A Hybrid Forecasting Approach Integrating Statistical and Machine Learning Models for Indoor Air Quality Prediction Using Multi-Pollutant Sensor Data

Jeevitha P., Department of Computer Science and Applications, The Oxford College of Science, Arts, Commerce and Management, Bengaluru, Karnataka 560102

Vinita Tapaskar, Department of Computer Science and Applications, The Oxford College of Science, Arts, Commerce and Management, Bengaluru, Karnataka 560102

M M Ganavi, Department of Computer Science and Applications, The Oxford College of Science, Arts, Commerce and Management, Bengaluru, Karnataka 560102

Lokesh R, Department of Computer Science and Applications, The Oxford College of Science, Arts, Commerce and Management, Bengaluru, Karnataka 560102

Received:Oct 21, 2026 Accepted:Nov 18, 2026 Published:Nov 19, 2026

Abstract:

Indoor air quality (IAQ) refers to the condition of the air inside and around the building, and it directly affects the health and daily performance of people who live there. People spend 90% of their time indoors, where pollutant levels can be two to five times higher and occasionally up to 100 times higher than outdoor levels. The pollutants can be accumulated due to poor ventilation, occupancy, booking, cleaning products and outdoor infiltration. This study proposes a two stage framework for indoor air-quality and health-risk prediction using multi-pollutant sensor data. In the initial stage, Time series models are used to forecast future pollutant concentration. Environmental and pollutant variables include occupancy, temperature, humidity, PM1, PM2.5, PM10, CO2, CO, NO2, and VOC concentrations. Carbon dioxide (CO2) was selected as the primary forecasting variable because it responds strongly to occupancy and ventilation changes. In the second stage. Ensemble machine-learning models classify indoor health risk using pollutant concentrations, temperature, humidity, occupancy, and room type into high, medium and low. The dataset taken contains 25.004 indoor environment records from 5 types of rooms namely office, kitchen, classroom, living room, and bedroom. After cleaning, missing-value treatment, duplicate removal, type correction, and interquartile-range-based outlier handling, the processed data was used for exploratory analysis, forecasting, and classification. Models like Holt-Winters, ARIMA, and Prophet were trained and compared for CO_2 forecasting. The best forecasting model was chosen to be Holt-Winters producing the lowest error, with a mean absolute error of 118.231 ppm and a root mean square error of 150.053 ppm. For health-risk classification models trained and evaluated were Extra Trees, XGBoost, LightGBM, and CatBoost. LightGBM achieved the best overall classification performance, with 53.25% accuracy, 51.35% weighted precision, 53.25% weighted recall, and an F1-score of 0.4990. These results indicate that combining ensemble classification with pollutant forecasting can support early awareness of dangerous indoor air conditions and help building occupants or ventilation systems respond before exposure becomes harmful.

Keywords:*Indoor air quality(IAQ). Multi-pollutant sensing, Time-series forecasting, Ensemble learning, LightGBM, Health-risk classification, CO_2 forecasting*

I. Introduction

Air quality inside buildings directly affects comfort, productivity, and health. Unlike outdoor air pollution, indoor air pollution is affected by several factors, including occupant behaviour, room usage, ventilation, and nearby pollution sources. For example, when more people occupy a room, CO_2 levels can increase quickly. Activities such as cooking, cleaning, and the use of furniture and other household materials can also release pollutants such as particulate matter, carbon monoxide, nitrogen dioxide, and volatile organic compounds (VOCs). Continuous exposure to poor indoor air quality may lead to respiratory irritation, headaches, fatigue, allergic reactions, and difficulties in concentration and cognitive performance.

With the availability of low-cost sensors, it has become easier to monitor different indoor air pollutants continuously. However, simply measuring pollutant levels may not provide enough information about how indoor air quality is likely to change over time or what risks it may pose to occupants. For instance, a high CO_2 reading indicates poor ventilation at a particular moment, but it does not explain whether the concentration will increase or decrease in the coming hours. In addition, analysing individual pollutant levels may not provide a complete picture of indoor air quality, as factors such as particulate matter, gaseous pollutants, humidity, and occupancy can influence overall conditions.

To address these limitations, this study proposes a two-stage prediction framework for indoor air quality assessment. In the first stage, historical sensor readings are analysed using time-series forecasting techniques to predict future pollutant concentrations. In the second stage, machine-learning classification algorithms use pollutant levels and environmental parameters to identify the corresponding indoor health-risk category. This approach aims to extend indoor air-quality monitoring beyond the measurement of current conditions by providing an estimate of future air quality and possible health risks.

The main objectives of this study are:

1. To clean and preprocess a multi-pollutant indoor air-quality dataset for further analysis.
2. To study the relationships between occupancy, environmental conditions, pollutant concentrations, and health-risk categories.
3. To compare Holt-Winters, ARIMA, and Prophet models for forecasting future indoor CO_2 concentrations.
4. To evaluate the performance of Extra Trees, XGBoost, LightGBM, and CatBoost algorithms for indoor health-risk classification.
5. To demonstrate the proposed framework by predicting air-quality levels and health-risk categories for new indoor observations.

II. Literature Review

Indoor air quality has become an important research area because poor indoor environmental conditions can affect human health, comfort, productivity, and decision-making. Recent research has increasingly used sensor-based monitoring, statistical forecasting, and machine-learning methods to understand and predict indoor pollution. This section reviews studies related to multi-pollutant sensing, indoor pollutant forecasting, occupancy-driven CO_2 variation, and machine-learning-based health-risk prediction.

A. Machine Learning for Indoor Air Quality

Machine learning has been widely applied to indoor air-quality research because pollutant concentrations are influenced by nonlinear interactions among occupancy, ventilation, temperature, humidity, and emission sources. A

recent review by Machine Learning Applications with Sensors for Indoor Air Quality Research examined the use of machine learning with indoor air-quality sensors and highlighted the importance of multi-pollutant datasets, environmental variables, occupancy information, and data quality. The review supports the use of sensor-derived features such as PM1, PM2.5, PM10, CO_2, CO, NO_2, VOCs, temperature, and humidity for predictive modelling . This is directly relevant to the present study, which uses these variables to classify indoor health risk.[2]

Similarly, Indoor Air Quality Forecasting With Machine Learning: A Systematic Review analysed 99 peer-reviewed studies published between 2010 and 2026. The review examined data sources, modelling approaches, forecasting horizons, and performance metrics used in indoor air-quality prediction. It found that PM2.5 and CO_2 are among the most frequently studied indoor pollutants and that machine-learning and deep-learning methods are increasingly used for short-term forecasting. This supports the decision in the present study to forecast CO_2 concentrations and evaluate model performance using error metrics such as MAE and RMSE.[1]

Indoor air quality prediction systems for smart environments: A systematic review reviewed IAQ prediction systems designed for smart buildings and indoor environments. The authors highlighted the potential of sensor networks for real-time monitoring but also identified challenges related to sensor reliability, calibration, data availability, and model deployment . These findings are closely related to the present study, as the proposed framework uses continuous measurements from multiple indoor air-quality sensors. Proper data preprocessing is also important to ensure that the data is reliable before applying forecasting and classification techniques. [7]

A broader review titled Artificial Intelligence and Machine Learning Approaches for Indoor Air Quality Prediction explored the use of AI and ML techniques for predicting indoor air quality in different environments, including residential, educational, commercial, and public buildings. It discussed various models used to predict pollutants such as CO_2, PM2.5, PM10, VOCs, and formaldehyde. The review also highlighted how these techniques can support ventilation management and help identify potential health risks. These findings provide a foundation for the present study, which aims to use predicted pollutant levels and health-risk classifications to support timely decisions related to ventilation and occupancy management. [8]

Earlier work of Machine learning and statistical models for predicting indoor air quality reviewed statistical and machine-learning models for predicting indoor pollutants, including PM2.5, PM10, CO_2, and radon. The review reported that neural networks, regression models, and tree-based approaches were commonly applied to indoor air-quality prediction. The present research extends this by comparing several ensemble classifiers, including Extra Trees, XGBoost, LightGBM, and CatBoost. [6]

B. Multi-Pollutant Forecasting

Several studies have focused on forecasting multiple indoor pollutants rather than predicting a single parameter. Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: A case study compared Rolling Average, Random Forest, Gradient Boosting, and LSTM models for predicting eight indoor pollutant concentrations, including CO_2, NO_2, PM1, PM2.5, PM10, formaldehyde, and VOCs. This study is particularly relevant because it demonstrates that low-cost sensors and machine-learning models can be used for multi-pollutant prediction in real buildings. The present study follows a similar direction by using multiple pollutant measurements as inputs for health-risk classification.[2]

Another study compared multilayer perceptron, convolutional neural network, and LSTM models for indoor air-quality prediction. The study showed that deep-learning models can capture temporal dependencies in pollutant data and improve forecasting performance. Although the present study uses Holt-Winters, ARIMA, and Prophet rather than deep neural networks, this research supports the importance of time-series modelling for indoor pollutant forecasting.onlinelibrary.wiley[12]

Research on harmful indoor air pollution discussed a GRU-based air-quality prediction system using VOCs, CO_2, particulate matter, temperature, humidity, and light data. This study is relevant because it links multiple environmental measurements to indoor pollution prediction and emphasises pollutants that can affect human health. The present study similarly combines pollutant and environmental variables to estimate health-risk categories.[3]

Another study investigated the use of low-cost sensors for measuring pollutant concentrations, humidity, temperature, and other indoor environmental variables. The authors proposed combining physical knowledge with machine learning to improve prediction and interpretability . This is relevant to the present study because indoor CO_2 is strongly affected by occupancy and ventilation, while particulate matter and VOCs are influenced by emission sources and room characteristics.

A study on indoor air-pollution estimation using ANN and SVR used particulate-matter measurements and smart-building data, including temperature, humidity, occupancy-related variables, and window state. The study demonstrates that building-context features can improve pollution estimation. In the present study, room type and occupancy are included as predictive features along with pollutant concentrations and environmental conditions.[10]

C. CO_2, IoT, and Air-Quality Index Prediction

CO_2 is one of the most important indoor air-quality indicators because it is closely related to occupancy, ventilation, and human activity. The Impact of Occupancy Dynamics on Indoor CO_2 Forecasting benchmarked 11 forecasting models, including statistical methods, classical machine learning, deep learning, ensemble models, and foundation models, using 18 weeks of IoT sensor data. The study explored how changes in occupancy levels influence the accuracy of CO_2 forecasting. Its findings are relevant to the present work, as they highlight the importance of considering occupancy patterns when predicting CO_2 concentrations. This also supports the inclusion of occupancy as an important feature in the proposed health-risk classification model. [14]

The study titled IoT-Based Indoor Air Quality Monitoring and AI-Based Prediction Systems reviewed the use of IoT technologies for indoor air-quality monitoring and examined how AI techniques can improve pollutant prediction, sensor reliability, and system performance. It covered various pollutants, including PM2.5, PM10, NO_2, CO, and O_3, and discussed how AI-based systems can support ventilation control and generate health-related alerts. These findings are relevant to the present study, which aims to go beyond monitoring current air-quality conditions by using prediction techniques to support better indoor air-quality management and timely interventions. [15]

Another review examined IoT-based IAQ assessment methods and hybrid models combining artificial neural networks, autoregressive moving average models, and statistical techniques. The review is relevant because the present study also combines time-series forecasting with machine-learning classification in a hybrid framework.onlinelibrary.wiley[5]

A review of indoor environmental quality monitoring discussed recent work on data-driven models, statistical learning, and smart sensor technology. The study emphasised the role of data-driven models in identifying unhealthy indoor conditions and supporting health-oriented building management . This supports the health-risk classification component of the present study.

Finally, a study on indoor AQI monitoring using LSTM discussed sequence-based forecasting of indoor air-quality levels and dangerous gas pollution. The study is relevant because it demonstrates the use of deep-learning models for indoor air-quality forecasting. The present study complements this work by comparing classical forecasting models and then applying ensemble classifiers to estimate health risk from pollutant and environmental features.[13]

D. Research Gap

The studies reviewed indicate that machine-learning and time-series techniques are commonly used for indoor air-quality prediction. However, most existing research focuses either on forecasting future pollutant concentrations or on classifying the current air-quality condition. Only a limited number of studies consider both tasks together within a single framework.

Another limitation is that several studies rely on a single pollutant or consider only a few environmental factors during analysis. In contrast, the present study uses a multi-pollutant dataset that includes PM1, PM2.5, PM10, CO_2, CO, NO_2, VOCs, temperature, humidity, occupancy, and room type. The study also evaluates different time-series forecasting methods and ensemble-based classification algorithms to compare their performance. The key contribution of this work is the development of a two-stage prediction framework, where future CO_2 concentrations are first forecasted, followed by the classification of indoor health-risk levels using ensemble learning techniques.

III. Dataset and Study Design

The dataset used in this study contains 25,005 original records and 13 attributes. After removing one duplicate record, 25,004 observations remained. Each record represents an hourly indoor air-quality reading associated with a timestamp, room type, occupancy level, environmental conditions, pollutant measurements, and a health-risk label.

Table I

Description of Dataset

Variable	Description	Unit
Timestamp	Date and time of the sensor reading	Hourly
RoomType	Indoor location category	Classroom, Living Room, Kitchen, Bedroom, Office
Occupancy	Number of people present	Persons
TemperatureC	Indoor temperature	°C
HumidityPercent	Relative humidity	%
PM1	Particulate matter with diameter up to 1 µm	µg/m³

PM2.5	Particulate matter with diameter up to 2.5 µm	µg/m³
PM10	Particulate matter with diameter up to 10 µm	µg/m³
CO2	Carbon dioxide concentration	ppm
CO	Carbon monoxide concentration	ppm
NO2	Nitrogen dioxide concentration	ppb
VOC	Volatile organic compound concentration	ppb
HealthRisk	Health-risk category	High or Low

The dataset covers readings from January 2024 to November 2026. The records are arranged at hourly intervals, which makes the data suitable for time-series analysis. The health-risk target contains two classes: High risk, with 13,753 records, and Low risk, with 11,251 records

Timestamp	Room_Type	Occupancy	Temperature_C	Humidity_Percent	PM1_ug_m3	PM2_5_ug_m3	PM10_ug_m3	CO2_ppm	CO_ppm	NO2_ppb	VOC_ppb	Health_Risk
10/24/2024 7:00	Classroom	2.0	26.76	54.46	15.89	23.46	30.93	769.7	0.393	21.26	233.14	High
1/13/2024 10:00	Classroom	7.0	30.59	57.5	17.98	27.16	40.68	1353.3	0.785	32.15	322.51	High
7/8/2024 3:00	Living Room	1.0	23.77	68.42	8.99	12.35	26.66	495.3	0.485	20.38	223.12	Low
3/5/2026 0:00	Kitchen	1.0	NaN	74.19	21.70	35.91	57.69	607.1	0.720	24.7	268.98	High
6/3/2025 8:00	Classroom	1.0	27.95	53.07	6.69	14.67	28.86	717.8	NaN	18.11	194.66	High

Figure 1: Sample Records from the Indoor Air Quality Dataset

IV. Methodology

The research follows a hybrid workflow consisting of data preparation, exploratory analysis, time-series forecasting, classification modelling, and evaluation. The complete process is summarized below.

4.1 Data preprocessing

The raw dataset contained some missing values, one duplicate record, inconsistent data types, and invalid numerical entries. As part of the preprocessing stage, numerical columns were converted into suitable numeric formats. Missing

values in numerical features were replaced using the median, while missing values in categorical features were filled using the mode. The duplicate record was also identified and removed to avoid repetition in the dataset.

Outlier detection was performed using the Interquartile Range (IQR) method to identify unusually high or low values in the numerical features. The lower and upper limits were calculated using the following formulas:

Lower limit=Q1-1.5×IQR ……(i)
Upper limit=Q3+1.5×IQR ……(ii)

Where Q1 represents the first quartile, Q3 represents the third quartile, and IQR is the difference between the third and first quartiles (Q3 - Q1). Values falling outside these limits were treated as outliers. Rather than removing these records from the dataset, the extreme values were clipped to the corresponding lower and upper limits. This helped reduce the effect of extreme sensor readings while retaining all the observations for further analysis.

```
for column in numeric_columns:
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    lower_limit = Q1 - 1.5 * IQR
    upper_limit = Q3 + 1.5 * IQR

    outliers = ((df[column] < lower_limit) | (df[column] > upper_limit)).sum()

    print(f"{column}: {outliers} outliers")
```

```
Occupancy: 267 outliers
Temperature_C: 20 outliers
Humidity_Percent: 58 outliers
PM1_ug_m3: 534 outliers
PM2_5_ug_m3: 408 outliers
PM10_ug_m3: 491 outliers
CO2_ppm: 375 outliers
CO_ppm: 1293 outliers
NO2_ppb: 316 outliers
VOC_ppb: 256 outliers
```

Figure 2: Python Code and Output for Calculating the Interquartile Range (IQR) for Outlier Detection

4.2 Exploratory data analysis

Exploratory data analysis was performed to understand pollutant distributions, correlations, and differences across room types. The analysis included summary statistics, pollutant histograms, correlation heatmaps, PM2.5 box plots by room type, occupancy- CO2 scatter plots, and health-risk distribution charts.

The dataset shows an average occupancy of 1.77 persons, mean temperature of 25.17 °C, mean relative humidity of 62.50%, mean PM2.5 concentration of 19.34 µg/m³, and mean CO2 concentration of 652.95 ppm. The average CO

concentration was 0.472 ppm, while average NO2 and VOC concentrations were 22.47 ppb and 240.32 ppb, respectively

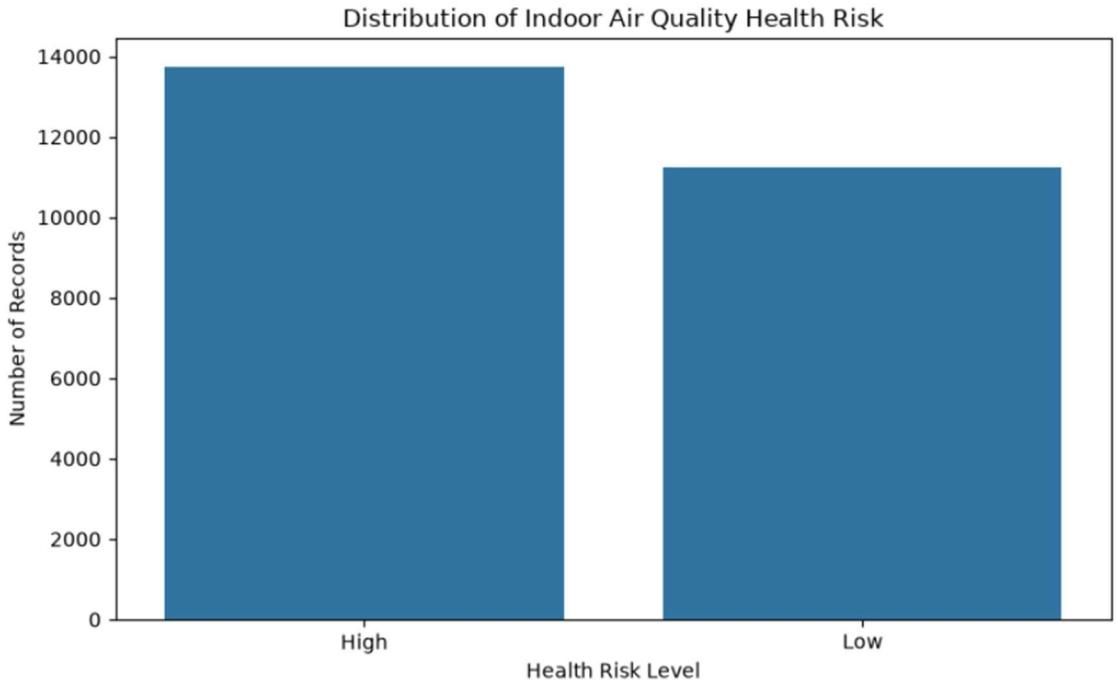


Figure 3:Nearly Balanced Distribution of High-Risk and Low-Risk Categories in the Dataset

The High-risk class represents approximately 55% of the cleaned dataset, while the Low-risk class represents approximately 45%. This relatively balanced distribution reduces the severity of class imbalance, although evaluation metrics beyond accuracy remain necessary.

The correlation heatmap illustrates the relationship between different indoor air quality parameters using the Pearson correlation coefficient. The correlation values range from -1 to +1, where positive values indicate that two parameters tend to increase together, while negative values indicate an inverse relationship. Values closer to zero represent weaker linear relationships.

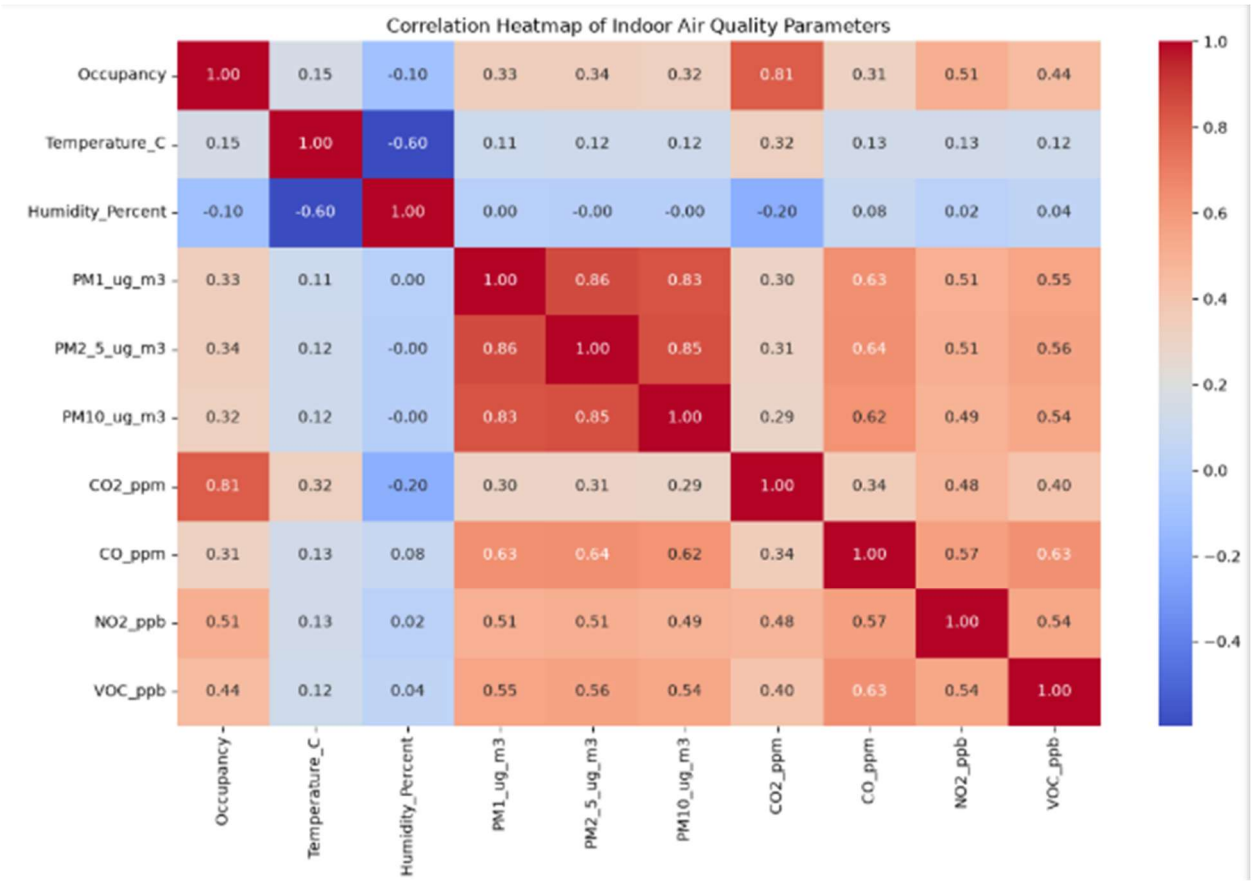


Figure 4: Correlation Heatmap of Indoor Air Quality Parameters

As observed in figure above, attributes Occupancy and CO_2 show a strong positive correlation of 0.81, indicating that higher occupancy is associated with increased CO_2 levels. Another strong positive correlation is observed between PM1.0 and PM2.5 (0.86), PM2.5 and PM10 (0.85), and PM1.0 and PM10 (0.83) suggesting the particulate matter concentrations tend to vary together.

Temperature and Humidity exhibit a moderate negative correlation of -0.60, indicating an inverse relationship between these parameters. Additionally, CO and VOC show a moderate positive correlation of 0.63, while NO_2 and CO show a correlation of 0.57.

Overall, the correlation analysis helps identify relationships and dependencies among indoor air quality parameters. These findings provide useful insights into the dataset and support further feature analysis and machine learning model development.

4.3 Time-series forecasting

CO2 was selected as the forecasting target because it is strongly associated with occupancy, ventilation, and indoor human activity. The CO2 readings were resampled to hourly averages, interpolated where necessary, and sorted chronologically.

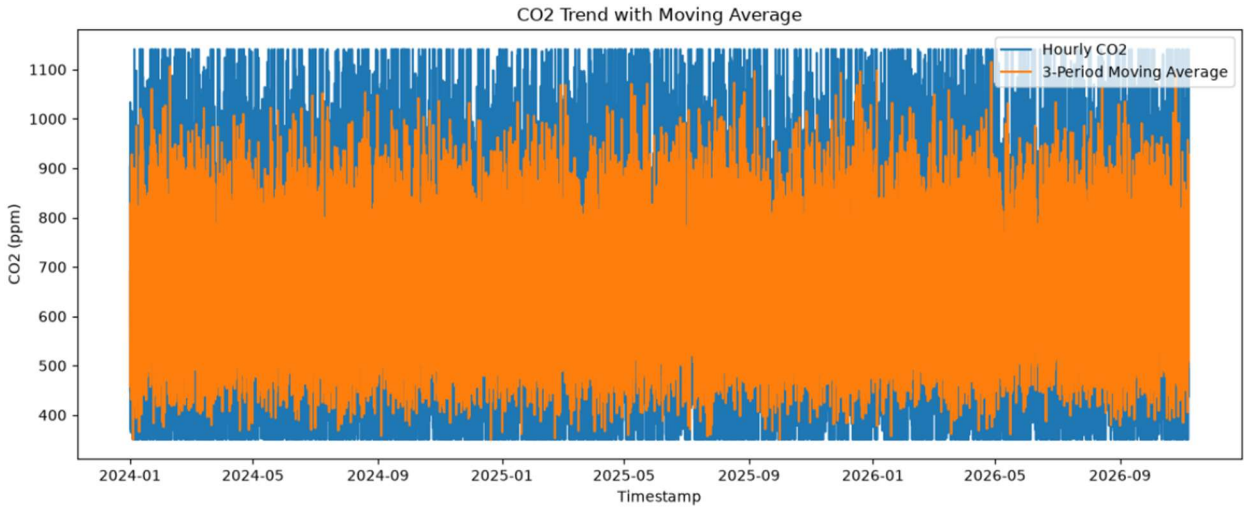


Figure 5: CO2 trend with moving average

The above figure shows the hourly CO_2 concentrations recorded between 2024 and 2026, along with a 3-period moving average. The raw CO_2 readings (blue line) show frequent fluctuations and sudden increases or decreases, indicating changes in indoor air quality over time. To make these variations easier to analyse, a 3-period moving average (orange line) was applied to smooth the data. This helps reduce short-term fluctuations and the effect of sudden changes while making the overall trend and recurring patterns in CO_2 concentrations more visible. The smoothed data can therefore provide a clearer view of the underlying trend for further forecasting and analysis.

The data were divided using a chronological 80:20 split, with 20,000 observations used for training and 5,000 for testing.

Three forecasting approaches were evaluated:

- Holt-Winters exponential smoothing: Captures level, trend, and seasonal components. A 24-hour seasonal period was used to represent daily variation.
- ARIMA: Uses autoregressive, integrated, and moving-average components to model temporal dependence.
- Prophet: A decomposable time-series model that incorporates trend, daily seasonality, and weekly seasonality.

4.4 Health-risk classification

The classification task predicts whether an indoor environment belongs to a High or Low health-risk category. The input features included room type, occupancy, temperature, humidity, and all pollutant measurements. The categorical room-type variable was encoded, and the health-risk target was label-encoded for model training.

Table II

Four ensemble classifiers used

Model	Main characteristic
Extra Trees	Highly randomized decision-tree ensemble
XGBoost	Gradient-boosted trees with regularization
LightGBM	Fast gradient-boosting method using leaf-wise growth
CatBoost	Gradient boosting with ordered target statistics for categorical features

The models were trained using the same preprocessed feature set and evaluated on the held-out test set. Performance was measured using accuracy, weighted precision, weighted recall, and weighted F1-score.

V. Results and Discusiion

5.1 Forecasting performance

The forecasting results show that Holt-Winters produced the lowest error among the three models. Its ability to model both trend and daily seasonality likely contributed to its stronger performance on the hourly CO2 series.

	Model	MAE	RMSE
0	Holt-Winters	118.231277	150.052766
1	Prophet	119.874541	152.025712
2	ARIMA	144.939729	180.766878

```
Best Forecasting Model: Holt-Winters
MAE: 118.231
RMSE: 150.053
```

Figure 6: Comparison of Forecasting Models Based on MAE and RMSE Scores, Highlighting Holt-Winters as the Best-Performing Model

From the above figure we can see that Holt-Winters achieved an MAE of 118.231 ppm and RMSE of 150.053 ppm, making it the best-performing forecasting model. Prophet performed similarly, with an MAE of 119.875 ppm and RMSE of 152.026 ppm. ARIMA produced the highest error, with an MAE of 144.940 ppm and RMSE of 180.767 ppm.

The small difference between Holt-Winters and Prophet suggests that both methods captured the dominant daily pattern in CO2 concentration. However, Holt-Winters was selected as the preferred forecasting model because it achieved the lowest RMSE.

Then using Holt-Winters forecasting was made of CO2 levels for the next 24 hrs.

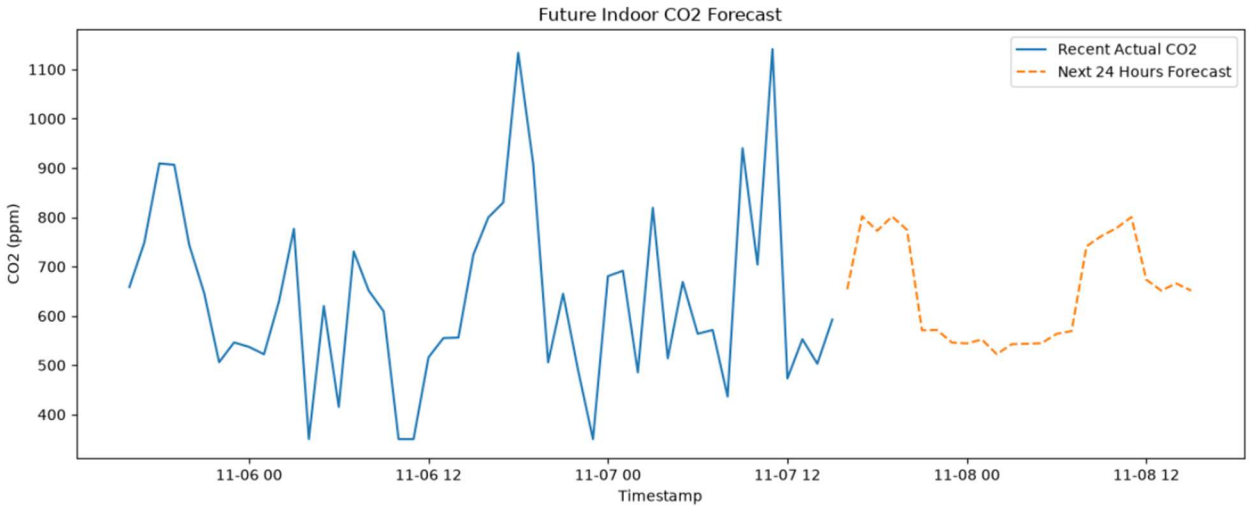


Figure 7: Future Indoor CO2 Forecast using Holt-Winters

To predict short-term air quality trends, the optimized Holt-Winters exponential smoothing algorithm was deployed to generate a 24-hour future indoor CO_2 forecast. As illustrated in above figure, the model effectively captures the cyclical, time-dependent nature of the actual historical data. The 24-hour forecast (represented by the dashed orange line) successfully replicates the expected diurnal variations, projecting key stabilization periods around 550 ppm during low-occupancy hours, alongside distinct peak surges matching historical occupant activity patterns

5.2 Classification performance

The classification results indicate that LightGBM provided the strongest overall performance among the four classifiers. Although the difference between models was moderate, LightGBM achieved the highest accuracy and F1-score.

Model	Accuracy	Precision	Recall	F1 Score
LightGBM	0.532494	0.513475	0.532494	0.499045
XGBoost	0.526495	0.508462	0.526495	0.498947
Extra Trees	0.517297	0.496494	0.517297	0.487920
CatBoost	0.539092	0.510228	0.539092	0.466863

Figure 8: Comparison of Machine Learning Models Based on Accuracy, Precision, Recall and F1 Scores, Highlighting LightGBM as the Best-Performing Model

From the above figure we can see that LightGBM achieved 53.25% accuracy, 51.35% weighted precision, 53.25% weighted recall, and an F1-score of 0.4990. CatBoost recorded slightly higher accuracy, at 53.91%, but obtained a lower weighted F1-score of 0.4669. Since F1-score balances precision and recall, LightGBM was selected as the best classification model.

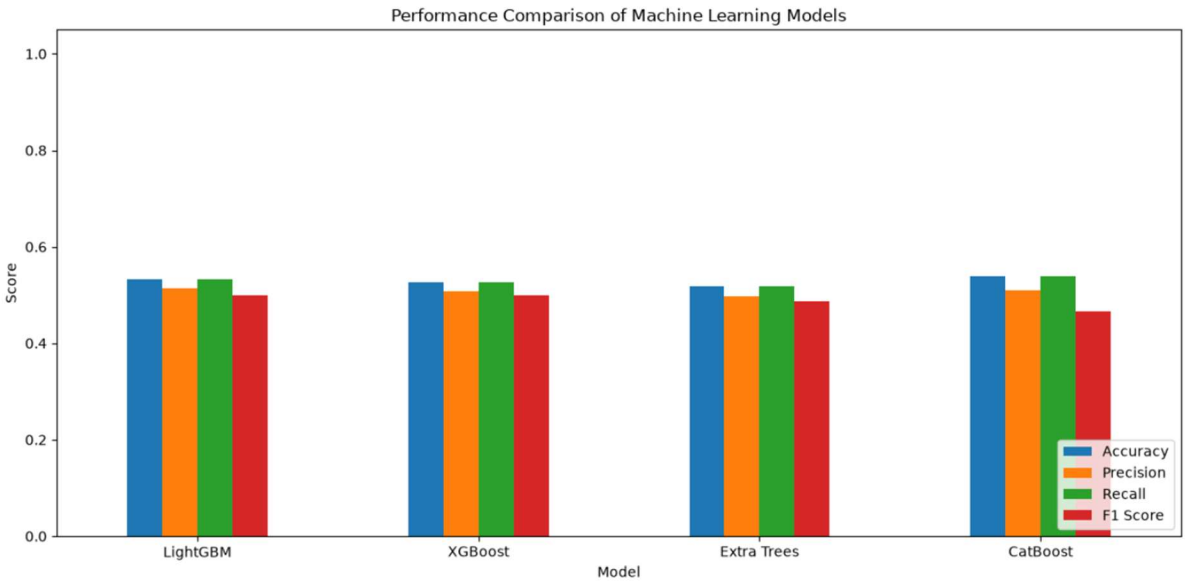


Figure 9: Comparison of Machine Learning Models Based on Accuracy, Precision, Recall and F1 Scores, Highlighting LightGBM as the Best-Performing Model

The classification report for LightGBM shows stronger performance for the Low-risk class than for the High-risk class. The model achieved a recall of 0.77 for Low-risk records but only 0.25 for High-risk records. This indicates that the model more frequently identifies low-risk conditions correctly but misses a substantial number of high-risk cases.

5.3 Interpretation of results

The forecasting results indicate that CO_2 concentrations in indoor environments follow certain patterns over time. Variations in hourly readings, changes in occupancy, and differences in ventilation appear to contribute to these patterns. The Holt-Winters model was able to capture these recurring variations and provide useful forecasts of future CO_2 levels.

However, the classification results showed comparatively lower performance. One possible reason is that indoor health-risk categories are influenced by several factors, including pollutant concentrations, room conditions, and occupancy levels. The available features may not fully capture the combined effect of these factors. The High-risk category was also more difficult for the models to identify. This is an important concern, as failing to detect high-risk conditions could affect the reliability and usefulness of an indoor air-quality warning system.

Future work can focus on improving the detection of high-risk conditions by applying techniques such as class weighting, resampling, feature engineering, and classification threshold adjustment. The model can also be improved by including additional information, such as outdoor air quality, ventilation status, weather conditions, and building characteristics. These improvements may help enhance classification performance and make the proposed framework more useful for real-world indoor air-quality monitoring.

VI. Conclusion

This study developed a hybrid forecasting and classification framework for indoor air-quality prediction using multi-pollutant sensor data. The forecasting results showed that Holt-Winters exponential smoothing performed better than the other tested models in predicting hourly CO_2 concentrations, with the lowest MAE and RMSE values. In the classification stage, LightGBM achieved the best overall performance among the evaluated models for predicting

indoor health-risk categories. However, identifying High-risk conditions remains a challenge and needs further improvement.

The findings show that combining time-series forecasting with machine-learning classification can be useful for smart indoor air-quality monitoring systems. This approach can help identify possible changes in pollutant levels in advance and provide early warnings when indoor conditions may become unhealthy. Future work can explore forecasting multiple pollutants, including more temporal features, and considering external factors such as outdoor air quality and weather conditions. Explainable AI techniques and more detailed health-risk categories can also be considered to improve the reliability and practical usefulness of the proposed framework.

References

[1] Rosca CM, Stancu A. Machine Learning Applications with Sensors for Indoor Air Quality Research. Sensors (Basel). 2026 May 6;26(9):2909. doi: 10.3390/s26092909. PMID: 42122630; PMCID: PMC13165709.

[2] Hall Rudford, Lewis, Sharifzadeh, Sara, Micallef, Nicholas, Lewis, Paul, Webb, Philip, Doneddu, Daniele, *Indoor Air Quality Forecasting With Machine Learning: A Systematic Review*, *Indoor Air*, 2026, 8524332, 24 pages, 2026. https://doi.org/10.1155/ina/8524332

[3] Saini J, Dutta M, Marques G. Indoor air quality prediction systems for smart environments: A systematic review. *Journal of Ambient Intelligence and Smart Environments*. 2020;12(5):433-453. doi:10.3233/AIS-200574

[4] Latoń, D., Grela, J., Ożadowicz, A., & Wisniewski, L. (2025). Artificial Intelligence and Machine Learning Approaches for Indoor Air Quality Prediction: A Comprehensive Review of Methods and Applications. Energies, 18(19), 5194. https://doi.org/10.3390/en18195194

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

[6] Mohammadshirazi, A., Ahmadi Kalkhorani, V., Humes, J., Speno, B., Rike, J., Ramnath, R., & Clark, J. D. (2022). Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: A case study. *Building and Environment, 213*, Article 108833. https://doi.org/10.1016/j.buildenv.2022.108833

[7] Altamirano-Astorga, J., Santiago-Castillejos, I. A., Hernández-Martínez, L., & Roman-Rangel, E. (2022). Indoor air pollution forecasting using deep neural networks. In O. O. Vergara-Villegas, V. G. Cruz-Sánchez, J. H. Sossa-Azuela, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, & J. A. Olvera-López (Eds.), *Pattern recognition (MCPR 2022)* (Lecture Notes in Computer Science, Vol. 13264). Springer. https://doi.org/10.1007/978-3-031-07750-0_12

[8] Mun, E., & Cho, J. (2023). Review of Internet of Things-based artificial intelligence analysis method through real-time indoor air quality and health effect monitoring: Focusing on indoor air pollution that are harmful to the respiratory organ. *Tuberculosis and Respiratory Diseases, 86*(1), 23–32. https://doi.org/10.4046/trd.2022.0087

[9] Mohammadshirazi, A., Nadafian, A., Monsefi, A. K., Rafiei, M. H., & Ramnath, R. (2023). Novel physics-based machine-learning models for indoor air quality approximations. *arXiv preprint arXiv:2308.01438*. https://doi.org/10.48550/arXiv.2308.01438

[10] Gabriel, M., & Auer, T. (2022). Indoor air pollution estimation using machine learning (ANN and SVR) in smart buildings. https://doi.org/10.26868/29761662.2022.24

[11] Garcia-Pinilla, P., Jurio, A., Figols, M., & Paternain, D. (2026). The impact of occupancy dynamics on indoor CO_2 forecasting: A cross-scenario evaluation. *Forecasting, 8*(2), Article 26. https://doi.org/10.3390/forecast8020026

[12] Ahmed, Shamima, Pasha, Anila, Kumer, Tusher, Akhtaruzzaman, Md, Hossain, Moktar, *Advancements in Air Quality Monitoring Systems: A Comprehensive Review of Emerging Technologies for Enhancing Environmental Health*, *Indoor Air*, 2026, 3080684, 23 pages, 2026. https://doi.org/10.1155/ina/3080684

[13] Wang, Z., Chen, Z., Shahid, I., Asif, Z., & Haghighat, F. (2025). Indoor air quality assessment through IoT sensor technology: A Montreal–Qatar case study. *Atmosphere, 16*(5), Article 574. https://doi.org/10.3390/atmos16050574

[14] Kidari, R., & Tilioua, A. (2024). Improving health through indoor environmental quality monitoring: A review of data-driven models and smart sensor innovations. *ITM Web of Conferences, 69*, Article 02007. https://doi.org/10.1051/itmconf/20246902007

[15] Talati, I., Shah, K., Patel, O., Tanna, J., Jain, A., Oza, A. D., Yadav, A. A., Alshayeb, M. J., Khan, M. A., & Islam, S. (2025). Study of AQI monitoring system of indoor environment using machine learning model and IoT device. *Rocznik Ochrona Środowiska, 27*, 152–163. https://doi.org/10.54740/ros.2025.013