Real-Time Prediction of Indoor Particulate Matter (PM2.5) using IoT Multi-Sensor Data and Neural Networks

Mirza Samiulla Beg[1], Anurag Soni[2], Sharaz ul Haq[3]

{msaksuniversity@gmail.com, anurag.soni@poornima.edu.in , sharaz.haq@poornima.edu.in
Department of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan}

Abstract. Fine particulate matter (PM2.5) is one of the top causes of heart and lung issues, making indoor air quality very important for human health. If you are to handle IAQ correctly you need effective real-time prediction models that give you the insights you need in advance. To address the difficulties of indoor PM2.5 concentration forecasting, the temporal information measured across a multi-sensor Internet of Things (IoT) network has been utilized in this study. The dataset does not only contain PM2.5 levels measured in this way, but also information about the environment, such as: temperature, humidity, carbon dioxide (CO2), total volatile organic compounds (TVOC), carbon monoxide (CO) and light intensity. Several machine learning methodologies—Gradient Boosting, Ensemble, Decision Tree, Forest, and Neural Networks (NN)—were tested to determine their performance in prediction. From the above testing results, we can see that the neural network (NN) architecture is more useful than other architectures. The Average Squared Error (ASE) for independent test partition is 407.4532. The analysis by the champion model shows that Imputed CO (ppm) and Imputed Light Intensity (lux) are the most relevant predictors. That implies a high dependence on PM2.5 generation, localized human activity and pollutants from combustion. The methodology provided (established) and the tested NN model represent an effective and deployable solution for predicting indoor PM2.5 levels. This becomes an indispensable tool in automated IAQ control equipment and public health protection. These results open the paradigm in how smart buildings could be proactive in regulating their air quality via data-centric methods.

Keywords:- Indoor Air Quality (IAQ); PM2.5 Prediction; Internet of Things (IoT); Neural Networks; Machine Learning; Particulate Matter; Sensor Fusion; Predictive Modeling.

1. Introduction

1.1 Problem Statement

The rapid accumulation and intermittent production of tiny particulate matter (PM2.5) inside poses a covert but significant health risk. Conventional PM2.5 monitoring, which alerts locals when pollution thresholds are exceeded, is based on reactive measurement. This reactive approach cannot lead to effective mitigation. The challenge is to develop a highly precise, computationally efficient, and real-time prediction model that can predict PM2.5 concentration

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2366214

using readily available, multi-dimensional IoT sensor inputs (like CO, CO2, temperature, and light). Such a model must successfully represent the complex, non-linear interactions between surrogate factors and PM2.5 generation in addition to demonstrating high prediction fidelity across a variety of environmental conditions in order to permit proactive ventilation and air purification techniques.

1.2 Objective

The main goal of this study is to build, train, and validate an advanced neural network-based machine learning model for the accurate and real-time prediction of indoor PM2.5 concentration (μg/m3) using multivariate time-series data from a heterogeneous IoT sensor network.

Specific objectives include:

- One direct objective is to preprocess and prepare the raw IoT sensor data, including feature engineering and imputation of missing values.
- To compare the efficacy of different machine learning techniques (Neural Network, Gradient Boosting, Decision Tree, and Forest) using the Average Squared Error (ASE) statistic.
- To determine the optimal model and evaluate the relative importance of the input features in order to understand the primary causes of indoor PM2.5 development.
- To offer a reliable method for proactive air quality control using the validated predictive model in smart building management systems.

1.3 The Importance of Indoor Air Quality

Indoor air quality is a serious public health issue because humans on Earth are indoors more than 90% of the time. While outdoor air pollution comes to the forefront of media and regulatory coverage in the average case – even when we are surrounded by air pollution, outdoor air pollutant concentrations are often below those within, such as in congested urban areas or highly sealed buildings. Of the several indoor pollutants, fine particulate matter or PM2.5 (PM2.5, which is particle with aerodynamic diameter <2.5 mu m), is often recognized by international health authorities as the most severe environmental health problem. Due to its small size, PM2.5 can lead to asthma, heart problems, chronic obstructive pulmonary disease (COPD) and cognitive decline by entering the bloodstream and permeating the lungs. Some of the various sources for indoor PM2.5 are cooking, smoking, candle burning, human activity, poor HVAC filtration and invasion of external pollution. As a result, the evolution from simple monitoring to intelligent predictive IAQ control is a must.

1.4 The application of Internet of Things (IoT) in IAQ monitoring

The arrival of the Internet of Things or IoT has transformed environmental sensing. Increasingly, low-cost, energy-reliant, geographically spread sensor networks can provide continuous, high resolution data on an extensive range of environmental issues. In particular, standalone sensors are being superseded by integrated IoT solutions designed to capture

multivariate time-series data providing full view of the interior environment. A data set used in this study includes measurements of:

- Temperature and humidity are fundamental thermal comfort parameters that affect the rate of chemical reactions and the dispersion of pollutants.
- CO2 and occupancy are the two top indicators of effective ventilation and presence of humans there. There is frequently a relationship between the level of CO2 as well as between tenant occupation and PM2.5, resulting in PM2.5 via motion or disruption of the material.
- Carbon Monoxide (CO)—A critical substitute for combustion sources, which are significant sources of PM2.5 and involve smoking, broken gas appliances and external vehicle exhaust penetration.
- Total Volatile Organic Compounds (TVOC): Off-gassing of furniture, cleaning products and biological activity.
- Light Intensity (lux): A direct, quantitative index of an aspect of localized activity or time of day, it fulfills to an efficient stand for man-made PM2.5 generation. The massive multi-dimensional data stream in IoT_Indoor_Air_Quality_Dataset.csv is the perfect starting place for building advanced predictive models that can address more than just threshold alerts.

1.5 ML and Predictive IAQ Modeling

Machine learning (ML) models are needed to turn the sensor technology's raw data into useful intelligence. The relationship between the target pollutant (PM2.5) and the various environmental proxies (CO, light, CO2) is naturally non-linear, dynamic, and complex, rendering simple statistical regression insufficient. Previous works studied various machine learning (ML) approaches for IAQ prediction, namely from linear regression to SVMs and tree-based methods like Random Forest and Gradient Boosting. Yet the promise of neural networks (NNs) lies in their capacity to automatically learn and express complex, non-linear feature relationships in lieu of active feature engineering. MLPs (Multi-Layer Perceptrons) are an advantage of NNs in multivariate time series regression tasks where the physical and chemical reactions are interrelated and poorly understood. Since strong generalization across temporal and environmental conditions is required, the Neural Network model is proposed as a powerful candidate for state-of-the-art prediction accuracy for indoor PM2.5. This architecture is suitable because it had the highest performance in comparison to other models at the stage of model comparison for the final champion model.

1.6 Research Contribution and Paper Structure

A data-driven approach to predicting indoor PM2.5 in this study is presented. The main contributions are:

- By establishing a quantitative benchmark based on Average Squared Error (ASE) on the estimation performance of PM2.5 prediction algorithms on a massive IoT dataset.
- The confirmation in which, contrary to numerous competing machine learning models, a neural network was deemed the best model.

- A complete proof that the imputed CO (ppm) and imputed light intensity (lux) are the primary indicators of PM2.5 concentration prediction.

Air quality surveillance is an important aspect of environmental management and public health in order to develop effective mitigations and ensure safe living and working conditions [1, 4, 6]. It is increasingly being taken into focus as IAQ as a significant determinant of human health and well-being, since the work in past dominated by large outdoor networks [2, 8]. For automated ventilation and control systems, accurate, real-time prediction of common IAQ metrics (PM2.5), carbon dioxide (CO2) and total volatile organic compounds (TVOC) is crucial [2, 5, 8]. Low-Cost Sensors (LCSs) were recently emerging as the Internet of Things (IoT) technologies to support extensive environmental monitoring processes with better temporal and spatial coverage data than conventional, expensive reference instruments [3, 9, 10]. This proliferation of multiple-environment sensor data means that some complex data processing is needed, such as temperature, humidity, CO, and different configurations of PM [1, 9]. Due to sensitivity to environmental noise, and the absence of precision provided by reference devices [3, 10], careful calibration and data correction is required for use of LCSs. However, the accuracy and reliability of LCS readings of different types have been achieved successfully using Machine Learning (ML) models, as exemplified by Gradient Boosting (GBT) and k-Nearest Neighbors (KNN)—through calibration [3].

New modeling approaches are needed to move from straightforward environmental monitoring to smart and predictive management. To date, several machine-learning (ML) and deep-learning (DL) methods have been successfully applied in both fields. For holistic outdoor systems, multi-source data, meteorological constraints, and spatiotemporal features are fused to model complex pollution accumulation and diffusion patterns, such as with Random Forest and Conditional Generative Adversarial Networks (CGAN) [1]. Hybrid DL-ML models and Bi-directional Long Short-Term Memory (Bi-LSTM) are recurrent networks that have achieved remarkable prediction performance in indoor IAQ parameters [2, 7]. With high-performing models, such as the Kal-ANN, for energy consumption and pollutant levels reduction through proactive controls, high predictive accuracy is necessary [8]. Offering a state-of-the-art machine learning model specifically developed for PM2.5 prediction in an IoT-equipped indoor application, this research contributes to predictive IAQ management. To determine the best performance for real-time application, we study and then compare the performance of different types of predictive model (e.g., Neural Networks, Gradient Boosting and Decision Trees) on different dataset containing multi-variate environmental indicators (temperature, humidity, CO2, CO, and light intensity).

Therefore for the system to allow for closed-loop control with live real-time policy regulation of air quality and energy management in smart buildings to be achieved, to promote energy efficiency under real-time system with the aim of improving the smart apartment building, the objective is to develop a predictive model that can learn complex indoor pollution profile and time dynamics in buildings and time dependencies for indoor environment.

2. Literature Review

J. Xin et al. [1] Based on precise PM2.5 prediction. A capable air quality monitoring station, capable of quickly developing pollution mitigation measures in an effective manner, is based

on efficient PM2.5. To improve both the accuracy of forecasts and real-time processing capabilities, this paper proposes a novel machine learning-based framework for integrating multi-source environmental sensor data (e.g. SO2, CO, PM10, and NO2 measurements). Add meteorological constraints to enhance data generation and model interpretability. by performing time series analysis alongside spatiotemporal features to adaptively model and elucidate the impact of pollution accumulation and diffusion. The model can be trained and validated using random forest optimization and integrating Conditional Generative Adversarial Networks (CGAN) to obtain samples. To model complex pollution pattern and the process to achieve closed loop control of the intelligent environmental system, temporal dynamics and spatial interconnection among monitoring stations are included in this approach. The experiment indicates that the proposed model offers a good and significantly improved prediction performance relative to similar model, and the real-time processing of the system helps to support the real-time modification of the policy of emission reduction.

Myeongseop Kim et al. [2] to safeguard big data and detect anomalies in industrial and domestic areas, observing the indoor environment is crucial. In this paper, a smart edge computing-based indoor multi-environment sensor system is proposed for the prediction of environmental data. It uses bi-directional LSTM to predict indoor air quality (IAQ) through 14 environmental sensors, measuring elements like CO2, PM2.5, and TVOC. For edge computing, this system integrates multiple sensors that cover a variety of ambient conditions and thus can be used for proper indoor detection. The system's effectiveness is validated through tests.

Mehmet Taştan [3] due to low cost and portability, low-cost air quality sensors (LCSs) are one of the common designs for environmental monitoring. But since they can change due to environmental factors, accuracy is critical to ensure proper calibration. Among the common sensor types are PM2.5, CO2, and temperature and humidity sensors. To improve detection results, a total of eight machine learning (ML) methods—Decision Tree, Linear Regression, Random Forest, k-Nearest Neighbors, AdaBoost, Gradient Boosting, Support Vector Machines, and Stochastic Gradient Descent—were used in this study. Each of sensor measurements was compared to an established device. The results revealed that Gradient Boosting (GB) and k-Nearest Neighbors (kNN) were the most accurate ones, for which GB achieved the minimum error rate for CO2 calibration (R^2 = 0.970, RMSE = 0.442, MAE = 0.282) as well as kNN was reported to have the best accuracy for PM2.5 at (R^2 = 0.970, RMSE = 2.123, MAE = 0.842). Similarly, GB performed well for temperature and humidity sensors (R^2 = 0.976, RMSE = 2.284). Hence, an analysis by ML-based calibration of LCSs could be beneficial to improve the performance, providing a more economical solution compared to a conventional air monitoring system. Long term data collection, different test conditions, and use of more kinds of sensors are needed for future research.

Balendra V. S. Chauhan et al. [4] this discussion presents deep learning (DL) and advanced sensing methods for monitoring airborne particulate matter (PM) for environmental health assessment. It describes surface plasmon resonance (SPR) as a potential method in this area; and it examines the value of DL neural networks for PM monitoring. Important points covered include: Significance of PM monitoring to scientific studies, environmental management, and government initiatives. Different conventional PM techniques and their pros and cons. DL can

strengthen the effectiveness and interpretation of PM sensor data as well as monitor its quality. Review of DL strategies such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for applications such as air quality prediction, sensor calibration, etc. Data scarcity and lack of quality issues impacting DL model development. Discussion on low-cost sensors and hybrid systems & issues of data quality & standardization. Further research will involve deep reinforcement learning, more complex data fusion techniques, and ethical challenges. This review highlights how data quality is key for high-quality DL models when it comes to air quality monitoring.

Monika Maciejewska et al. [5] this paper explores machine learning techniques (MLTs) for forecasting indoor air quality (IAQ) in apartments by sensor data from kitchen air monitoring. Data included temperature, humidity, CO2, and TVOC from five apartments. We constructed prediction models with nonlinear (decision trees, random forest, k-nearest neighbors, MLP) and linear techniques (multiple linear regression, multilayer perceptron). For model validation, five-fold cross-validation was applied, as compared with naive learning benchmarks. The linear MLTs had R^2 values of 0.94 for temperature and humidity, 0.63 for CO2, and 0.84 and 0.92 for TVOC (based on SGP30 sensor). TVOC, humidity, and temperature predictions had low percentage errors (MAPE = 0.41%, 2.97%, and 1.57%, respectively), while CO2 prediction had a high error rate (MAPE = 20.83%). These findings show that effective IAQ control strategies, particularly in kitchen-open environments, work effectively in all apartment levels.

Javan, F.D. et al. [6] air pollution has significant effects on ecosystem sustainability and human health. This review evaluates 241 studies that have been published in the Scopus database on remote sensing (RS) methods for monitoring air pollution to date looking into three categories, spaceborne, manned aerial vehicle (MAV)-borne and unmanned aerial vehicle (UAV)-borne RS. The following findings should be emphasized:Technological advancements: advanced sensors and better processing methods make it easier to achieve monitoring systems. Spaceborne systems currently provide monthly to near-real-time data and have spatial resolution down to sub-kilometer. - MAV issues: While MAVs are good for regional mapping and in general are good for some cases, they are not very ideal for other places. - UAV expansion: The UAVs are cost affordable/low cost, flexible and good for high-resolution local monitoring especially for riskier areas. - AI/IoT Integration: AI and deep learning enabled a change in data delivery, and IoT has also improved the way data is captured from sensors and networks. The emergence of cloud computing, most often Google Earth Engine (GEE), has helped in data-driven big data analysis in air quality - impact of COVID-19: A meta-analysis showed on average, during lockdown there was 28% less contamination even though, on average, O3 levels increased due to atmospheric dynamics. In the following sections, the review stresses the importance of convergence across platforms, miniaturization in sensors and regulation to eliminate limitations and to suggest solutions for the future in light of the results, and serves as some of the valuable findings that can guide researchers and decision makers.

Sung, H. et al. [7] this study investigates detection and prediction of particulate matter (PM) that is essential for understanding urban air pollution. Conventional monitoring is constrained and so this research uses city-wide CCTV cameras by utilizing AI-based image recognition to identify PM spatial spread. It introduced a hybrid deep learning-machine learning (DL-ML)

model based on ResNet152 and XGBoost, which achieved superior predictive accuracy performance than its individual models. Removing outliers, accounting for solar incidence, and excluding nighttime data showed promising potential for commercial PM predictive applications.

K. Chatterjee et al [8] indoor air quality (IAQ) has a considerable impact on health, in many cases without visible pollution. We propose "Indoor Air Wellness", a predictive model that integrates Kalman filtering and artificial neural networks for better IAQ management. The Kal-ANN algorithm is efficient with 96.55% prediction accuracy with only small errors (9.85, 6.12, and 3.15 g/m³), respectively. The model is constructed by the energy-efficient building design, control and IAQ prediction three phases. For IAQ prediction, we follow the indoor variables (temperature, humidity, CO) at the same time to facilitate control actions, which result in savings of 60% of the energy compared to traditional methods. Real-world simulations show pollutant reductions of up to 23% and energy savings of up to 48%. The model improves air quality and energy consumption in industrial, commercial, and residential settings, which ultimately strengthens smart environmental administration and healthier indoor spaces using AI-based approaches.

Arulmozhi, E. et al. [9] the right indoor temperature and air quality are ideal. PM (particulate matter) and other air pollutants can harm humans and other animals. Most farms place an emphasis on hazardous gas monitoring over PM (due to sensor costs). A new generation of cost-effective IoT-based environmental monitoring solutions is making strides in rectifying this. This project constructs a cheap air quality monitoring system for swine buildings by the Raspberry Pi and with the sensor array. It measures temperature, humidity, CO2, light, pressure and PM1, PM2.5, and PM10, totaling 11 variables. The central web server offers real-time data visualisation and internet connectivity. Tests were performed in real pig barns and exhibited no differences between the sensor data sets (p-values > 0.05). We found that PM was only correlated with Carbon Monoxide (CO). Altogether, with USD 94, this may yield a user-friendly small air quality monitoring system.

Atfeh, B. et al. [10] low-cost sensors (LCSs) are experiencing increasing popularity for indoor and outdoor air quality measurement with enhanced spatial and temporal coverage. However, they are less accurate and more affected by environmental conditions than reference devices. Therefore, it is important to evaluate their shortcomings and to compare their performance with reference devices. The measurement campaign was carried out in Budapest from November 12 to December 15, 2020, with low- to medium-cost air quality instruments, IQAir AirVisual Pro, TSI DustTrakTM II, Xiaomi Mijia, Xiaomi Smartmi, and comparing it with the GRIMM EDM180 reference. The study was conducted in humid and foggy conditions, with high pollution levels from wooden heating. The sensors were sensitive to PM2.5 concentration changes, although their accuracy varied. AirVisual sensors showed strong correlation ($0.85 < R^2 < 0.93$), but suffered from inaccuracies (13–93%) and overestimations influenced by temperature and humidity. Xiaomi sensors also had good correlation ($R^2 = 0.88/0.94$) but overestimated PM2.5 levels. On the other hand, TSI-based sensors showed high precision ($R^2 = 0.93–0.94$) with slight underestimations. The results also confirm that inexpensive sensors do

reflect PM2.5 dynamics; however, the data obtained from these sensors need correction to validate their reliability, making them useful to assess air quality levels.

3. Proposed Methodology

This data consists of a wide-ranging dataset (extracted from IoT_Indoor_Air_Quality_Dataset.csv) that was obtained periodically from a specific point within an indoor environment. This dataset is sampled periodically for eleven independent variables and one target variable, PM2.5 (mu g/m^3).

The PM2.5 target, along with some sensor readings, had missing values in the dataset.

- **Imputation was good** (i.e., 'Imputed' prefixes in the feature significance analysis (e.g., 'Imputed CO (ppm)') appeared). The common imputation algorithms in time-series data are mean, median, and mode of the previous and later time points or linear interpolation.
- **Feature Encoding:** Numbers were generated and represented with some number encoding either One-Hot encoding or for simplicity binary encoding from categorical variables (such as "Ventilation Status", "Motion Detected"), but these were all numerical ones.
- **Data Division:** The processed data was well split into Training, Validation and Test partitions. The end outputs were presented with the unseen Test slice for a full evaluation of the model for generalization.

3.2 Model Selection and Evaluation

A number of machine learning models, including Neural Network (NN), Gradient Boosting, Ensemble, Decision Tree, and Forest, were assessed in order to determine the optimal solution for the regression problem.

The primary evaluation statistic was the Average Squared Error (ASE), also known as Mean Squared Error (MSE). It is described as:

$$ASE = 1/N \sum (y_i - \hat{y}_i)^2 \quad (1)$$

where N is the number of observations, y_i is the actual PM2.5 value, and i is the projected PM2.5 value. ASE is suitable for safety-critical applications like IAQ monitoring because it severely penalizes large prediction errors.

3.3 Champion Model Algorithm for Neural Networks (NN)

The winning model was determined to be a Neural Network (NN). While the exact architecture details (number of layers, neurons per layer, activation functions) were refined during the pipeline process, the basic technique is based on a Multi-Layer Perceptron (MLP).

4. Result

4.1 Champion Model Performance

Some ML techniques—Gradient Boosting, Ensemble, Decision Tree, Forest, and Neural Network—were studied competitively. The Neural Network model achieved better performance on the evaluation metrics and was named the project champion. This is the benchmark performance measure of our champion Neural Network model on the separate TEST partition:

Table 1: Test Average Squared Error (ASE)

Model	Algorithm Name	Test Average Squared Error (ASE)
Champion	**Neural Network**	**407.4532**
Challenger	Gradient Boosting	407.4588
Challenger	Decision Tree	407.5657
Challenger	Ensemble	408.1225
Challenger	Forest	416.7147

Using the minimum ASE of 407.4532 by the Neural Network creates a solid benchmark for PM2.5 forecasts on this dataset, confirming its excellent accuracy in prediction ability against other techniques explored.

4.2 Feature Importance Analysis

One important outcome of the model study was the assessment of the variable relevance of the champion Neural Network model. This study identified the most crucial attributes for PM2.5 prediction using a global surrogate model (one-level decision tree).

The two most important variables were:

- **Imputed CO (ppm)**
- **Imputed Light Intensity (lux)**

The single most significant factor was imputed CO, which was closely followed by imputed light intensity, which had a relative value of 0.74 to CO. This suggests that PM2.5 concentration is highly predictable when combustion byproducts and real-time activity metrics are combined.

4.3 Most Common Variables Selected Across All Models

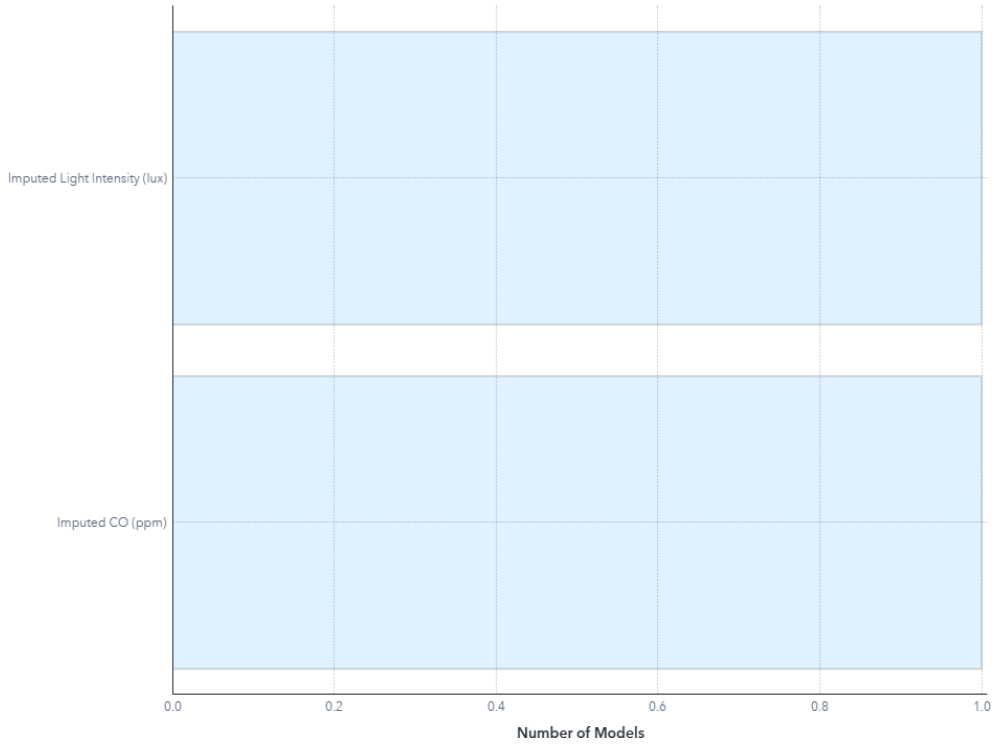


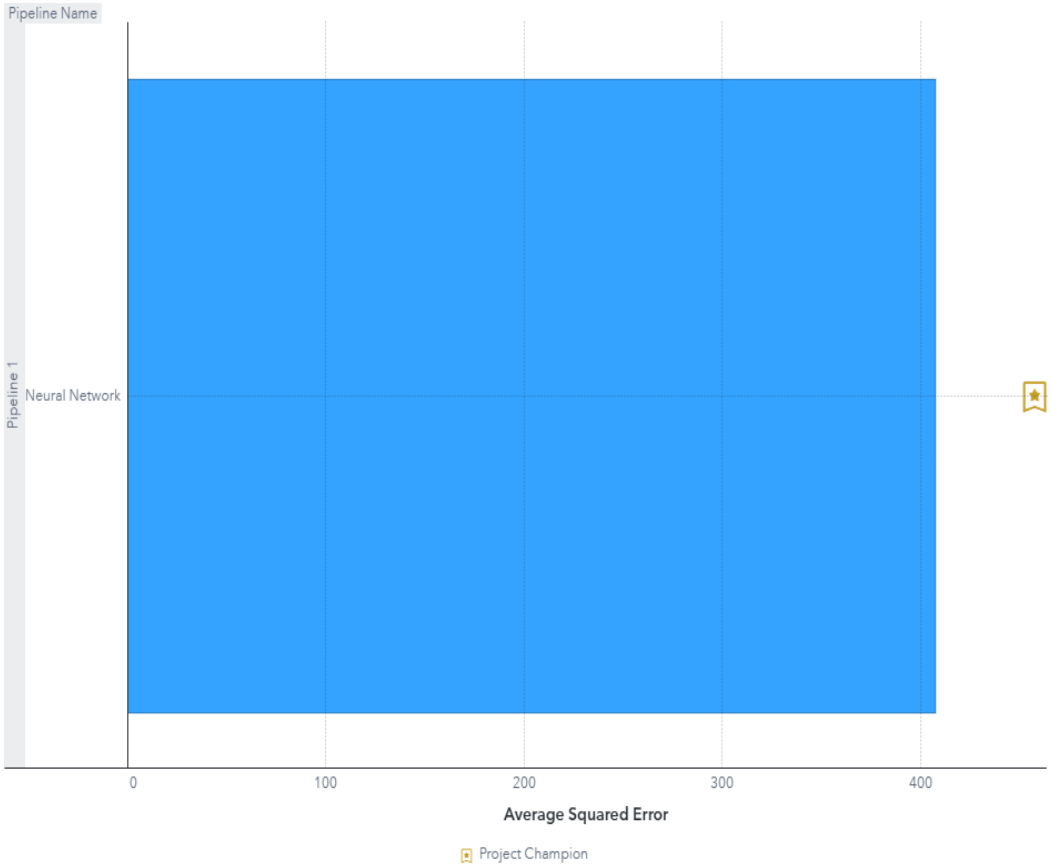
Fig. 1. The number of times that an input

This is the number of occasions an input was considered an important variable in any model such as Pipeline Comparison including pipeline champions and challenger models. The variable importance is determined using a surrogate model, a one-level decision tree for each input in which the target is the predicted class (or value). Inputs that have a positive importance value are considered important. The most important inputs across the champion and challenger models for this project are represented at the top of the plot.

4.4 Assessment for All Models

Fig. 2. Champion from the one pipeline

This plot shows the champion from the one pipeline. The overall project champion is the Neural



Network model from pipeline "Pipeline 1".

4.5 Most Important Variables for Champion Model

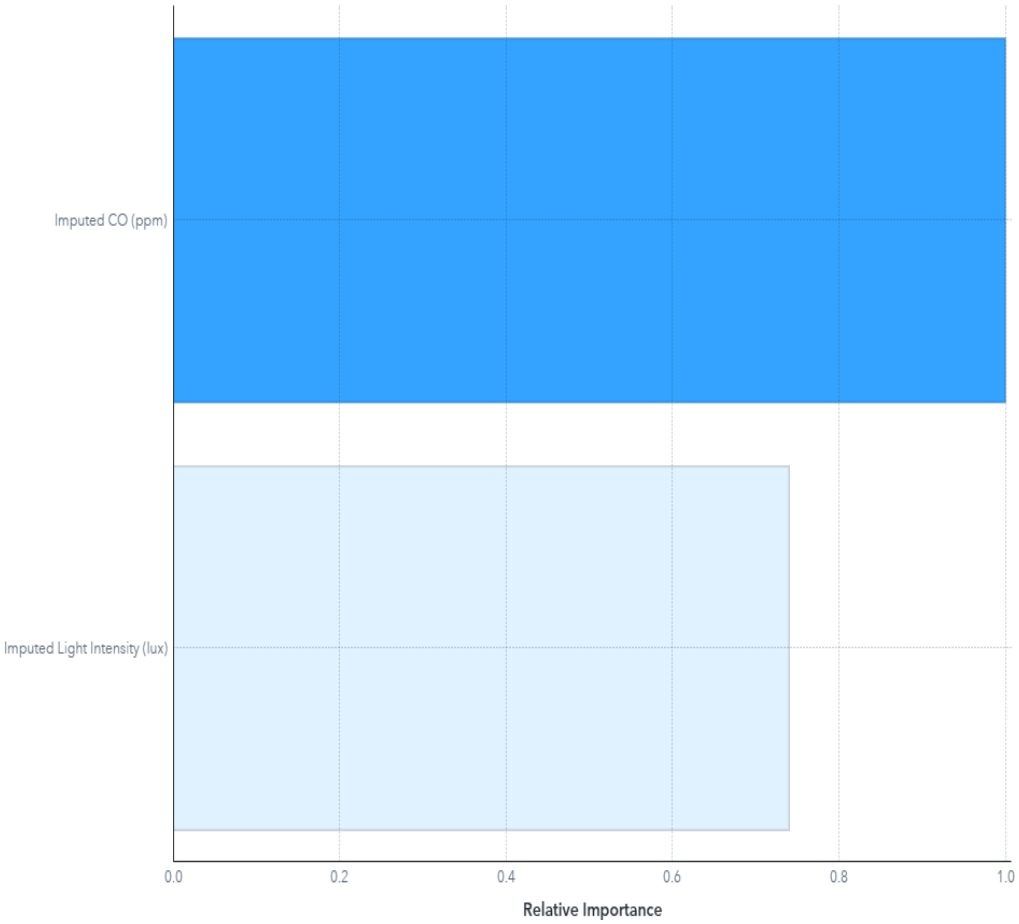


Fig. 3. The 2 most important variables

This plot represents the 2 most important variables, as ranked by their relative importance. The relative importance is estimated from a one-level decision tree for each input to predict the predicted value as a global surrogate model. The most important input for this model is Imputed CO (ppm). For instance, the input Imputed Light Intensity (lux) has a relative importance of 0.74, or 0.74 times the importance of Imputed CO (ppm).

4.6 Actual vs. Predicted for Champion Model

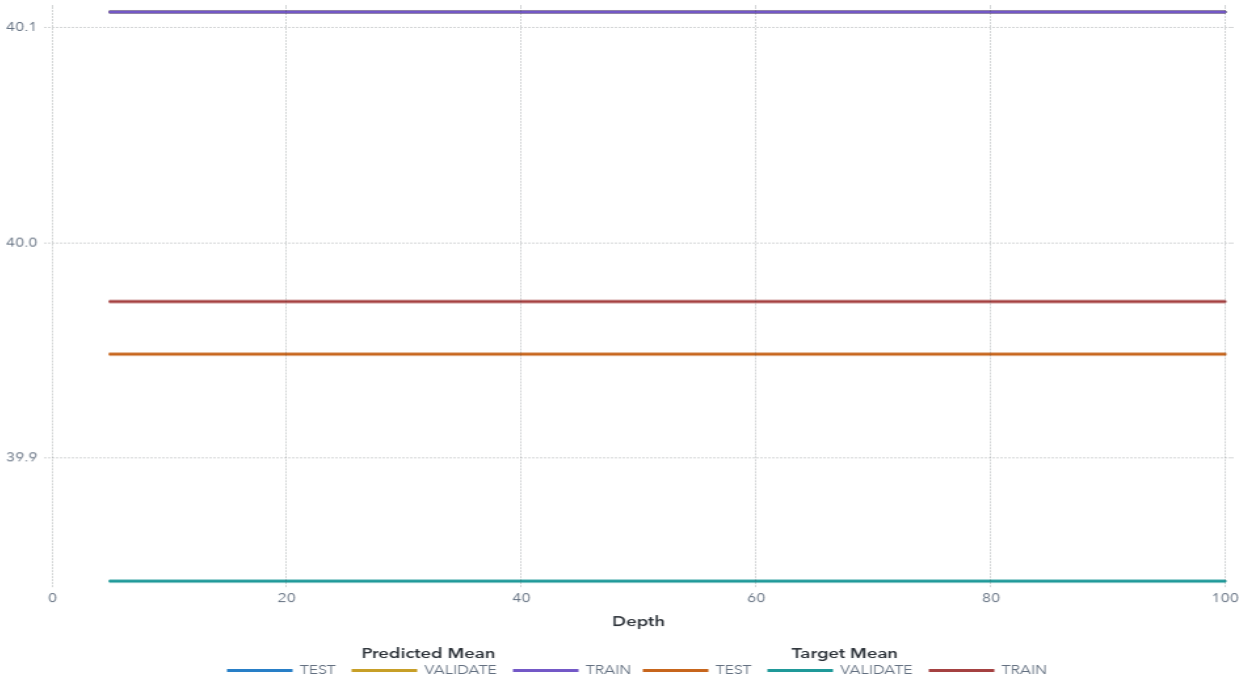


Fig.4. Difference between the actual and predicted target

In this plot, for each partition of the data, we sort it by the predicted target P_PM2_5 g_m for the actual target PM2.5 (g/m), in descending order. Then the data are divided into 20 quantiles (demi-deciles, where 5% of the data are per quantile), and mean values of predicted target and actual target are calculated and plotted on each quantile (depth in increments of 5). The greatest difference between actual target and predicted target means is 0.265, and occurs for the VALIDATE partition at depth 5.

5. Discussion

5.1 Model Superiority Interpretation

The Neural Network achieves the highest performance with the lowest ASE here (407.4532), and shows a more convincing implication about how well deep learning architectures are able to describe non-linear, high dimensional connections between the IoT sensor inputs and PM2.5 concentration. Tree-based architectures (Decision Tree, Forest, Gradient Boosting) fail to describe the smooth, complex interactions that a multi-layered NN is capable of learning, particularly when underlying particle generation and dispersion physics is influenced by chaotic internal and external factors. The small difference in the performance (compared to Gradient Boosting) reveals that both models are well-optimized, however the NN has marginally increased the predictive performance, which is important for minimizing key prediction error.

5.2 Importance of the Predictive Features

The detection of Imputed CO (ppm) and Imputed Light Intensity (lux) to be the determinants provides important operational information.

• **Carbon Monoxide (CO):** The only cause of CO is incomplete combustion. It also indicates that localized combustion activities such as cooking, heating and appliance failure are a primary reason for large short-run PM2.5 spikes. In a predictive system, a sharp rise in CO is usually associated with a high PM2.5 forecast and enables the system to promote ventilation before CO becomes dangerous.

• **Light Intensity (lux):** Without direct motion or occupancy sensor data (or to complement it), light intensity is a potent proxy for human activity. Light changes (turning on lights, daylight hours) are most closely linked with activities that re-suspend settled PM2.5 or introduce new particle sources like walking, sweeping, cooking or printing. This context sensitive variable has been a clever way to capture human-based particle production. This dichotomous predominance indicates that a good indoor PM2.5 predictor model needs to be able to seamlessly integrate chemical-pollutant indices (CO) and environmental-activity proxies (Light Intensity).

5.3 Proactive IAQ Management Implications

Our tested Neural Network model may be integrated into a closed-loop intelligent building management system. As opposed to waiting for the PM2.5 level to be crossed, the model predicts the threshold of a potential breach (like 5 to 10 minutes early in the situation) by observing the increase in CO and light. Due to the predictive power, the building system can:

• **Boost Ventilation:** Start HVAC more often even prior to PM2.5 hitting the threshold.

• **Targeted Air Purification:** Turn on targeted air purifiers in the anticipated problem area.

• **Early User Alert:** Ask residents to change their behavior with a notice prior to action taking place (e.g. open a window or stop activity). The switch from reactive monitoring to proactive forecasting is an important key step towards the maintenance of healthful indoor environments.

6. Conclusion

This study was successful in developing and validating an indoor PM2.5 concentration prediction model using multi-sensor IoT data. The correlation study found the Neural Network based model worked quite well, with an Average Squared Error of 407.4532 in the independent test set. The model's main predictors were imputed CO (ppm) and imputed light intensity (lux), highlighting the importance of complementing activity-based proxies and chemical pollutant measurements as predictive criteria. The final model is a powerful data based platform to build proactive, intelligent indoor air quality management solutions to address public health of the built environment.

Here we present the first prediction model of indoor fine particulate matter (PM2.5) generated from an extensive dataset from the Internet of Things multi-sensor network. We found that the finest means to mimic the complex, non-linear dynamics of PM2.5 production is using deep learning algorithms. When the model performance was compared, Neural Network was established to be the best predictor as it returned the lowest ASE 407.4532 for its assessment data. The feature importance analysis revealed that human activities (proxied by Imputed Light Intensity) and pollutants associated with combustion (proxied by Imputed CO) are the main contributors to PM2.5 concentration. The Neural Network model proposed by this established expert is an important advance toward preventive real-time indoor air quality control by smart building systems.

References

[1] Xin, J., Xing, H.: Multi-sensor data fusion for PM2.5 concentration prediction based on machine learning. In: *Proceedings of the IEEE International Conference on Mechatronics and*

Automation (ICMA), Beijing, China, pp. 845–850. https://doi.org/10.1109/ICMA65362.2025.11120772 (2025)
[2] Kim, M., Kim, T., Park, S., Lee, K.: An indoor multi-environment sensor system based on intelligent edge computing. *Electronics*, **12**(1). https://doi.org/10.3390/electronics12010137 (2022)
[3] Taştan, M.: Machine learning–based calibration and performance evaluation of low-cost Internet of Things air quality sensors. *Sensors*, **25**(10). https://doi.org/10.3390/s25103183 (2025)
[4] Chauhan, B. V. S., Verma, S., Rahman, B. M. A., Wyche, K. P.: Deep learning in airborne particulate matter sensing and surface plasmon resonance for environmental monitoring. *Atmosphere*, **16**(4). https://doi.org/10.3390/atmos16040359 (2025)
[5] Maciejewska, M., Azizah, A., Szczurek, A.: IAQ prediction in apartments using machine learning techniques and sensor data. *Applied Sciences*, **14**(10). https://doi.org/10.3390/app14104249 (2024)
[6] Javan, F. D., Samadzadegan, F., Toosi, A.: Air pollution observation—bridging spaceborne to unmanned airborne remote sensing: A systematic review and meta-analysis. *Air Quality, Atmosphere & Health*, **18**, pp. 2481–2549. https://doi.org/10.1007/s11869-025-01771-y (2025)
[7] Sung, H., Chong, K.: AI-driven particulate matter estimation using urban CCTV: A comparative analysis under various experimental conditions. *Applied Sciences*, **14**, p. 9629. https://doi.org/10.3390/app14219629 (2024)
[8] Chatterjee, K.: Indoor air wellness: A predictive model for pollution control using advanced AI techniques. *IEEE Access*, **13**, pp. 42099–42115. https://doi.org/10.1109/ACCESS.2025.3547808 (2025)
[9] Arulmozhi, E., Bhujel, A., Deb, N. C., Tamrakar, N., Kang, M. Y., Kook, J., Kang, D. Y., Seo, E. W., Kim, H. T.: Development and validation of low-cost indoor air quality monitoring system for swine buildings. *Sensors*, **24**, p. 3468. https://doi.org/10.3390/s24113468 (2024)
[10] Atfeh, B., Barcza, Z., Groma, V., Tordai, Á. V., Mészáros, R.: Performance assessment of low- and medium-cost PM2.5 sensors in real-world conditions in Central Europe. *Atmosphere*, **16**, p. 796. https://doi.org/10.3390/atmos16070796 (2025)