Angers, Polytech Angers, F-49000 Angers, France 62 avenue Notre Dame du Lac, 49000 Angers**
**[4] Indiggo, 44000 Nantes, France 4 avenue Millet, 44000 Nantes**
**[5] Laboratory of Processes for Sustainable Energy & Environment (ProcEDE), Cadi Ayyad University, Marrakech, Morocco.**
*** ayoub.hannad@um6p.ma**

*RESUME. La qualité de l'air intérieur est un élément très important pour un environnement sain et confortable. L'utilisation des capteurs à faible coût, enregistrant les taux de $CO_2$ ou d'autres substances, s'est répandue ces dernières années. Cependant, la quantité et la complexité des données mesurées posent des problèmes pour l'identification des anomalies et l'extraction d'informations significatives. C'est là que les techniques d'apprentissage automatique excellent. Dans le cadre du projet ACQA, un micro-capteur composé de différents capteurs de qualité de l'air a été développé et déployé dans une bibliothèque universitaire. Cette étude examine la détection d'anomalies dans les niveaux de $CO_2$ enregistrés, en utilisant différents modèles d'apprentissage automatique : K-Nearest Neighbors, Random Forest (RF), Gradient Boosting Regressor et Decision Tree Regressor. Les modèles sont évalués en fonction de leurs précision et de leur efficacité à détecter des anomalies. Les résultats sont quantifiés à l'aide des indicateurs $R^2$, RMSE et MAE, et montrent que le modèle RF est le plus précis.*

*MOTS-CLÉS : Qualité de l'air intérieur, Apprentissage automatique, Détection d'anomalies.*

*ABSTRACT. Indoor air quality is a very important element of a healthy and comfortable environment. The use of low-cost sensors recording $CO_2$ or other substances has become popular in recent years. However, the quantity and complexity of the data measured by these sensors present challenges for identifying anomalies and extracting meaningful information. This is where machine-learning techniques excel. As part of the ACQA project, a micro sensor composed of various air quality sensors was developed and deployed in a university library. This study examines the detection of anomalies in $CO_2$ levels recorded, using various machine-learning models: K-Nearest Neighbors, Random Forest (RF), Gradient Boosting Regressor and Decision Tree Regressor. The models are evaluated based on their accuracy and efficiency in detecting anomalies. The results, quantified using $R^2$, RMSE and MAE indicators, and show that the RF model is the most accurate.*

*KEYWORDS : Indoor air quality, Machine learning, Anomaly detection*

## 1. INTRODUCTION

Energy efficiency in buildings has become a major concern worldwide. The need to reduce energy consumption and greenhouse gas emissions has led to a constant search for solutions to make buildings energy efficient. In addition, as the world become increasingly interconnected, people's lifestyles have

shifted towards spending a significant portion of their time indoors. Studies indicate that people are spending more than 80% of their time indoors (Diffey 2011). This includes time spent in residential buildings, offices, schools, and other indoor environments, which is why the quest for energy efficiency must not be at the expense of Indoor Air Quality (IAQ).

IAQ has become the objective of several studies especially after the pandemic of Covid-19, most of the studies focus on the public spaces such as classrooms (Di Gilio et al. 2021), aircraft cabines (Gameiro et al. 2023), Subway train (Park et Ha 2008), etc. There are two types of studies, first ones use high accurate expensive sensors installed for a short time to investigate specific IAQ issues, and the second ones use low-costs sensors for continuous monitoring. The low costs-sensors offer a cost-effective solution for continuous monitoring of various IAQ parameters, including $CO_2$ levels, temperature, humidity, VOCs, and particulate matter. They provide real-time data collection and enable the generation of large datasets for analysis. This data can be used in various applications such as building management and energy efficiency. By continuously monitoring IAQ parameters, building managers can make informed decisions regarding ventilation, heating, and cooling systems to create a healthier and more energy-efficient environment (Li et al. 2020).

However, the volume and complexity of the data generated by low-cost sensors present challenges in identifying anomalies and extracting meaningful insights. This is where machine learning (ML) techniques excel. ML is a branch of artificial intelligence that focuses on the development of algorithms and statistical models, enabling computer systems to learn from data and make predictions or decisions without being explicitly programmed (Senders et al. 2018). In the context of IAQ anomaly detection, ML models can be trained to predict the expected values of various IAQ parameters based on historical data. These predicted values can then be compared with the actual sensor readings to identify any significant deviations or anomalies. By continuously monitoring and comparing the predicted and actual data, the ML system can automate the detection and alerting of anomalies in real-time, enhancing the efficiency and accuracy of IAQ monitoring systems. Many studies focuses on the prediction of $CO_2$ levels using different ML models. Taheri et Razban (2021) develop and compare six ML models for the prediction of indoor $CO_2$ concentration to control the HVAC systems in classrooms. Kallio et al. (2021) investigate four models for the prediction of future office $CO_2$ concentration. Chen et al. (2018) tests four ML models for the prediction of $CO_2$, TVOC and HCHO. Our paper, do not focus only on the comparison between models and access their performance, but also on applying ML techniques to detect $CO_2$ anomalies. By training ML models on historical $CO_2$ sensor data, we aim to estimate and detect deviations from normal $CO_2$ levels.

## 2. MATERIALS AND METHODOLOGY

### 2.1. INSTRUMENTATION AND STUDY SITE

The device ( Figure 2) used in this work was developed by the LARIS laboratory of Polytech Angers using low-cost micro-sensors. These sensors have been designed to measure a wide range of parameters, such as $CO_2$, temperature and humidity with a nondispersive infrared sensor, Carbone monoxide (CO), Ethanol ($C_2H_5OH$), and Nitrogen dioxide ($NO_2$) using microelectromechanical systems based metal oxide semiconductors gas sensors, and particulate matter (PM) using a laser scattering sensor, making them suitable for an IAQ study. Temperature, relative humidity, noise and brightness are also measured to inform on the occupants' overall comfort.

*Figure 1: Device developed by LARIS lab.*

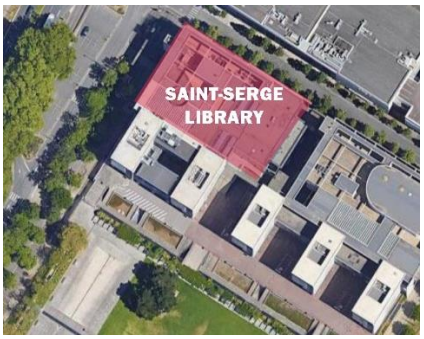


*Figure 2: Saint-Serge library location.*

The location chosen for the installation of the sensors is the Saint Serge library at the University of Angers (Figure 2); it was built in 1998, with a total surface area of 5004 m², and has an annual energy consumption of approximately 378.9 MWh. Additionally, the library has an average occupancy of 120 individuals, varying according to the time of the day and season. The selection of this site stems from the desire of the manager of the university libraries for carefully monitor IAQ in this location in response to the challenges posed by the Covid19 pandemic. The main objective of this monitoring is to detect pollutants levels in the air, an essential indicator for assessing ventilation and air circulation in the indoor environment. A high $CO_2$ concentration may indicate poor ventilation, which could potentially increase the risk of virus transmission (WHO 2021) (Ferrari et al. 2022). The sensors were strategically placed in the library to cover a variety of locations, ensuring comprehensive monitoring of IAQ. The sensor ENTRANCE was positioned on the first floor, near the main door to monitor variations due to entries and exits. The sensor PRINTZONE was placed on the first floor, close to the library printers, to assess their impact on air quality. The sensor STUDYSPOT was installed on level 1, near a work area most frequented by students. Finally, the sensor BOOKHALL was fixed on the second level, in a corridor near the book section.

## 2.2. METHODOLOGY

The study methodology is based on a systematic approach in three key steps (Figure 3), designed to (i) collect, (ii) analyze, (iii) estimate $CO_2$ data accurately, and detect anomalies. The first step focuses on collecting data from the four sensors. In this study, the only data used from our IAQ device was $CO_2$ concentration: we aim to predict the CO2 concentration. Data pre-processing is also carried out in this step, involving the cleaning of the collected data. In this step, identifying outliers or incorrect values in the data involved by applying specific thresholds to each type of data, these thresholds represent values outside the measurement range of the sensors. The second step, dedicated to in-depth data analysis, comprises two parts: the first part involves measurements visualization and correlation with meteorological data, as well as data regarding the presence of occupants indoors. Here, correlations are utilized to identify parameters correlated with $CO_2$ data, by calculating the Pearson correlation coefficient between each parameter and the $CO_2$ data, insights are gained into which parameters are most strongly correlated with $CO_2$ levels. Which are later on employed as training parameters for the models. The second part entails the comparison of ML techniques for estimation. In this study, we employed Python programming language, leveraging the scikit-learn library, a popular ML toolkit in Python, to implement and compare various ML models for the estimation task. To assess the estimation performance of different ML models, three key metrics were used, which are coefficient of determination ($R^2$), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics

allow us to assess the precision, dispersion and accuracy of our models. A set of regression models is used, namely the Gradient Boosting Regressor (GBR), the K-Nearest Neighbors Regressor (KNN), the Random Forest Regressor (RF) and the Decision Tree Regressor (DTR). For each model, a 10-fold cross-validation was performed by using the appropriate features for each sensor, Training values were taken randomly from the available year of data. During each iteration, nine parts were used for training and one part for validation, repeating the process 10 times. This method allows for rigorous assessment and comparison of each model's performance, after which the best-performing model for the estimation is selected. The third and final step is reserved for the estimation of IAQ values using the selected model, and the detection of any anomalies. This step is separated from the comparison of models and involves the training of our model on 11 months of data and making predictions for the last available month. Those 12 training and testing sets have been used for the anomaly detection step.

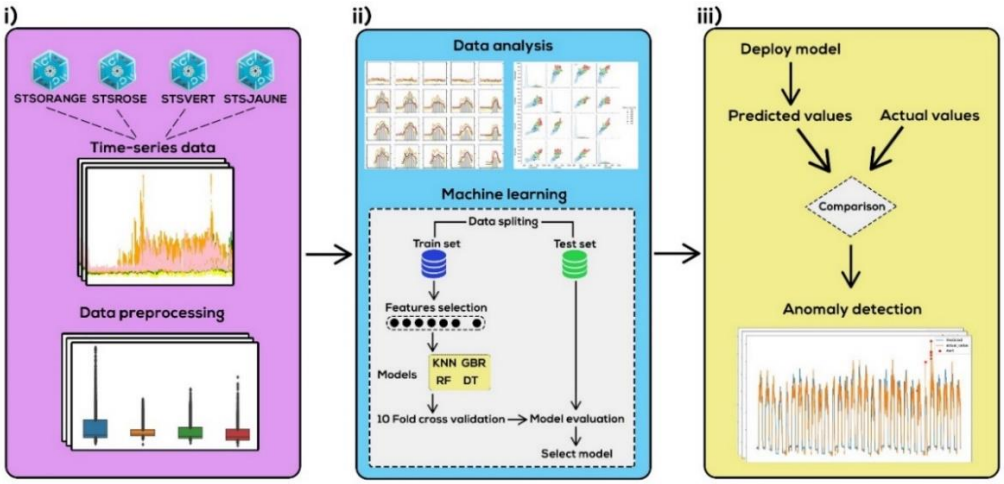


*Figure 3: Methodology (i) collect, (ii) analyze, (iii) estimate.*

# 3. RESULTS AND DISCUSSION

## 3.1. DATA DESCRIPTION

During this project, two types of data have been used. Firstly the $CO_2$ data is collected approximatively one year, from 13 July 2022 to 30 June 2023. The data is recorded for each minute and then averaged by computing the mean over every 10-minute interval. After the preprocessing, the percentage of data cleaned for each sensor was calculated enabling us to assess the loss of data resulting from the cleaning process. These percentages vary according to each sensor. It is important to note that despite cleaning, the percentage of data cleaned remains relatively low (always less than 0.74%). The second type of data is the occupancy. This data is collected by presence sensors at regular 30-minute intervals. However, for greater consistency with $CO_2$ data, the occupancy data was adapted by converting it into 10-minute intervals. This means that we have taken the presence value measured during each 30-minute interval and spread it over three consecutive 10-minute intervals.

The features used for estimation included temporal information such as day of year, hour, day of week, quarter, month, year, day of month, as well as number of occupants. The data predicted in this study specifically pertains to $CO_2$ levels, while the consideration of the other substance measured by our device (as input or output) will be explored in future works.

### 3.2. MODELS COMPARISON

Figure 4, Figure 5, and Figure 6 represent a set of boxplots for each model, illustrating the performance of the estimation in terms of $R^2$, RMSE, and MAE. Table 1 shows the mean of each boxplot, representing the average values across the 10 cross-validation iterations. In 10-fold cross-validation, the fold of testing is taken randomly from the data, ensuring that the evaluation is not influenced by any specific ordering of data. Results indicates that the Random Forest Regressor is the best performing model for estimating $CO_2$ levels, with mean $R^2$ of 0.96, 0.95, 0.98 and 0.97 respectively for the STUDYSPOT, ENTRANCE, PRINTZONE and BOOKHALL sensors. The Gradient Boosting Regressor model on the other hand, comes in last, with mean $R^2$ of 0.88, 0.78, 0.89 and 0.87. In terms of RMSE and MAE, the Random Forest Regressor also showed the lowest values of 33.7, 25.7, 17.2, 19.3 ppm, and 20.5, 17, 11, 12.1 ppm respectively, indicating better overall accuracy.

| | STUDYSPOT | | | ENTRANCE | | | PRINTZONE | | | BOOKHALL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE (ppm) | MAE (ppm) | $R^2$ | RMSE (ppm) | MAE (ppm) | $R^2$ | RMSE (ppm) | MAE (ppm) | $R^2$ | RMSE (ppm) | MAE (ppm) |
| **GBR** | 0.88 | 61.9 | 40 | 0.78 | 54.4 | 37.6 | 0.89 | 45.3 | 31.4 | 0.87 | 41.6 | 28.4 |
| **KNN** | 0.92 | 50 | 29.9 | 0.86 | 44 | 27.3 | 0.94 | 34.4 | 21 | 0.91 | 32.9 | 19.9 |
| **RF** | 0.96 | 33.7 | 20.5 | 0.95 | 25.7 | 17 | 0.98 | 17.2 | 11 | 0.97 | 19.3 | 12.1 |
| **DTR** | 0.96 | 35.5 | 21.22 | 0.94 | 27.1 | 17.7 | 0.98 | 17.93 | 11.1 | 0.96 | 20.9 | 12.65 |

*Table 1 : Model Performance by Sensor.*

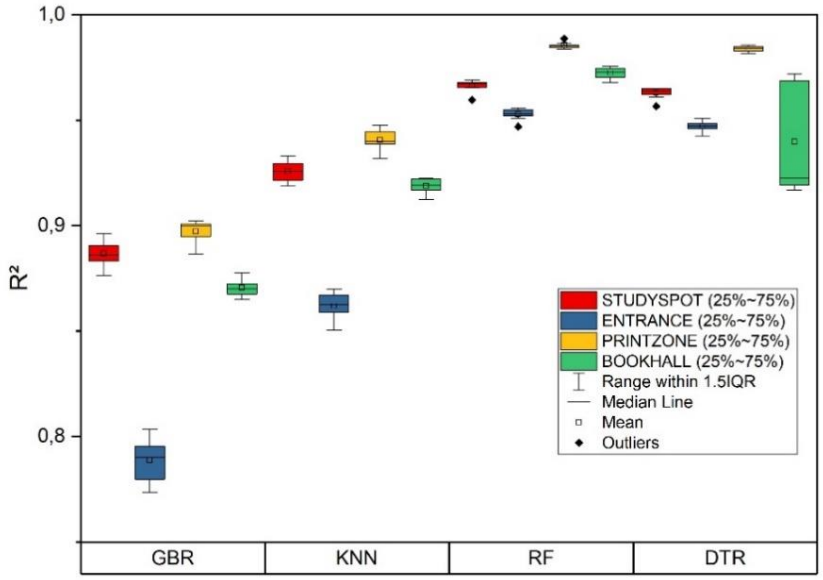


*Figure 4: Boxplots of $R^2$ scores for estimation by sensors.*

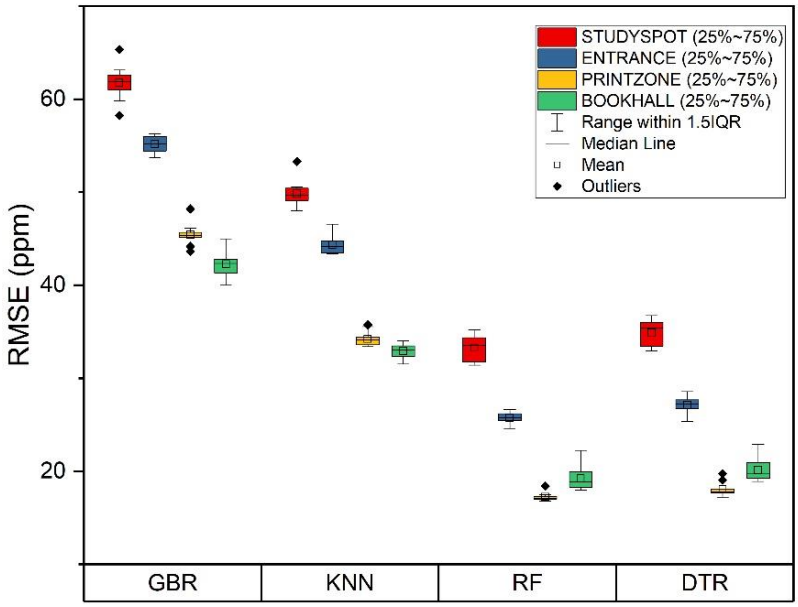


*Figure 5: Boxplots of RMSE scores for estimation by sensors.*

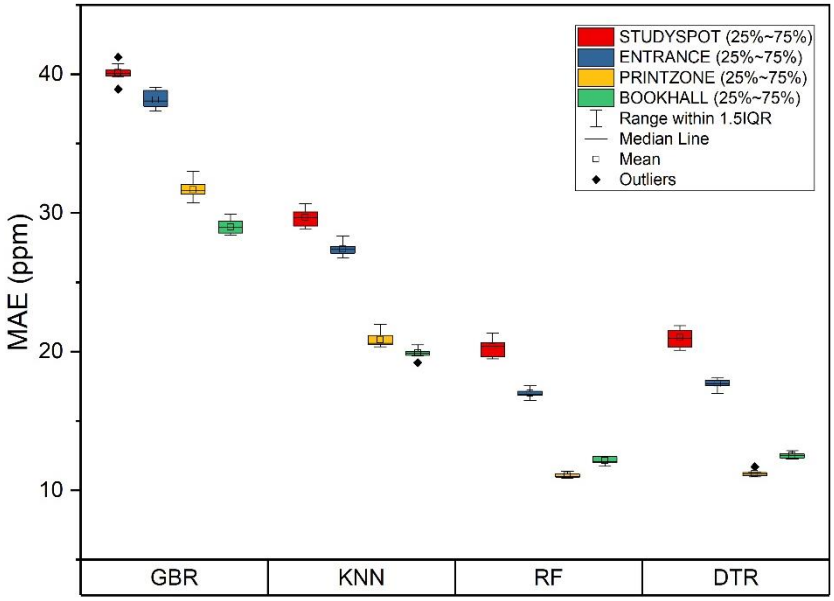


*Figure 6: Boxplots of MAE scores for estimation by sensors.*

## 3.3. ANOMALIES DETECTION

Anomaly detection involves identifying $CO_2$ values that deviate significantly from established estimation models. Anomalies can signal potential problems in the indoor environment, such as unexpected pollution levels or ventilation system malfunctions. To this end, the Random Forest model was used to estimate expected values of IAQ parameters, as it was the best-performing model in our case study.

Figure 7 and Figure 8 shows an example of August 2022, in which data of July 2022 and data from September 2022 to June 2023 are used for training, while data of August 2022 is used for detecting anomalies. Once the Random Forest model has been trained, we apply it to August data to estimate expected $CO_2$ levels knowing for the same time step the number of occupants, the year, the day of year, the quarter, the month, the day of month, the day of the week, and the hour. When the measured value

exceeds the bounds of the 95% prediction interval given by the RF model, it is considered as an anomaly. The results reveal that the STUDYSPOT sensor detected anomalies in most of occupied days. Figure 8 show graph of August18th, where 11 anomalies are detected, we can remark that almost all anomalies are detected when the occupancy decrease significantly such as the midday break or at the end of work hours. Let's remind that the STUDYSPOT sensor is located in a highly occupied workspace where the CO2 level can be high due to human activity. Additionally, it's important to note that we only have an idea about the occupancy of the entire library, not by specific zones. Therefore, when the total number of occupants decrease in the whole library, the area around the STUDYSPOT sensor may still have high occupancy levels. This can be seen during lunch time and on August 18, were the grey bars indicates a decrease of occupancy at the library scale and an increase CO2 is measured by the STUDYSPOT sensor (orange line). In addition, at the end of the day, the CO2 levels decreases more slowly than expected, as the share of students staying in the library may be larger in the STUDYSPOT than in the rest of the library. This approach enables us to identify the days and times when $CO_2$ levels are higher than expected, providing crucial information that can be useful to improve the IAQ. In our case, the solution might be to distribute the sensors more widely throughout the library, or to assess the number of occupants by zone. Additionally, our methodology distinguishes itself from the conventional threshold approach, such as using a threshold of 1000 ppm, because even below 1000 ppm, we can discern anomalous sources of $CO_2$ in the environment, allowing a more comprehensive assessment of IAQ.

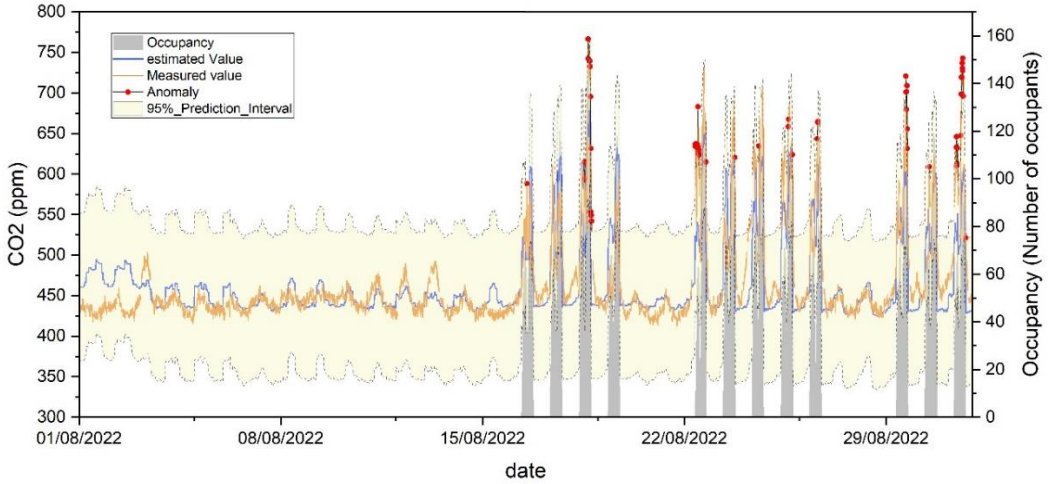


*Figure 7: Graph of STUDYSPOT sensor anomalies for $CO_2$ in August.*

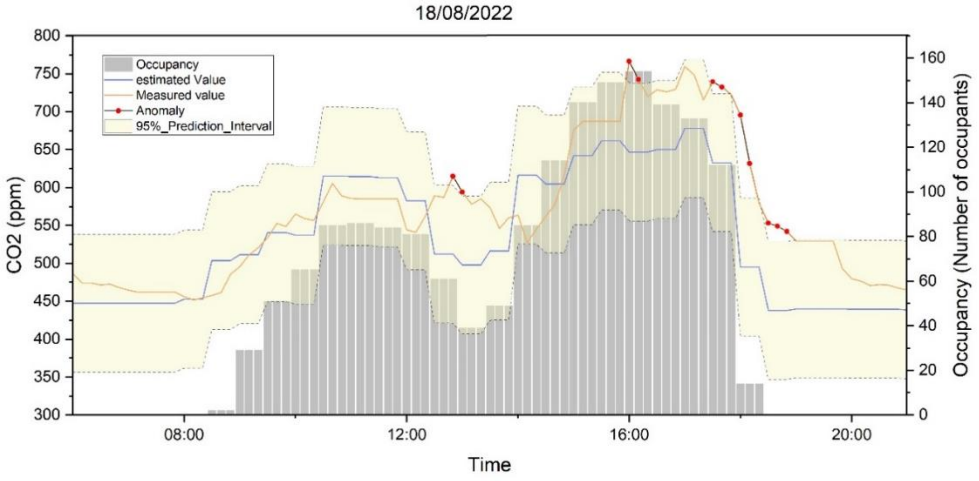


*Figure 8: Graph of STUDYSPOT sensor anomalies for $CO_2$ in August 18th.*

## 4. CONCLUSION

This study focused on developing an anomaly detection technique to identify $CO_2$ values that significantly deviate from normal concentrations, indicating potential issues within indoor environments. Four models namely K-Nearest Neighbors, Random Forest, Gradient Boosting Regressor and Decision Tree Regressor were compared. Results showed that Random Forest Regressor is the best performing model for estimating $CO_2$ levels, with mean $R^2$ of 0.96, 0.95, 0.98 and 0.97 respectively for the STUDYSPOT, ENTRANCE, PRINTZONE and BOOKHALL sensors, and lowest RMSE and MAE values of 33.7, 25.7, 17.2, 19.3 ppm, and 20.5, 17, 11, 12.1 ppm respectively, indicating better overall accuracy. It is important to note that this study was based on one year's data. Therefore, to further refine this work, it would be necessary to extend this analysis over several years and to add relevant features. Such an extension would enable better adaptation and training of the models, resulting in more accurate predictions. Additionally, expanding our evaluation to include other IAQ parameters will provide a more comprehensive understanding of indoor environmental quality. Furthermore, integrating derivatives into the models would allow predicted values to be calculated based on previous measurements, enhancing prediction accuracy and overall performance of the system.

## 5. ACKNOWLEDGEMENT

This work was performed within the frame of the research projects : CoLoC (Comfort of Connected Dwellings) granted by the “France Relance” plan for the preservation of R&D employment, and ACQA (Air Quality Acquisition and Characterization) granted by the University of Angers as part of the call for projects on participatory science.

## 6. BIBLIOGRAPHIE

Chen, Shisheng, Kuniaki Mihara, et Jianxiu Wen. 2018. « Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points ». *Building and Environment* 146: 238-46. https://doi.org/10.1016/j.buildenv.2018.09.054.

Di Gilio, Alessia, Jolanda Palmisani, Manuela Pulimeno, Fabio Cerino, Mirko Cacace, Alessandro Miani, et Gianluigi de Gennaro. 2021. « CO2 concentration monitoring inside educational buildings as a strategic tool to reduce the risk of Sars-CoV-2 airborne transmission ». *Environmental Research* 202 : 111560. https://doi.org/10.1016/j.envres.2021.111560.

Diffey, B.L. 2011. « An Overview Analysis of the Time People Spend Outdoors: Time Spent Outdoors ». *British Journal of Dermatology* 164 (4): 848-54. https://doi.org/10.1111/j.1365-2133.2010.10165.x.

Ferrari, S., T. Blázquez, R. Cardelli, G. Puglisi, R. Suárez, et L. Mazzarella. 2022. « Ventilation strategies to reduce airborne transmission of viruses in classrooms: A systematic review of scientific literature ». *Building and Environment* 222: 109366. https://doi.org/10.1016/j.buildenv.2022.109366.

Gameiro Da Silva, Manuel, Evandro Eduardo Broday, et Celestino Rodrigues Ruivo. 2023. « Indoor Climate Quality Assessment in Civil Aircraft Cabins: A Field Study ». *Thermal Science and Engineering Progress* 37: 101581. https://doi.org/10.1016/j.tsep.2022.101581.

Kallio, Johanna, Jaakko Tervonen, Pauli Räsänen, Riku Mäkynen, Jani Koivusaari, et Johannes Peltola. 2021. « Forecasting Office Indoor CO2 Concentration Using Machine Learning with a One-Year Dataset ». *Building and Environment* 187: 107409. https://doi.org/10.1016/j.buildenv.2020.107409.

Li, Wenzhuo, Choongwan Koo, Taehoon Hong, Jeongyoon Oh, Seung Hyun Cha, et Shengwei Wang. 2020. « A Novel Operation Approach for the Energy Efficiency Improvement of the HVAC System in Office Spaces through Real-Time Big Data Analytics ». *Renewable and Sustainable Energy Reviews* 127: 109885. https://doi.org/10.1016/j.rser.2020.109885.

Park, Dong-Uk, et Kwon-Chul Ha. 2008. « Characteristics of PM10, PM2.5, CO2 and CO monitored in interiors and platforms of subway train in Seoul, Korea ». *Environment International*, Assessment of Urban and Regional Air Quality and its Impacts, 34 (5): 629-34. https://doi.org/10.1016/j.envint.2007.12.007.

Senders, Joeky T., Mark M. Zaki, Aditya V. Karhade, Bliss Chang, William B. Gormley, Marike L. Broekman, Timothy R. Smith, et Omar Arnaout. 2018. « An Introduction and Overview of Machine Learning in Neurosurgical Care ». *Acta Neurochirurgica* 160 (1): 29-38. https://doi.org/10.1007/s00701-017-3385-8.

Taheri, Saman, et Ali Razban. 2021. « Learning-Based CO2 Concentration Prediction: Application to Indoor Air Quality Control Using Demand-Controlled Ventilation ». *Building and Environment* 205 (novembre): 108164. https://doi.org/10.1016/j.buildenv.2021.108164.

WHO. 2021. « Roadmap to Improve and Ensure Good Indoor Ventilation in the Context of COVID-19 ». 2021. https://www.who.int/publications-detail-redirect/9789240021280.