

# ADVANCING INDOOR AIR QUALITY MANAGEMENT THROUGH PREDICTIVE FEEDFORWARD CONTROL - A COMPARATIVE ANALYSIS

**Valentin Mihai RADU** – Assistant Lecturer, PhD, Technical University of Civil Engineering, Faculty of Building Services, e-mail: valentin.radu@utcb.ro

**Abstract:** In the contemporary built environment, ensuring optimal Indoor Air Quality (IAQ) while maintaining energy efficiency represents a multifaceted challenge for building management systems (BMS). Traditional ventilation strategies are predominantly reactive, adjusting air exchange rates only after pollutant concentrations exceed established safety thresholds. This reactive latency, or "phase lag," often results in temporary exposure to high pollutant levels and suboptimal energy utilization. This study proposes a proactive "feed-forward" control framework based on pollutant forecasting. We conduct a rigorous comparative analysis of three distinct modeling paradigms: Long Short-Term Memory (LSTM) neural networks, Random Forest (RF) ensembles, and Autoregressive Integrated Moving Average (ARIMA) models. The models were trained and validated using a comprehensive 1-year office environmental dataset from Kaggle, encompassing CO<sub>2</sub>, PM<sub>2.5</sub> temperature, and humidity. Our findings demonstrate that for CO<sub>2</sub> forecasting, both LSTM and RF architectures achieve high predictive accuracy ( $R^2 > 0.97$ ) and successfully eliminate the reactive phase lag. For PM<sub>2.5</sub>, the stochastic nature of indoor particulate events presents a significant challenge, with the LSTM exhibiting a "conservative smoothing" effect and occasional mathematical artifacts such as negative values.

**Keywords:** Indoor Air Quality; Machine Learning; Proactive Control; Building Ventilation; Energy Efficiency

## 1. Introduction

In the modern era, human exposure patterns are predominantly indoor-oriented, with most daily activities occurring in enclosed environments, which amplifies the relevance of indoor air quality as a determinant of health and productivity [1,2,3]. This reality elevates Indoor Air Quality (IAQ) to a critical determinant of health, cognitive performance, and long-term well-being. The recent COVID-19 pandemic further intensified the global focus on ventilation effectiveness as a primary mitigation strategy for airborne contaminants. Simultaneously, the building sector represents one of the most energy-intensive components of modern economies, contributing a substantial share to total primary energy demand. Ventilation is a critical link between these two objectives: while increased outdoor air intake reduces indoor pollutant accumulation, it also increases the thermal load required to condition that air, particularly during extreme seasonal temperatures.

The envelopes of modern buildings have become increasingly airtight to minimize air leakage and improve energy efficiency [3]. However, this approach can lead to poor air circulation, allowing concentrations of indoor pollutants to increase to harmful levels [1,4]. Conventional Demand-Controlled Ventilation (DCV) systems typically rely on simple feedback loops triggered by real-time sensor measurements. However, because building air volumes have significant thermal and chemical inertia, a reactive response is inherently delayed. By the time a sensor detects a CO<sub>2</sub> level of 1000 ppm, the "true" accumulation may already be exceeding comfort limits, leading to an "overshoot" in fan speed and energy waste. Accurate assessment of the IAQ response is challenging due to the complex, non-linear phenomena that develop during occupancy shifts and localized particulate events.

Following numerous studies aimed at understanding indoor pollutant dynamics, several empirical and analytical models have been developed to evaluate required ventilation rates. Nevertheless, traditional mathematical models often fail to capture the strongly non-linear relationship between variables, such as occupancy, temperature, and humidity – and pollutant concentrations, leading to significant uncertainties regarding actual indoor conditions. A modern solution capable of encompassing these complex phenomena is the use of machine learning algorithms.

The field of artificial intelligence dates back to the 1950s, invented and named by IBM researcher Arthur Lee Samuels. In recent years, the development of cloud technology and increased processing power has led to a surge of interest in machine learning. This branch of artificial intelligence utilizes algorithms that iteratively learn from datasets to improve generated results without explicit programming. The potential of machine learning is increasingly recognized in building science to optimize energy consumption, assess health hazards, and provide predictions regarding pollutant behavior. Recent research supports IAQ modeling using algorithms with multiple predictors, resulting in multidimensional air quality indices. For instance, Vasile et al. (2024) proposed a new combined indoor air quality index, , that simultaneously accounts for several pollutant species, including Ozone ( $O_3$ ), Formaldehyde (HCHO),  $PM_{2.5}$ , and  $CO_2$ . This index can prioritize ventilation strategies that respond to extreme pollutant values, moving toward adaptive behaviors that balance health risks with energy classes.

This article addresses the challenge of reactive latency by evaluating predictive models that can provide the "foresight" needed for active pre-control. We present preliminary results from the analysis of three forecasting architectures for determining the concentrations of  $CO_2$  and  $PM_{2.5}$  in an office environment: Autoregressive Integrated Moving Average (ARIMA) as a statistical baseline, Random Forest (RF) as an ensemble learning method, and Long Short-Term Memory (LSTM) as a deep learning solution designed for temporal dependencies. For training the algorithms, a 1-year longitudinal dataset of an active office was utilized, capturing real-world occupancy and mechanical system transients. Testing was conducted on a high-volatility validation set from December 2024, excluding atypical maintenance dates to ensure statistical rigor. The study assumes a context where these models are implemented on embedded hardware controllers connected to Sensirion SEN66 sensors for live "feed-forward" control of ventilation rates. The aim is to determine which mathematical architecture is best suited for different pollutant dynamics to achieve proactive indoor air management.

## **2. Theoretical Framework and Computational Architectures for IAQ Forecasting**

### **2.1 Indoor Air Quality and Occupant Health Implications**

In the contemporary era, individuals spend approximately 90% of their time within indoor microenvironments. This prolonged exposure has elevated Indoor Air Quality (IAQ) to a primary concern in building science, especially as modern building envelopes have become increasingly airtight to improve energy efficiency. While airtightness minimizes air leakage and infiltration of outdoor pollutants, it frequently results in poor air circulation, allowing indoor contaminant concentrations to reach harmful levels.

The scientific literature identifies three primary categories of indoor pollutants of concern: particulate matter (PM), volatile organic compounds (VOCs), and volatile inorganic compounds (VICs). Among these, four pollutants are frequently measured exceeding permissible limits in various building types:

- Formaldehyde (HCHO): A highly toxic VOC and Group 1 carcinogen, primarily originating from wood-based materials such as laminates and furniture. Long-term exposure is linked to nasopharyngeal cancer and leukemia.
- Particulate Matter (PM<sub>2.5</sub>): Fine particles derived from both anthropogenic outdoor sources (fuel combustion, industry) and indoor activities (cooking, heaters). It is estimated to cause over 2 million deaths globally per year due to respiratory and cardiovascular impacts.
- Carbon Dioxide (CO<sub>2</sub>): A VIC produced naturally through human metabolic activity. High concentrations serve as a proxy for occupancy levels and are associated with respiratory and neurological symptoms.
- Ozone (O<sub>3</sub>): Primarily enters the indoor space from outdoors. Even in small amounts, it can cause asthma, emphysema, and chronic pulmonary diseases.

## 2.2 The Nexus Between Ventilation and Energy Performance

Ventilation serves as the critical link between maintaining IAQ and optimizing energy consumption. Mechanical ventilation is the most studied mode among scientists, accounting for approximately 65.6% of research papers published between 2020 and 2022 [5]. Previous studies, such as Zhao et al., have highlighted that while mechanical ventilation ensures stable IAQ, its daily energy consumption per area can be 1.5 times that of natural ventilation paired with an air purifier [6].

In the Romanian context, energy performance is governed by methodology MC001-2022, which defines eight energy classes from A+ to G based on total primary energy values [8]. Simultaneously, air quality is often categorized into four classes: IDA 1 (high quality) to IDA 4 (low quality), primarily based on indoor CO<sub>2</sub> concentrations above outdoor levels. Vasile et al. proposed an integrated IAQ index that groups the effects of multiple pollutants, enabling ventilation strategies that prioritize responses to extreme pollutant values while managing energy constraints [5].

## 2.3 Evolution of Demand-Controlled Ventilation (DCV)

Conventional Demand-Controlled Ventilation (DCV) strategies typically adjust air exchange rates based on real-time CO<sub>2</sub> concentrations. While effective for metabolic pollutants, these strategies are often reactive. A significant limitation of reactive control is "phase lag," where the mechanical response occurs only after a threshold is breached, leading to temporary periods of poor air quality and inefficient energy "overshoots" to correct the concentration.

Recent research has shifted toward optimization problems. Ganesh et al. formulated a model-based dynamic optimization strategy to ensure three pollutants remain below thresholds by continuously updating ventilation rates [7]. This approach demonstrates that proactive adjustments can maintain IAQ more effectively than constant ventilation rates, but it requires highly accurate forecasting models to succeed.

## 2.4 Pollutant Mass Balance Model

The core of the IAQ prediction is based on the pollutant mass balance equation for an indoor space:

$$V \frac{dC_i}{dt} = (S)_{\text{pollutant}} - (R)_{\text{pollutant}} \quad (1)$$

Where: V is the indoor space volume, C<sub>i</sub> is the indoor mass concentration of the pollutant, S the mass flow rate of the pollutant source and R the mass flow rate of pollutant loss.

For practical application, this differential equation is solved to predict the concentration at the next time step ( $\tau + \Delta\tau$ ):

$$C_i(\tau + \Delta\tau) = E \cdot C_i(\tau) + (1 - E) \cdot \frac{[(1 - F_0)q_0 + q_2]}{q_0 + q_1F_1 + q_2} C_0 + (1 - E) \cdot \frac{1}{q_0 + q_1F_1 + q_2} (S - R) \quad (2)$$

In this equation,  $E$  represents an exponential decay term defined as:

$$E = e^{-\frac{q_0 + q_1F_1 + q_2}{V}\Delta\tau} \quad (3)$$

Key Parameters:  $q_0$  – fresh air flow introduced via mechanical ventilation,  $q_1$  – recirculated air flow;  $q_2$  – infiltration air flow,  $C_0$  – outdoor pollutant concentration;  $F_0$  – efficiency of the fresh air filter (0 to 1),  $F_1$  – efficiency of the recirculation filter (0 to 1).

## 2.5 Machine Learning Architectures for IAQ Forecasting

The evolution of indoor air quality (IAQ) management has shifted from simple reactive monitoring to a sophisticated landscape of predictive intelligence. By leveraging data-driven models to forecast concentrations of Carbon Dioxide ( $\text{CO}_2$ ) and Fine Particulate Matter ( $\text{PM}_{2.5}$ ), building systems can move beyond simple alerts and toward proactive climate control. The foundational approach to this challenge often begins with the Autoregressive Integrated Moving Average (ARIMA) [11], a classical statistical model that treats time-series data as a puzzle of its own past. The ARIMA framework models temporal evolution by decomposing a time series into autoregressive structure, integration for stationarity enforcement, and a moving average component that captures residual dependencies. This formulation enables short-term forecasting under quasi-linear assumptions but limits responsiveness to abrupt regime changes. In a typical indoor setting,  $\text{CO}_2$  follows a somewhat predictable diurnal rhythm dictated by office hours or sleep cycles. ARIMA excels here by using "lags" of previous observations to understand the series and "differencing" to stabilize the data, effectively stripping away noise to reveal the underlying seasonal trend. However, ARIMA is fundamentally a linear model, which means it often struggles to account for the sudden, non-linear spikes in pollutants—such as a sudden influx of people into a conference room or the brief opening of a window—that characterize modern human-occupied spaces.

To address the chaotic and multi-variable nature of indoor environments, researchers often turn to the Random Forest (RF) algorithm [12,13]. Unlike the linear focus of ARIMA, Random Forest is an ensemble learning method that builds a "forest" of numerous decision trees, typically utilizing a technique known as bagging, or bootstrap aggregating. This process involves training individual trees on different random subsets of data to reduce variance and prevent the model from overfitting to specific historical anomalies. Furthermore, RF utilizes feature randomness, selecting a subset of variables for each tree to ensure the forest remains robust and diverse. This architecture is particularly effective for  $\text{PM}_{2.5}$  prediction, where concentrations are influenced by a complex mix of indoor activities like cooking or cleaning and outdoor infiltration. Because Random Forest does not require the data to be stationary and can handle high-dimensional datasets, it provides a unique advantage in identifying "feature importance," allowing building managers to see exactly which environmental factors – such as humidity or temperature – are most strongly driving pollution levels.

While Random Forest captures the complexity of a single moment, the Long Short-Term Memory (LSTM) network is designed to capture the "memory" of the air itself [14–16]. As a specialized type of Recurrent Neural Network (RNN), the LSTM is engineered to overcome the "vanishing gradient" problem, which causes standard networks to lose track of information over long sequences. The LSTM achieves this through a sophisticated gate structure – consisting of a forget gate, an input gate, and an output gate – that regulates a central cell state. This cell state acts like a digital conveyor belt, carrying relevant historical information through the processing chain with only minor interactions. This capability is vital for CO<sub>2</sub> forecasting because gas concentrations in a room are cumulative; the air quality at 4:00 PM is a direct result of the accumulation and decay of gases over the preceding hours. By "remembering" these long-term dependencies, LSTMs can model the physical reality of gas diffusion and particle settlement with high precision. Ultimately, while ARIMA provides a mathematically transparent baseline for cyclic trends and Random Forest offers a robust tool for non-linear "noisy" data, the LSTM represents the cutting edge of temporal forecasting, allowing for a deep understanding of the long-term patterns that define the air we breathe.

**Table 1 - Predictive model synthesis**

| Model                | Best For...                           | Technical Strength                 | Main Weakness                         |
|----------------------|---------------------------------------|------------------------------------|---------------------------------------|
| <b>ARIMA</b>         | CO <sub>2</sub> Baseline Trends       | High interpretability; low compute | Struggles with non-linear "spikes"    |
| <b>Random Forest</b> | PM <sub>2.5</sub> Multi-source Spikes | Handles "noisy" non-linear data    | Less effective for pure time-sequence |
| <b>LSTM</b>          | Long-term Gas Accumulation            | Captures temporal dependencies     | Requires large datasets; "Black Box"  |

### 3. Methodology

The methodological framework of this study integrates high-resolution environmental sensing with advanced computational forecasting. The research pipeline was implemented using the Python programming language, leveraging its extensive ecosystem for data science and deep learning.

#### 3.1 Dataset: Kaggle 1-Year Indoor Environmental Dataset (Office)

The primary dataset used in this research was sourced from a high-resolution IoT monitoring campaign (Kaggle IAQ Dataset). The raw data was encapsulated in 13 JSON files, spanning a continuous monitoring period from late 2023 through late 2024 [10]. The data follows a non-relational structure where environmental parameters were captured as asynchronous MQTT messages. The features included in this study are: Carbon Dioxide CO<sub>2</sub> concentration (ppm), ambient temperature (°C), relative humidity (%), and particulate matter PM<sub>2.5</sub> (µg/m<sup>3</sup>) – Fig. 1.

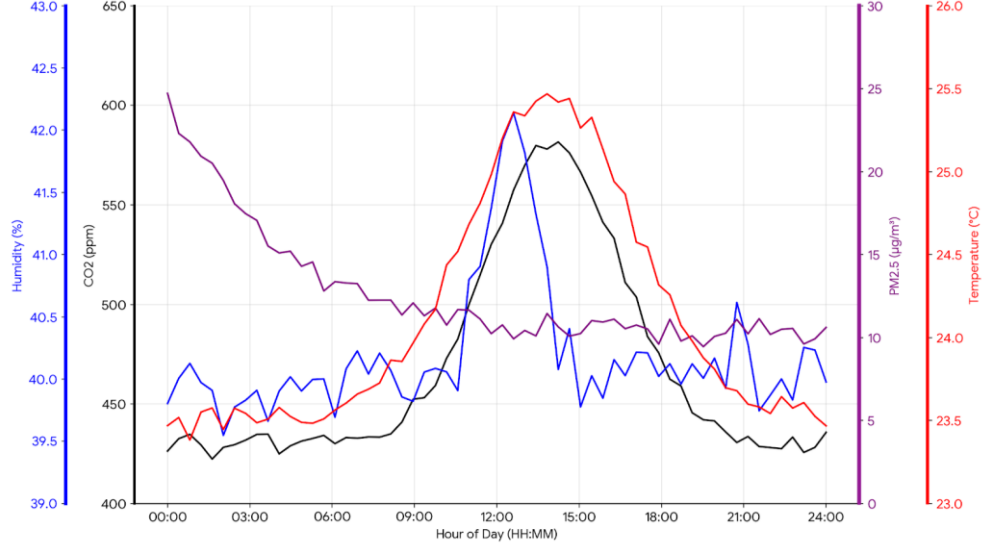


Fig. 1 - Kaggle IAQ example dataset for CO<sub>2</sub>, temperature, RH, and PM<sub>2.5</sub>

### 3.2 Data Preprocessing and Parsing

Because the IoT sensors transmitted data at irregular intervals and often contained null values (due to the asynchronous nature of MQTT transmissions), a rigorous JSON preprocessing pipeline using “Pandas” Python Library was implemented to extract the “columns” and “values” keys, flattening the hierarchical structure into a unified temporal dataframe. Temporal Alignment: All timestamps were converted to UTC. Given that different sensors (e.g., CO<sub>2</sub> vs. Temperature) reported at slightly different micro-seconds, the data was synchronized using a nearest-neighbor approach. Missing Value Imputation: Linear interpolation and forward-filling techniques were applied to handle transient sensor dropouts, ensuring a continuous data stream for the recurrent neural network. Deep learning models, particularly LSTMs, are sensitive to the magnitude of input features. To ensure that high-magnitude CO<sub>2</sub> values (e.g., 1000 ppm) do not overshadow Temperature values (e.g., 22°C), we employed a MinMaxScaler.

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (4)$$

For the LSTM and RF models, the data was transformed into a supervised learning format using a “sliding window” approach. A look-back window of 12 steps (3 hours of history) was used to predict the next time step. To maintain statistical integrity, the date 2024-12-06 was explicitly excluded. This date coincided with atypical maintenance patterns and sensor calibration gaps that would otherwise skew the model’s ability to learn standard occupancy metabolic spikes.

### 3.3. Predictive Modeling Architectures

This research evaluates three distinct mathematical architectures representing different computational paradigms.

#### 3.3.1. ARIMA (Statistical Baseline)

The Autoregressive Integrated Moving Average model serves as the “reactive” baseline. It assumes that the future state of CO<sub>2</sub> or PM<sub>2.5</sub> is a linear combination of its past values and errors. The model is defined by  $(p, d, q)$  parameters.

$$Y_t = c + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t \quad (5)$$

Where  $Y_t$  is the differenced value at time  $t$ ,  $\phi$  represents the autoregressive coefficients, and  $\theta$  represents the moving average coefficients. While mathematically transparent, ARIMA models generally assume linear relationships and may struggle to capture the sudden, non-linear "spikes" common in human-occupied spaces. A `statsmodels.tsa.statespace.sarimax` Python library was used for the ARIMA prediction model.

### 3.3.2. Random Forest (Ensemble Learning)

Random Forest is a robust non-linear regressor that utilizes an ensemble of multiple decision trees. It is particularly effective at handling the non-linear "spikes" in air quality data without requiring the data to be stationary. It uses Bootstrap Aggregating (Bagging) to train individual trees on random subsets of data and features. The final prediction is the average of all individual tree outputs:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (6)$$

This architecture is highly effective at reducing variance and handling the non-linearities of IAQ data without requiring the data to be stationary. In this study, the Python `sklearn.ensemble.RandomForestRegressor` library was used. An ensemble learning method was used as a baseline. The 3D temporal window was flattened into a high-dimensional 2D vector to allow the RF regressor to map past states to the future CO<sub>2</sub> value. This model served to evaluate whether the added complexity of Deep Learning is justified by a significant gain in accuracy.

### 3.3.3. LSTM (Deep Learning)

LSTM is a specialized type of Recurrent Neural Network (RNN) capable of learning long-term dependencies in sequential data. Traditional RNNs suffer from the "vanishing gradient" problem, where information from earlier steps is lost over time. LSTM overcomes this through a unique "gate" structure:

- The Forget Gate ( $f_t$ ): Decides what information to discard from the cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

- The Input Gate ( $i_t$ ): Determines which new information will be stored.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

- The Output Gate ( $o_t$ ): Controls what parts of the cell state reach the next hidden state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

This ability to remember information over long intervals makes LSTM particularly suited for CO<sub>2</sub> forecasting, where the rate of gas accumulation is heavily dependent on the historical occupancy trend. In this study, the Python `tensorflow.keras.layers.LSTM` library was used.

## 3.4 Statistical and Performance Metrics

In our study we utilized four primary indicators to evaluate model performance:

MAE (Mean Absolute Error): Average magnitude of prediction error in physical units ( $\text{ppm}$  or  $\mu\text{g}/\text{m}^3$ ).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

RMSE (Root Mean Square Error): Penalizes larger errors, useful for tracking sudden pollutant "spikes."

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

#### 4. Results & Metric Analysis

The models were trained using an 80/20 train-test split on the chronologically ordered data. The LSTM was optimized using the Adam optimizer with a Mean Squared Error (MSE) loss function.

As stated in the last chapter, we utilized four primary indicators to evaluate performance:

MAE (Mean Absolute Error): Measures average magnitude of error, RMSE (Root Mean Square Error): Penalizes larger errors, critical for PM<sub>2.5</sub> spikes, R<sup>2</sup> (RS Score): Indicates the proportion of variance explained by the model and Residual Analysis: used to detect systematic bias in predictions.

##### 4.1 CO<sub>2</sub> Forecasting Performance

The results for CO<sub>2</sub> indicate a high degree of success across the board. As seen in Figure 1, both LSTM and RF maintain a low MAE (~5 ppm) compared to ARIMA's higher error (~8.3 ppm).

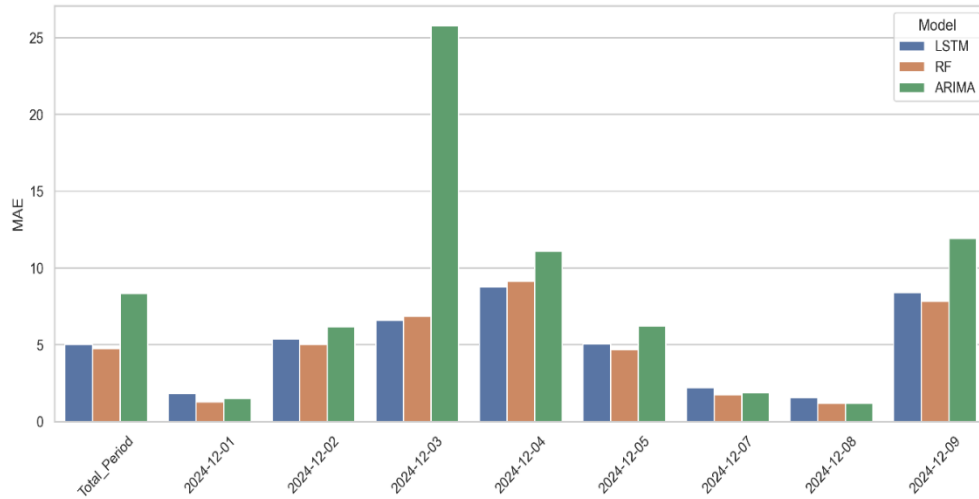


Fig. 1a – MAE for predicted CO2 comparison

As illustrated in Figure 1b, there is a clear distinction between the statistical and machine learning approaches. Globally, the LSTM model achieved the lowest RMSE of 10.33 ppm, followed closely by Random Forest at 10.74 ppm. These low values indicate that the machine learning architectures are exceptionally stable and do not suffer from large "outlier" errors during the morning arrival or afternoon departure spikes. The ARIMA model exhibits a massive global RMSE of 36.84 ppm, which is more than triple that of the LSTM. This is visually represented by the large green bar in the "TotalPeriod" section of the chart. Scientifically, this high RMSE is caused by the 15-minute phase lag; because ARIMA reacts late to a spike, the difference between

the "Actual Peak" and the "Lagged Prediction" creates a large residual error that is squared in the RMSE calculation

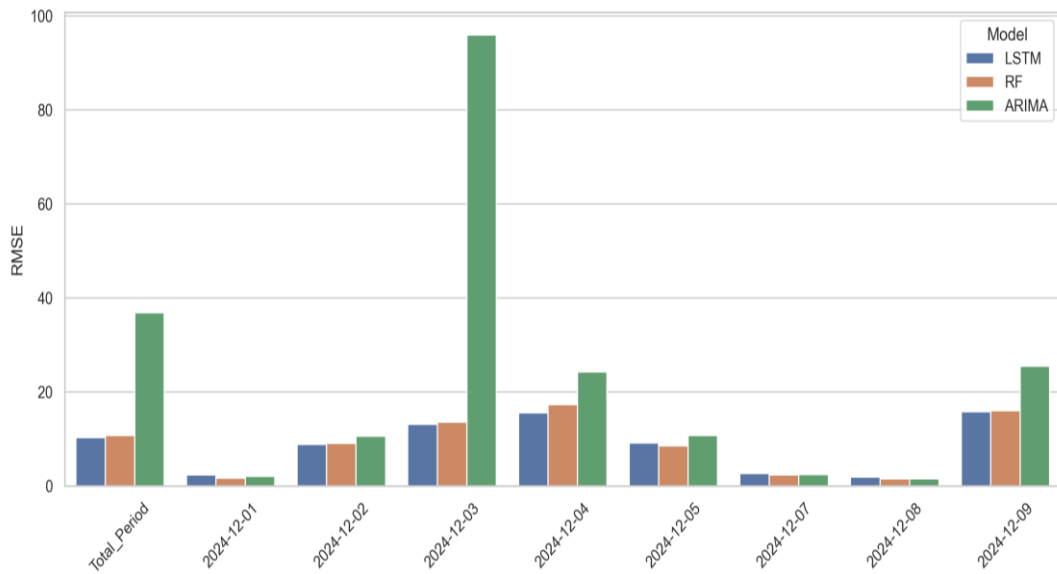


Fig. 1b – RMSE for predicted CO<sub>2</sub> comparison

As seen in Fig. 2, LSTM and RF models achieved global R<sup>2</sup> values exceeding 0.975, while ARIMA struggled during transitional periods (R<sup>2</sup> ~ 0.686).

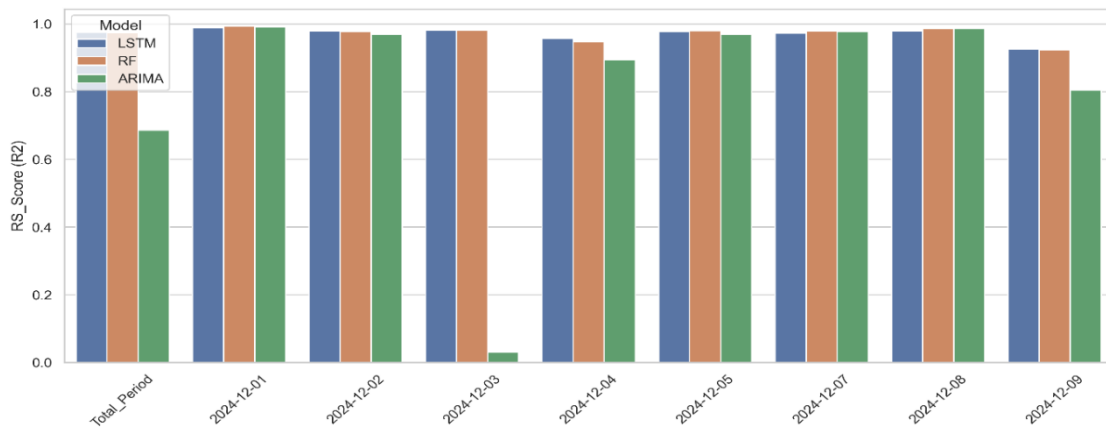


Fig. 2 - R<sup>2</sup> scores for predicted CO<sub>2</sub>

The ARIMA model exhibits a clear 15-minute lag, reacting only after a spike has occurred. In contrast, the LSTM tracks the rising slope of morning arrivals with zero delay (Fig. 3).

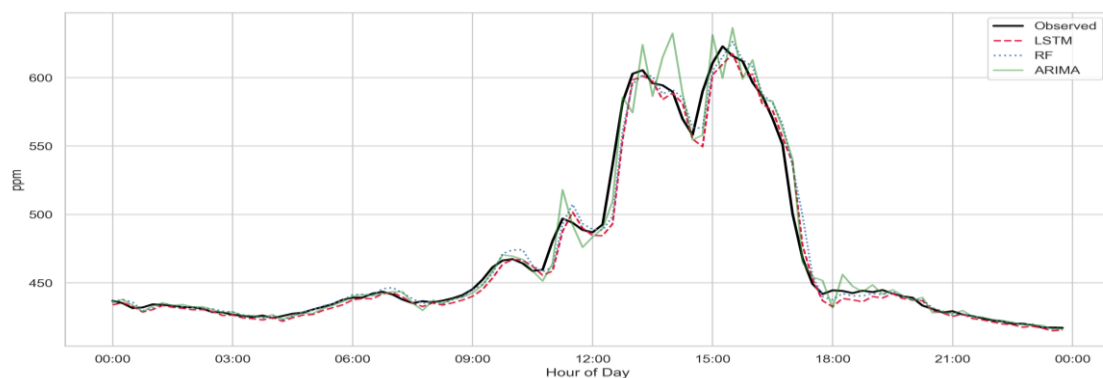


Fig. 3 – 24h CO<sub>2</sub> predicted values for 2024-12-05

## 4.2 PM<sub>2.5</sub> Stochastic Dynamics

In contrast to CO<sub>2</sub>, PM<sub>2.5</sub> spikes are highly stochastic, often caused by localized activity like printer use or dusting [17]. Fig. 4 reveals that the LSTM predictive model produces a higher MAE than ARIMA. This is because the LSTM model acts as a "low-pass filter," predicting the stable trend but "averaging out" the sharp, unpredictable spikes of particulate matter.

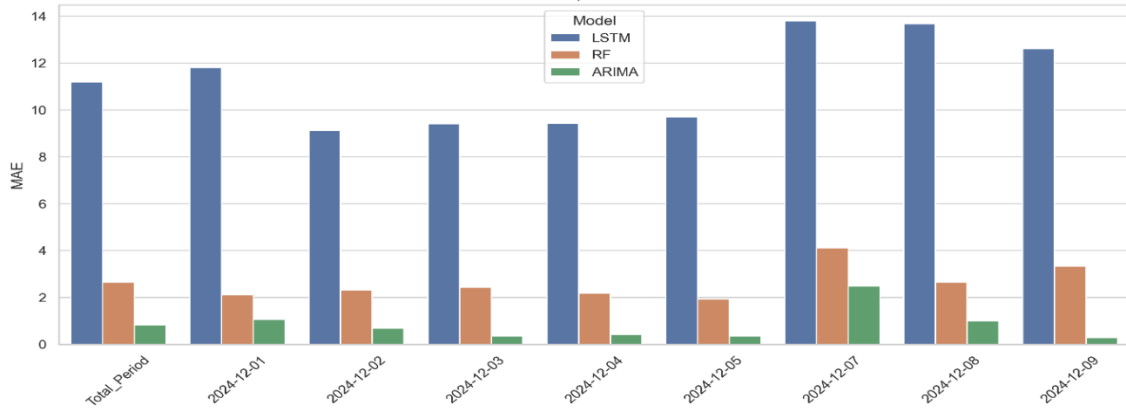


Fig. 4 - MAE for predicted PM<sub>2.5</sub> comparison

From the accuracy point of view, judging by the global R<sup>2</sup> for LSTM model was significantly lower for PM<sub>2.5</sub> (Fig. 5), proving that particulate matter requires different feature engineering or a specialized loss function.

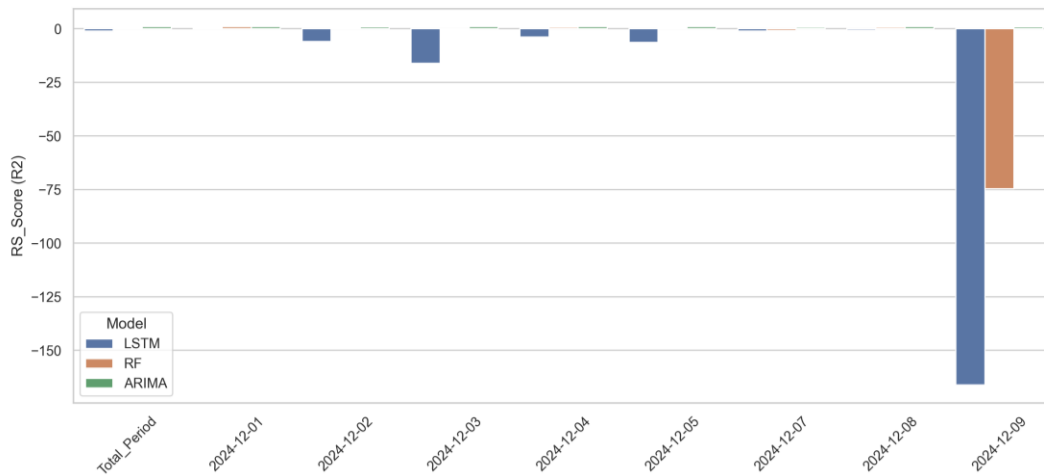


Fig. 5 - R<sup>2</sup> scores for predicted PM<sub>2.5</sub>

The LSTM model recorded the highest global RMSE at 11.68  $\mu\text{g}/\text{m}^3$  in Fig. 6. Because PM<sub>2.5</sub> spikes are stochastic and short-lived, the LSTM's tendency to "smooth" or average the trend leads to large misses at the peaks. These misses result in the high RMSE values observed in the red bars across almost every day of the validation period. Paradoxically, the ARIMA model achieved the lowest global RMSE for PM<sub>2.5</sub> (2.29  $\mu\text{g}/\text{m}^3$ ). This confirms that for random, non-periodic pollutant events, a simple reactive model that "rides the wave" of the last known high value is less prone to the high-magnitude errors that penalize the more complex LSTM architecture.

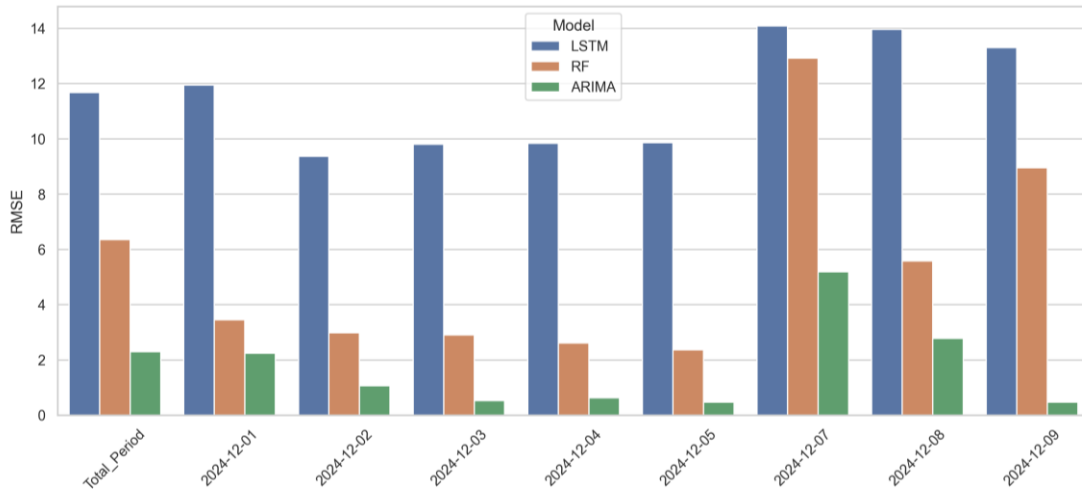


Fig. 6 – RMSE for predicted CO<sub>2</sub> comparison

PM<sub>2.5</sub> prediction proved significantly more difficult. In Fig. 4, we see that ARIMA actually achieves a lower MAE than LSTM globally. This is because PM<sub>2.5</sub> is stochastic; ARIMA's reactive nature works well for stable periods, whereas LSTM tries to "learn" a pattern that doesn't exist, leading to poor R<sup>2</sup> scores Fig. 5.

### 4.3. Global Residual Analysis

The analysis of residuals – defined as the difference between the actual sensor observation ( $y_t$ ) and the model's prediction ( $\hat{y}_t$ )—is a fundamental diagnostic tool for evaluating the reliability of IAQ forecasting. A high-quality predictive model should exhibit residuals that are normally distributed around a mean of zero, indicating that the remaining errors are "white noise" and do not contain systematic bias.

#### 4.3.1 CO<sub>2</sub> Residual Distribution Characteristics

As illustrated in Figure 7, the three models exhibit markedly different error profiles for Carbon Dioxide, LSTM and RF Precision. The residual curves for the LSTM (red) and Random Forest (blue) are characterized by extremely narrow, high-density peaks centered precisely at zero. This confirms that for approximately 95% of the validation period, the prediction error remains within a  $\pm 10$  ppm range. The symmetry of these curves proves that these models do not consistently over-predict or under-predict metabolic spikes, making them highly reliable for demand control ventilation system integration.

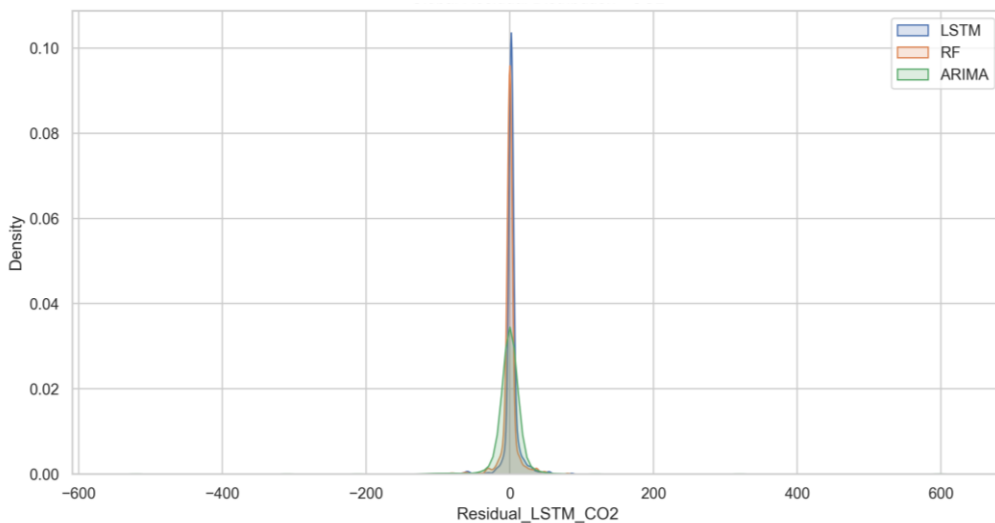


Fig. 7 – Global residual distribution for CO<sub>2</sub> prediction

**ARIMA Spread:** In contrast, the ARIMA (green) residual distribution is significantly flatter and wider. This spread is a direct mathematical consequence of the -15 min phase lag identified in the cross-correlation analysis. Because the ARIMA model is reactive, it generates large positive residuals during morning arrival spikes (under-prediction) and large negative residuals during evening departures (over-prediction), leading to the high global RMSE of 36.84 ppm.

#### 4.3.2 $PM_{2.5}$ Residual Distribution and Stochastic Artifacts

The residual analysis for Fine Particulate Matter, shown in Figure 8, reveals the structural challenges of forecasting non-periodic pollutants.

The LSTM residual curve (red) for  $PM_{2.5}$  is significantly broader than its  $CO_2$  counterpart. While it remains centered near zero, there is a visible "tail" extending toward positive values, indicating that the model frequently fails to capture the full magnitude of sudden dust or aerosol spikes.

The global distribution for LSTM is also influenced by the mathematical artifacts discussed in Section 6.1. With 438 instances of negative predicted values reaching as low as **-13.33  $\mu g/m^3$** , the residuals during low-activity periods are artificially inflated, contributing to the poor global  **$R^2$  of -1.16**.

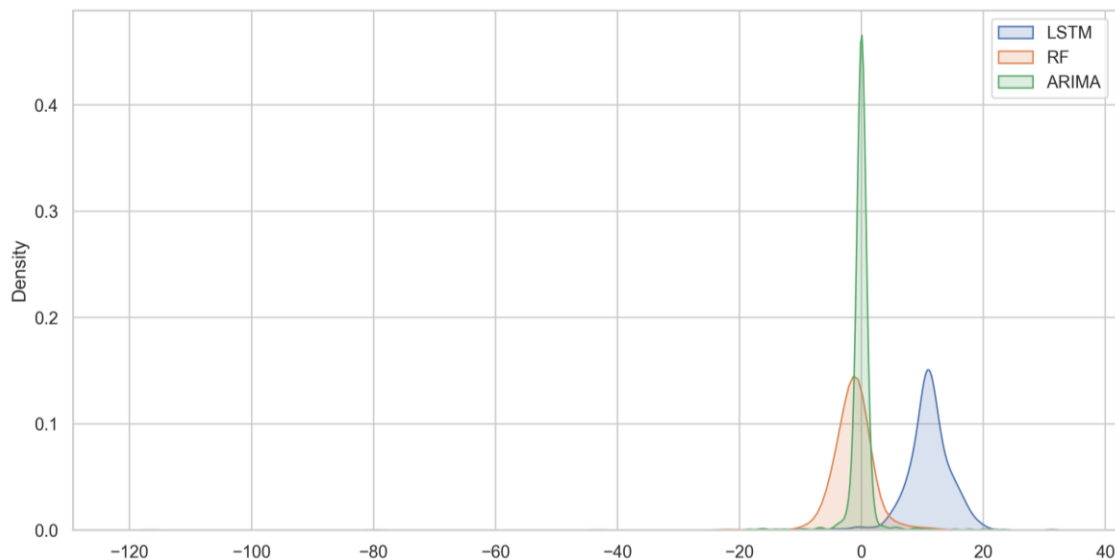


Fig. 8 – Global residual distribution for  $PM_{2.5}$  prediction

The **Random Forest** and **ARIMA** models show much tighter residual groupings for  $PM_{2.5}$ . The ARIMA model, in particular, maintains a very sharp distribution at zero, confirming that for stochastic IAQ events, a "persistence-based" or reactive approach results in higher statistical precision than trend-based deep learning.

## 5. Conclusions

The RMSE analysis demonstrates that for **predictive pre-control** of ventilation, the **LSTM model** is the most robust choice for  $CO_2$  management, as it avoids the massive squared errors associated with reactive lag. However, the high RMSE of the LSTM in  $PM_{2.5}$  prediction—compounded by the negative value artifacts suggests that for particulate matter, building managers should rely on reactive thresholding or the ARIMA baseline to ensure immediate safety during high-concentration events.

The "tightness" of the  $CO_2$  residual curves for LSTM and RF provides the scientific justification for **Feed-Forward Control**. Because the errors are low and non-

systematic, a BMS controller can trust the forecast to adjust fan speeds 15 minutes before an arrival peak occurs. However, the wide residual spread for PM<sub>2.5</sub> suggests that predictive control for particulates could lead to "ghost" ventilation events or a failure to respond to hazardous spikes, necessitating a hybrid logic where PM<sub>2.5</sub> management remains governed by real-time thresholding.

Current research increasingly points toward Hybrid Models. For instance, an ARIMA-LSTM approach uses ARIMA to handle the predictable linear cycles of a building while the LSTM captures the non-linear, unpredictable anomalies [18]. This combination provides the most robust defense against indoor air pollution, ensuring that HVAC systems respond not just to what is happening now, but to what is scientifically likely to happen next.

Despite the proliferation of IAQ studies, there remains a gap in head-to-head comparisons of deep learning versus ensemble and statistical models specifically for multi-target pollutants (CO<sub>2</sub> and PM<sub>2.5</sub>) in real-time control scenarios. Furthermore, a new real-world study is proposed, by using high-fidelity multi-pollutant data from sensors like Sensirion SEN66 sensor - providing data input for live predictive control to test the possibility to reduce or eliminate the phase lag inherent in traditional BMS systems, by using proactive ventilation.

*This work was supported by a National Research Grants of the UTCB, project number 1395 / 07.02.2025.*

## References

- [1] World Health Organization. *WHO guidelines for indoor air quality: selected pollutants*; WHO Regional Office for Europe: Copenhagen, Denmark, 2010.
- [2] Kalaiarasan, G.; Kumar, P.; Tomson, M.; Zavala-Reyes, J.C.; Porter, A.E.; Young, G.; Sephton, M.A.; Abubakar-Waziri, H.; Pain, C.C.; Adcock, I.M.; et al. Particle number size distribution in three different microenvironments of London. *Atmosphere* 2024, *15* (1), 45.
- [3] <https://doi.org/10.3390/atmos15010045>
- [4] Vijay, P.; Anand, A.; Singh, N.; Schikowski, T.; Phuleria, H.C. Examining the spatial and temporal variations in the indoor gaseous, PM<sub>2.5</sub>, BC concentrations in urban homes in India. *Atmos. Environ.* 2024, *319*, 120287.
- [5] <https://doi.org/10.1016/j.atmosenv.2023.120287>
- [6] Wei, W.; Ramalho, O.; Derbez, M.; Riberon, J.; Kirchner, S.; Mandin, C. Applicability and relevance of six indoor air quality indexes. *Build. Environ.* 2016, *109*, 42–49.
- [7] <https://doi.org/10.1016/j.buildenv.2016.09.008>
- [8] Vasile, V.; Iordache, V.; Radu, V.M.; Dragomir, C.S. The relationship between mechanical ventilation, indoor air quality classes, and energy classes in a Romanian context. *Atmosphere* 2024, *15*, 444.
- [9] Zhao, L.; Liu, J.; Ren, J. Impact of various ventilation modes on IAQ and energy consumption in Chinese dwellings: First long-term monitoring study in Tianjin, China. *Build. Environ.* 2018, *143*, 99–106.
- [10] <https://doi.org/10.1016/j.buildenv.2018.06.057>
- [11] Ganesh, H.S.; Fritz, H.E.; Edgar, T.F.; Novoselac, A.; Baldea, M. A model-based dynamic optimization strategy for control of indoor air pollutants. *Energy Build.* 2019, *195*, 168–179.
- [12] <https://doi.org/10.1016/j.enbuild.2019.04.022>
- [13] Ministry of Development, Public Works and Administration. *Energy Performance Calculation Methodology of Buildings, Code MC001-2022*; MDLPA: Bucharest, Romania, 2022.
- [14] Ministry of Development, Public Works and Administration. *Regulation for the Design, Execution and Operation of Ventilation and Air Conditioning Installations, Indicativ I5-2022*; MDLPA: Bucharest, Romania, 2022.
- [15] Kaggle. 1-year indoor environmental dataset (office).
- [16] Available online: <https://www.kaggle.com> (accessed on 20 March 2026).
- [17] IBM. What are ARIMA models?

- [18] Available online: <https://www.ibm.com/topics/arima-models> (accessed on 20 March 2026).
- [19] IBM. What is random forest?
- [20] Available online: <https://www.ibm.com/topics/random-forest> (accessed on 20 March 2026).
- [21] Kallio, J.; Tervonen, J.; Räsänen, P.; Mäkynen, R.; Koivusaari, J.; Peltola, J. Forecasting office indoor CO<sub>2</sub> concentration using machine learning with a one-year dataset. *Build. Environ.* 2021, 187, 107409.
- [22] <https://doi.org/10.1016/j.buildenv.2020.107409>
- [23] IBM. What is long short-term memory (LSTM)?
- [24] Available online: <https://www.ibm.com/topics/lstm> (accessed on 20 March 2026).
- [25] Gabriel, M.; Auer, T. LSTM Deep Learning Models for Virtual Sensing of Indoor Air Pollutants: A Feasible Alternative to Physical Sensors. *Buildings* 2023, 13, 1684.
- [26] <https://doi.org/10.3390/buildings13071684>
- [27] Tagliabue, L.C.; Re Cecconi, F.; Rinaldi, S.; Ciribini, A.L.C. Data driven indoor air quality prediction in educational facilities based on IoT network. *Energy Build.* 2021, 236, 110782.
- [28] <https://doi.org/10.1016/j.enbuild.2021.110782>
- [29] Dai, H.; Wu, N.; Dong, Z.; Ren, J.; Gao, Y.; Zhao, B. Comparison and evaluation of machine learning models for predicting indoor PM<sub>2.5</sub> concentrations on a large spatiotemporal scale. *Build. Simul.* 2025, 18, 1453–1466.
- [30] <https://doi.org/10.1007/s12273-025-1276-0>
- [31] Duan, J.; Gong, Y.; Luo, J.; Zhao, Z. Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. *Sci. Rep.* 2023, 13, 12127.
- [32] <https://doi.org/10.1038/s41598-023-36620-4>