PREDICTING INDOOR $PM_{2.5}$ LEVELS USING DEEP LEARNING FOR ENHANCED DIGITAL TWIN APPLICATIONS

Juan Diego Castaño Molina[1], Zihao Zheng[1], and Borja García de Soto[1]
[1]S.M.A.R.T. Construction Research Group, Division of Engineering,
New York University Abu Dhabi (NYUAD), United Arab Emirates

Abstract

Monitoring and predicting indoor $PM_{2.5}$ levels is critical for ensuring healthy indoor environments. However, the integration of real-time data acquisition and instant $PM_{2.5}$ forecasting within digital twin systems remains underdeveloped in practical applications. This study presents a CNN-BiLSTM hybrid model for forecasting indoor $PM_{2.5}$ concentrations over a 72-hour horizon. In a real-world case study, the model achieved a root mean square error (RMSE) of 4.884 $\mu g/m^3$ and a Mean Absolute Error (MAE) of 4.092 $\mu g/m^3$. The integration of the model into a digital twin platform demonstrates its potential to enhance indoor air quality management through real-time, data-driven interventions.

Introduction

Indoor air quality monitoring and prediction have become increasingly critical as people reportedly spend 80% to 90% of their time indoors (Cai et al., 2023). $PM_{2.5}$ refers to fine particulate matter with a diameter of less than 2.5 micrometers, which is small enough to penetrate deep into the respiratory system and even enter the bloodstream. Prolonged exposure to these particles has been linked to a range of serious health issues, including respiratory conditions such as asthma and chronic bronchitis, as well as cardiovascular diseases such as heart attacks and strokes. These health risks make $PM_{2.5}$ a critical concern for air quality management and public health initiatives (Farahat, 2016). However, indoor $PM_{2.5}$ concentrations are influenced by multiple factors, including outdoor pollution levels, ventilation systems, and human activities (Syafei and Kurnianto, 2023). Accurately forecasting $PM_{2.5}$ requires capturing several interacting and time-varying factors, such as pollutant sources, ventilation dynamics, and human activities.

Machine learning techniques, including support vector machines (SVMs), random forests (RF), and artificial neural networks (ANNs) have been widely used in both outdoor and indoor air quality prediction (Dong et al., 2016; Shi et al., 2023; Xu et al., 2020). Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are also widely used in $PM_{2.5}$ prediction. CNNs excel at capturing spatial dependencies in time series data, while LSTM networks are adept at learning temporal dynamics (Putri et al., 2024). Bidirectional LSTM (BiLSTM) networks further enhance this capability by learning from past and future data points in a sequence (Sutskever et al., 2014). These advances in predictive modeling create opportunities to forecast indoor $PM_{2.5}$ and improve indoor environmental quality. However, as recommended by the U.S. Environmental Protection Agency (EPA), the Air Quality Index (AQI) for $PM_{2.5}$ is calculated based on a 24-hour average due to the long-term health risks posed by continuous exposure to $PM_{2.5}$ in the environment (U.S. Environmental Protection Agency, 2021). To this end, a long-term $PM_{2.5}$ prediction model, based on hourly reported $PM_{2.5}$ data, is necessary to capture AQI changes in the near future.

A Digital Twin is a virtual representation of physical environments that continuously updates with real-time data. It enables the analysis of current conditions, forecasting of future changes, and visualization of data in an accessible and interactive format. By providing actionable insights, a Digital Twin supports informed decision-making aimed at optimizing and improving the associated physical environments (Kanna et al., 2022). By integrating the $PM_{2.5}$ prediction model, which forecasts indoor air quality based on various environmental factors, the Digital Twin leverages real-time and historical datasets collected through IoT sensing platforms. This combination provides the potential to enhance the system's ability to monitor, analyze, and predict indoor air quality trends effectively (Opoku et al., 2024). Coupled with advanced visualization tools, these predictions can be presented in an intuitive and interactive manner, allowing occupants and decision-makers to easily interpret data and make informed decisions to enhance indoor air quality and occupant well-being (Gunatilaka et al., 2022).

This study develops a predictive model for indoor $PM_{2.5}$ concentrations using a hybrid CNN-BiLSTM architecture. Building on the Digital Twin platform previously proposed by the authors (Zheng et al., 2025), the integration of this model enhances the system's capabilities for dynamic and reliable indoor air quality monitoring and control. By incorporating multiple data streams, such as temperature, humidity, and outdoor $PM_{2.5}$ levels, the hybrid predictive framework delivers accurate 72-hour forecasts of indoor $PM_{2.5}$ concentrations, offering a robust tool for proactive air quality management.

Related Work

Indoor $PM_{2.5}$ Prediction

Indoor air quality (IAQ) management has garnered increasing attention in the post-pandemic era, with $PM_{2.5}$ being one of the most critical indoor pollutants. Various machine learning models have demonstrated their effectiveness in predicting indoor $PM_{2.5}$ levels, as shown in Table 1.

Table 1: Performance of Different Prediction Models for Indoor $PM_{2.5}$ Concentrations

Source	Forecast horizon (steps)	Models investigated	Best performing model*
(He et al., 2025)	1 (1 step = 5 min)	LSTM, GRU, CNN-LSTM, BiLSTM*	RMSE = 3.118 MAPE = 0.086
(Lee and Zhang, 2025)	1 (1 step = 2 min)	LSTM with Self-Attention, Bi-LSTM with Self-Attention*	MSE = 0.152 MAE = 0.224 RMSE = 0.389
(Minassian et al., 2025)	12 (1 step = 1 hour)	CNN, LSTM, CNN-LSTM*	MAPE = 0.0114
(Long et al., 2023)	24 (1 step = 1 hour)	DNN, LSTM, GRU, Transformer, Informer*	MSE = 0.017 MAE = 0.023

However, previous studies primarily emphasize model accuracy while lacking meaningful predictions that reflect air quality trends and provide actionable guidance for users. To the best of the authors' knowledge, a long-term indoor $PM_{2.5}$ prediction model (i.e., 72-hour forecast horizon) has not been evaluated using real-time datasets.

Digital Twin for Indoor Air Quality

Digital Twins are gaining emerging attention for indoor environmental quality (IEQ) management, offering significant potential to improve indoor air quality (IAQ) through the integration of advanced technologies, including the Internet of Things (IoT), platform services, data mining, and Building Information Modeling (BIM) (García de Soto et al., 2023). In particular, the proliferation of low-cost air quality sensors and off-the-shelf commercial devices has facilitated the monitoring and analysis of IAQ within IoT platforms. These platforms, in turn, target key pollutants highlighted in the World Health Organization (WHO) IAQ guidelines, such as particulate matter ($PM_{2.5}$, PM_{10}), CO, and others (Dai et al., 2023). Consequently, Digital Twin platforms can provide actionable guidance to occupants with data insights and intuitive visualization, helping them make informed decisions about their environment (Govindasamy et al., 2021). Altogether, this synergy enables advanced visualization, analysis, and decision-making for IAQ management (Opoku et al., 2024). This is particularly important since air quality issues, such as elevated $PM_{2.5}$ levels, are often imperceptible to human senses. For example, Hu et al. developed a Digital Twin-based platform for predictive IAQ maintenance that combines supervised and unsupervised learning models (Hu et al., 2023). Nonetheless, this model exhibits limitations in scalability across diverse indoor settings and lacks flexibility for applying to other air quality criteria due to the annotation of data.

While digital twin technology presents significant promise for IAQ prediction and management, models that generate actionable and intuitive insights on indoor air pollutants remain underexplored.. Additionally, the integration of real-time data acquisition and instant prediction within digital twin systems requires further proof-of-concept validation in real-world projects.

A Long-Term Forecasting Model for Indoor $PM_{2.5}$

The proposed model integrates the machine learning architectures of CNN and BiLSTM to generate 72-hour forecasts of $PM_{2.5}$ concentrations. Unlike single-step models, which predict only the next immediate value in a sequence, this model is designed for multi-step forecasting. This means that the model predicts a sequence of future values—in this case, the indoor $PM_{2.5}$ concentrations. After processing through the BiLSTM layers, the data is fed into a `Dense` layer with the number of neurons equal to the forecast horizon (in this case, 72 hours). The overall workflow is illustrated in Figure 1. It consists of two main components: data pre-processing and the ML workflow. The data pre-processing stage includes **Data Merging**, **Data Cleaning**, **Train-Test Splitting**, and **Scaling**. The ML workflow comprises the **CNN-BiLSTM Architecture**, **Training, Validation**, and **Evaluation**.

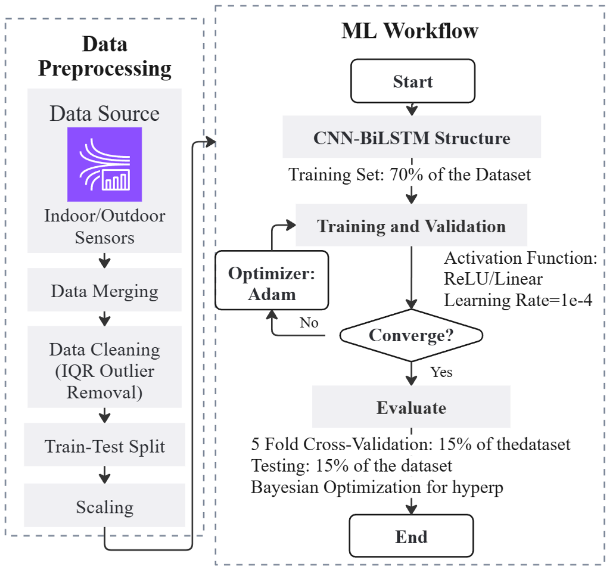

Figure 1: CNN-BiLSTM Training and Forecasting Pipeline

Data Pre-processing

Data Merging

During data pre-processing, critical parameters related to $PM_{2.5}$ are selected from the database and merged based on their timestamps. Indoor $PM_{2.5}$ levels and the corresponding environmental conditions (indoor $PM_{2.5}$, temperature, and humidity and outdoor $PM_{2.5}$) are to be investigated in this study. The merging process uses an inner join, ensuring that only records with matching timestamps in both datasets are included.

Data Cleaning (IQR-Based Outlier Removal)

An Interquartile Range (IQR) clipping method is used to exclude the outliers. The lower and upper bounds are defined in Equation (1a) and (1b).

$$\text{LowerBound} = Q_1 - 1.5 \times \text{IQR}, \quad (1a)$$

$$\text{UpperBound} = Q_3 + 1.5 \times \text{IQR}, \quad (1b)$$

$$\text{IQR} = Q_3 - Q_1 \quad (1c)$$

where Q_1 and Q_3 are the 25th and 75th percentiles of a feature, respectively.

Train–Test Split

After removing outliers, the data were split chronologically into training (first 70 %), validation (next 15 %), and test (final 15 %) sets. Preserving temporal order prevents the leakage of future information into the model during training and reflects real-world conditions, where models are trained on historical data and evaluated on previously unseen future observations.

Scaling

To prepare the data for model training, the features in the merged (and cleaned) dataset were normalized using the `MinMaxScaler` from the `sklearn` library. Scaling ensures that features are within a consistent range, typically between 0 and 1. This normalization process aids in faster model convergence during training and prevents features with larger numerical ranges from dominating the learning process.

ML Workflow

CNN-BiLSTM Structure

CNN Component. The CNN component applies two sequential 1-D convolutional layers—first with 32 filters, then with 64 filters (both kernel size = 2)—to the multivariate input sequence (indoor $PM_{2.5}$, temperature, humidity, outdoor $PM_{2.5}$), as shown in Figure 2 . Each convolution scans along the temporal axis to detect localized patterns. The output of the second convolution is then fed into a MaxPooling1D layer, which halves the sequence length, emphasizes the strongest activations, and reduces computational complexity.

BiLSTM Component. BiLSTM networks are the core of the model's ability to capture temporal dependencies in the data. Figure 3 shows a diagram of a BiLSTM network, illustrating the bidirectional flow of information,

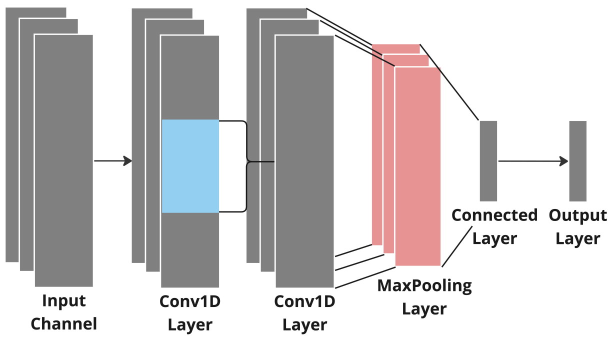

Figure 2: 1-D convolution and max-pooling workflow

which enables the model to capture dependencies from both past and future time steps in sequential data. Inputs $(X_{t-1}, X_t, \ldots, X_T)$ are sequential data points representing consecutive time steps. The forward LSTM processes data from past to future, and the backward LSTM from future to past, integrating context from both directions at each time step.

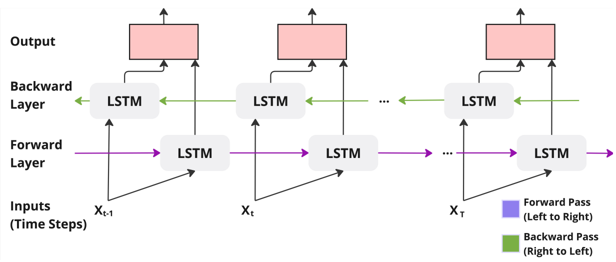

Figure 3: Architecture of the Bidirectional LSTM

Model Integration and Multi-Step Forecasting. Figure 4 illustrates the architecture of the hybrid CNN-BiLSTM model used for 72-hour indoor $PM_{2.5}$ forecasting. The input consists of a sequence of 48 time steps, each with four features: indoor $PM_{2.5}$, outdoor $PM_{2.5}$, humidity, and temperature. The model first applies two sequential 1D convolutional layers (with 32 and 64 filters, respectively) to extract short-term spatial features, followed by max pooling to reduce dimensionality. The resulting representation is passed through three stacked bidirectional LSTM layers with decreasing hidden units (256, 128, and 64), allowing the model to capture complex temporal dependencies at multiple levels of abstraction. This progressive reduction in the number of neurons serves as a form of compression, helping to mitigate overfitting while focusing the network's capacity on the most relevant patterns. The final dense layers (with 128 and 72 units) map the learned features to a 72-dimensional output, corresponding to the predicted $PM_{2.5}$ values for each hour of the next three days.

Training and Validation

The proposed CNN–BiLSTM model is trained using the `Adam` optimizer. Hyperparameters such as the learning rate, epochs, and dropout rate are automatically tuned via Bayesian optimization, allowing the model to converge ef-

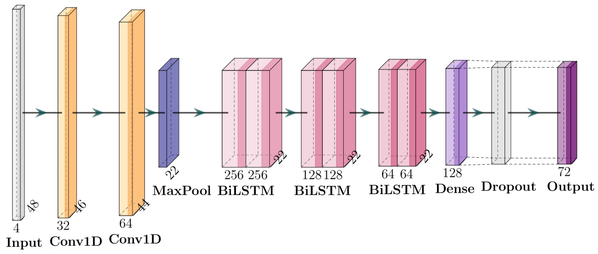

Figure 4: Architecture of the hybrid CNN-BiLSTM model for 72-hour indoor $PM_{2.5}$ forecasting

ficiently. The dataset is split chronologically, using 70% of the data for training while reserving 15% for validation and 15% for final testing.

Model performance is reported primarily via mean squared error (MSE) and mean absolute error (MAE), defined in Equations (2) and (3). During training, the model's performance is continuously monitored on the validation set, using MSE to guide hyperparameter updates and early stopping. Once the best hyperparameters are found, we retrain the model on the combined training and validation data and then evaluate it on the final 15% test set.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

We also apply 5-fold cross-validation to assess the model's generalization capabilities. In 5-fold cross-validation, the training data is divided into five subsets (folds).

Evaluation

After training and validation, the model undergoes a separate evaluation on the final 15% of unseen test data. This stage, distinct from validation, focuses on benchmarking by calculating root mean squared error (RMSE) alongside MAE. The RMSE is defined in Equation (4):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

Bayesian Optimization Setup. For each candidate set of hyperparameters, we train the CNN–BiLSTM model and evaluate it on the validation set. The Bayesian optimization algorithm seeks to minimize the validation MSE (or RMSE) as the objective function. Table 2 summarizes the primary hyperparameters and their respective search ranges.

At each iteration, the Bayesian optimizer selects a new point in the hyperparameter space based on the results of previous trials, thereby refining its understanding of how these parameters influence validation performance. This guided search often converges faster to an optimal or near-optimal combination of learning rate, batch size, number of epochs, and dropout rate than a traditional grid search.

Table 2: Hyperparameter search space for Bayesian optimization

Hyperparameter	Range / Values
Learning Rate	Continuous in [1e-4, 1e-2]
Batch Size	Discrete in 16, 32, 64, 128
Number of Epochs	Discrete in 10, 20, 30, 40, 50
Dropout Rate	Continuous in [0.0, 0.5]

Case Study

Prototype Framework of Digital Twin Integrated with CNN-BiLSTM

The proposed CNN-BiLSTM model is tested for indoor $PM_{2.5}$ prediction, emphasizing its integration into a digital twin application to ensure a healthy indoor environment. The digital twin platform structure, illustrated in Figure 5, includes **data acquisition** (indoor and outdoor environmental parameters), **correlation analysis and variable selection**, **model training and evaluation**, **visualization**, and **informed decision-making**. Within the digital twin framework, once the CNN-BiLSTM model is trained, it is deployed on the server as an inference model to forecast the future 72 hours $PM_{2.5}$ based on real-time readings.

Data Acquisition

In this study, data were collected using an IoT-based monitoring system. Indoor measurements were obtained from Sensirion S55 and BME280 sensors, each integrated with an ESP-WROOM-32 module for wireless data transmission to a local database via Wi-Fi. The sensors were put together by the authors and deployed in the S.M.A.R.T. Construction Research Group at NYU Abu Dhabi, serving as a testbed. The monitored space is located in the basement of the building and covers approximately 80 square meters. Ventilation is provided by a combination of fresh air units and a centralized HVAC system. The average readings from all sensors are used to represent the overall indoor air quality in the office. The layout of the room and the instrumentation of the indoor sensors are shown in Figure 6.

Outdoor $PM_{2.5}$ data were obtained from an IQAir AirVisual Outdoor monitoring sensor, accessed via an API provided by NYUAD's Center for Interacting Urban Networks (CITIES). The API is queried hourly, and the retrieved data are stored in a local database.

These datasets provide a robust foundation for the predictive modeling presented in this study and support further analysis of the relationship between indoor and outdoor air quality. The correlation between indoor and outdoor $PM_{2.5}$ concentrations is examined in the following section.

Correlation Analysis and Variable Selection

Hourly data were selected from the Digital Twin database, which has monitored the indoor environment since April 25, 2024. For this study, data collected up to January 18, 2025, were used. A total of 7,542 data rows of indoor $PM_{2.5}$, temperature, humidity, and outdoor $PM_{2.5}$ were obtained after merging and filtering. The selection

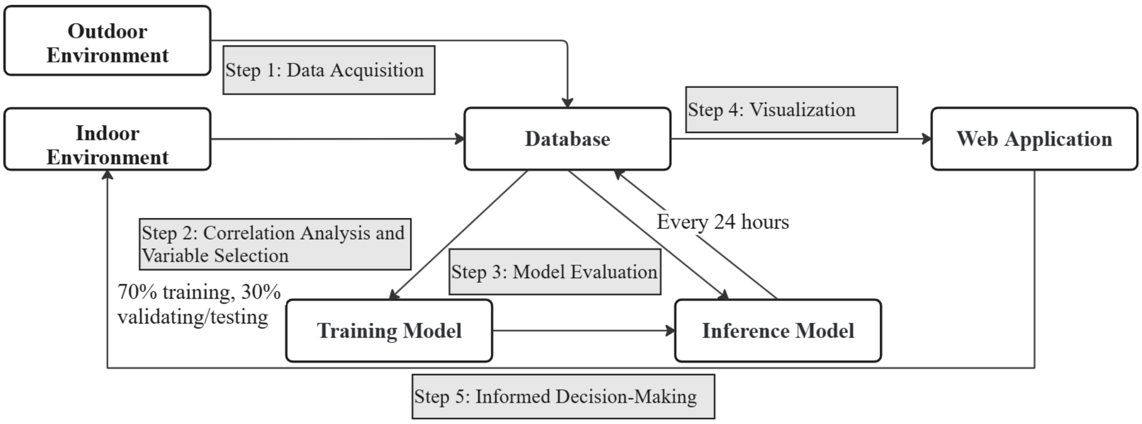

Figure 5: Overview of the process and key components of the proposed forecasting Digital Twin framework

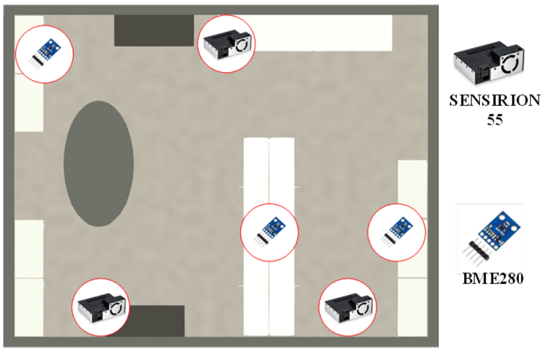

Figure 6: Room layout and indoor sensor instrumentation

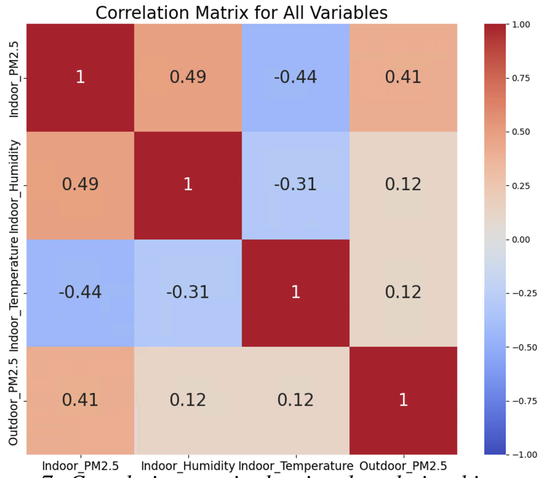

Figure 7: Correlation matrix showing the relationships among indoor $PM_{2.5}$, indoor humidity, indoor temperature, and outdoor $PM_{2.5}$

of variables for this study was guided by their correlations with indoor $PM_{2.5}$ levels, as shown in Figure 7. The correlation matrix reveals that outdoor $PM_{2.5}$ concentrations and indoor humidity have moderate to weak positive correlations with indoor $PM_{2.5}$, with correlation coefficients of 0.49 and 0.41, respectively. These relationships are plausible, as outdoor $PM_{2.5}$ could infiltrate indoor spaces through ventilation or leaks, while higher humidity levels may promote the hygroscopic growth of particulate matter, increasing both their size and concentration. It highlights the influence of outdoor air quality and indoor humidity on the indoor $PM_{2.5}$ level.

On the other hand, indoor temperature shows a negative correlation with indoor $PM_{2.5}$ (correlation coefficient of -0.44). This inverse relationship may arise because higher indoor temperatures are often associated with more active ventilation, which improves air filtration, thereby reducing $PM_{2.5}$ levels. As a result, indoor temperature is also included in the predictive model due to its indirect but significant role in influencing indoor air quality, particularly through its interaction with ventilation systems.

Model Evaluation

Hyperparameter Evaluation

The CNN-BiLSTM hybrid model was developed with Tensorflow and the results are evaluated using RMSE and MAE metrics for a 72-hour forecast horizon. Figure 8 shows how RMSE varies with batch size and look-back windows. The look-back values tested were 24, 48, 72, 120, and 240 hours, while batch sizes of 2, 16, 32, 64, and 128 were examined. Results indicate that medium look-back windows (48 and 72 hours) generally yield lower RMSE, particularly when combined with smaller batch sizes. However, RMSE tends to increase when the batch size exceeds 32. The optimal configuration for the dataset used was a look-back window of 48 hours and a batch size of 32.

Actual vs. Predicted $PM_{2.5}$ Concentrations

The selected CNN-BiLSTM model achieved a RMSE of 4.884 $\mu g/m^3$ and a MAE of 4.092 $\mu g/m^3$. These metrics indicate the model's strong predictive capability in estimating indoor $PM_{2.5}$ concentrations over a multi-step forecast horizon. Figure 9 shows a scatter plot of actual versus predicted indoor $PM_{2.5}$ concentrations. The diagonal line represents an ideal scenario where predictions match actual values perfectly. The clustering of data points around this line indicates a strong correlation.

Figure 10 presents the distribution of prediction errors,

Figure 8: Different Batch Size and Look-Back Values

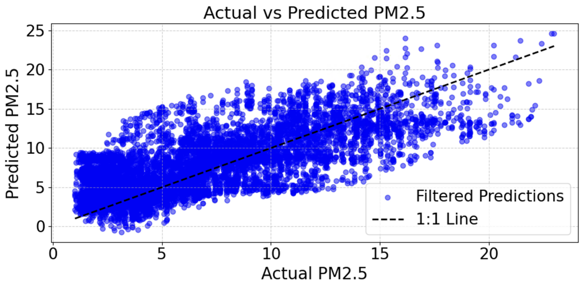

Figure 9: Actual vs Predicted $PM_{2.5}$

showing that most errors are near zero, indicating accurate predictions.

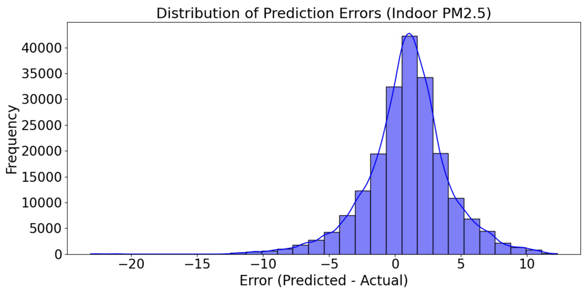

Figure 10: Distribution of Prediction Errors

Comparison with Other Models

To contextualize the performance of the proposed CNN-BiLSTM model, we compared it with other models previously used for forecasting indoor $PM_{2.5}$ concentrations, focusing specifically on those evaluated over similar forecast horizons. The results of this comparison are summarized in Table 3. The proposed CNN-BiLSTM model outperforms all other models considered, including RNN, LSTM, and various configurations of Dense LSTM and Convolutional LSTM.

Visualization

The 72-hour $PM_{2.5}$ forecasting model was deployed in our digital-twin platform, and the results are visualized in a web-based monitoring application (Digital-Twin Dashboard) (Castaño Molina et al., 2025).

Table 3: Performance Comparison of Deep Learning Models for $PM_{2.5}$ 72-hour Prediction

Model	**RMSE**	**MAE**
RNN	6.774	5.596
LSTM	7.900	6.067
Single Dense LSTM	8.409	6.667
Multi Dense LSTM	10.927	8.836
Convolutional LSTM	6.291	5.110
CNN-BiLSTM	**4.884**	**4.092**

As recommended by the U.S. Environmental Protection Agency (EPA), the Air Quality Index (AQI) for $PM_{2.5}$ is calculated based on a 24-hour average. The 24-hour average $PM_{2.5}$ concentration and the corresponding AQI scale is shown in Table 4.

Table 4: EPA AQI Guidelines and Corresponding Recommendations for 24-Hour $PM_{2.5}$ Averages

24-hour $PM_{2.5}$ Average Value ($\mu g/m^3$)	**USA AQI**	**Recommendation**
Low (0–9)	Good (0–50)	Air quality is satisfactory, no action is needed.
Moderate (9.1–35.4)	Moderate (51–100)	Air quality is acceptable. However, there may be a risk for some sensitive groups. Higher ventilation is recommended.
High (35.5–55.4)	Unhealthy for Sensitive Groups (101–150)	General groups may face risks. Masks are recommended for sensitive groups. Air purifiers should be turned on.
Very High (>55.5)	Unhealthy (>151)	Masks are recommended for all groups. Air purifiers should be turned on at the highest rate.

The forecasting model is scheduled to run at 12 am every day to provide a three-day $PM_{2.5}$ (and corresponding AQI) prediction alongside the outdoor AQI for comparison, as shown in Figure 11.

Informed Decision-Making

The digital twin platform provides occupants and decision-makers with real-time information on $PM_{2.5}$ concentrations, along with forecasted values, to support informed decision-making. For instance, when forecasts indicate Moderate Level (9.1–35.4 $\mu g/m^3$, AQI 51–100), facility managers can proactively adjust ventilation systems to increase airflow or inspect filtration systems as a preventive measure to mitigate indoor air pollution. In office environ-

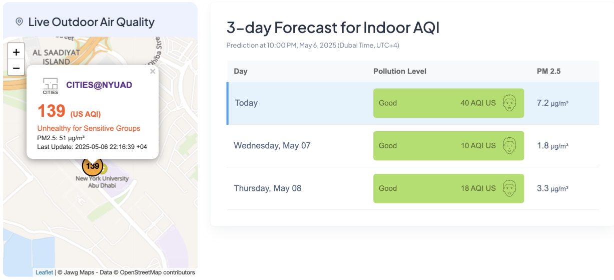

Figure 11: Web-based Implementation for the Air Quality Forecast Dashboard (Castaño Molina et al., 2025)

ments, occupants who are particularly sensitive to air quality can be notified in advance, enabling them to take personal precautions, such as wearing masks, using personal air purifiers, or opting to work remotely during periods of anticipated high $PM_{2.5}$ levels. The capacity to implement such proactive measures helps ensure a healthy indoor environment for all occupants. This is especially critical in buildings where $PM_{2.5}$ concentrations may fluctuate due to inadequate air circulation or inconsistent facility maintenance.

This predictive and integrated approach not only enhances occupant health and comfort but also contributes to greater operational efficiency. The extended 72-hour forecasting horizon provides facility managers with a forward-looking window to implement preventive measures and optimize building operations, enabling proactive maintenance of a healthy indoor environment based on predicted $PM_{2.5}$ concentrations or corresponding AQI levels.

Limitations and Future Work

Figure 12 presents the confusion matrix for the 72-hour forecasts against the observed indoor $PM_{2.5}$ levels in the test set. The model correctly classifies 55 314 of the 58 677 "Good" readings and 7 669 of the 21 387 "Moderate" readings, while misclassifying 3 363 "Good" as "Moderate" and 13 718 "Moderate" as "Good." No samples occur in the "Unhealthy for Sensitive Groups" (101–150 AQI) or "Unhealthy" (>150 AQI) categories, so the model's ability to predict those higher-risk regimes cannot be assessed. This absence of high-concentration examples limits its evaluation under rare but critical events (e.g., cooking peaks, outdoor infiltration). Future work should explore data-enrichment strategies, such as targeted sampling during pollution episodes, incorporating measurements from adjacent naturally ventilated spaces, or leveraging transfer learning, to bolster performance in those underrepresented AQI bands.

In addition, even within the Good and Moderate ranges, false positives or missed alarms may occur if unmodeled factors (e.g., occupant behaviors, HVAC system changes) shift conditions outside the training distribution. Incorporating additional sensors, building layout information, or seasonal variables could further enhance the model's robustness. Periodic retraining, via online learning or scheduled updates, would allow the model to adapt to evolving building usage patterns as captured by the digital twin.

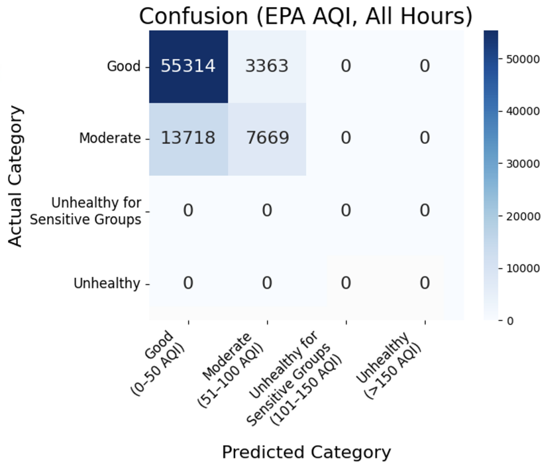

Figure 12: Confusion matrix (EPA AQI categories) for all 72-hour predictions.

Conclusion

This study presents a CNN-BiLSTM hybrid model for 72-hour indoor $PM_{2.5}$ forecasting, designed for integration with a digital twin platform. The model utilizes real-time sensor data, including indoor $PM_{2.5}$, temperature, humidity, and outdoor $PM_{2.5}$, which are preprocessed to exclude outliers. The combination of CNN layers for spatial feature extraction, and Bidirectional LSTM layers for temporal pattern recognition enables robust long-term forecasting. The best-performing model achieved strong results, with a RMSE of 4.884 $\mu g/m^3$ and a MAE of 4.092 $\mu g/m^3$, demonstrating improvements over existing deep learning approaches.

The significance of this research lies in its integration of IoT sensing, predictive modeling, and digital twin technologies. It advances the digital twin from a passive monitoring tool to a responsive, data-driven system capable of anticipating and reacting to real-world conditions. The deployment of the predictive model within a live-reporting web-based application further demonstrates its practical utility, with the potential to enhance indoor environmental quality and promote sustainable building operations.

Acknowledgments

This study was partially supported by the Center for Sand Hazards and Opportunities for Resilience, Energy, and Sustainability (SHORES), funded by Tamkeen under the NYUAD Research Institute Award CG013, and the Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001. The authors also acknowledge the support of the NYUAD Office of Student Research.

References

Cai, J., Chen, J., Hu, Y., Li, S., and He, Q. (2023). Digital twin for healthy indoor environment: A vision for the post-pandemic era. 10(2):300–318.

Castaño Molina, J., Zheng, Z., and García de Soto, B. (2025). Smart forecast. Available at: https://digitaltwin.smartconstructionresearch.com/app/forecast/ (Accessed: 7 May 2025).

Dai, X., Shang, W., Liu, J., Xue, M., and Wang, C. (2023). Achieving better indoor air quality with IoT systems for future buildings: Opportunities and challenges. Science of The Total Environment, 895:164858.

Dong, Y., Wang, H., Zhang, L., and Zhang, K. (2016). An improved model for PM2.5 inference based on support vector machine. In 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pages 27–31.

Farahat, A. M. (2016). Air pollution in the arabian peninsula (saudi arabia, the united arab emirates, kuwait, qatar, bahrain, and oman): causes, effects, and aerosol categorization. Arabian Journal of Geosciences, 9:1–17.

García de Soto, B., Sanchis, Z., Ezzeddine, A., Mengiste, E., Padilla, M., and Karmacharya, S. (2023). Developing a Digital Twin on a University Campus to Support Efficient and Sustainable Buildings. In Proceedings of the Creative Construction Conference 2023, pages 48–61, Keszthely, Hungary. Budapest University of Technology and Economics.

Govindasamy, H. S., Jayaraman, R., Taspinar, B., Lehner, D., and Wimmer, M. (2021). Air Quality Management: An Exemplar for Model-Driven Digital Twin Engineering. In 2021 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), pages 229–232.

Gunatilaka, D., Sanbundit, P., Puengchim, S., and Boontham, C. (2022). Airadar: A sensing platform for indoor air quality monitoring. 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), pages 1–6.

He, J., Zhang, S., Yu, M., Liang, Q., Cao, M., Xu, H., Liu, Z., and Liu, J. (2025). Predicting indoor PM2.5 levels in shared office using LSTM method. Journal of Building Engineering, page 112407.

Hu, W., Wang, X., Tan, K., and Cai, Y. (2023). Digital twin-enhanced predictive maintenance for indoor climate: A parallel LSTM-autoencoder failure prediction approach. Energy and Buildings, 301:113738.

Kanna, K., AIT Lachguer, K., and Yaagoubi, R. (2022). MyComfort: An integration of BIM-IoT-machine learning for optimizing indoor thermal comfort based on user experience. Energy and Buildings, 277:112547.

Lee, M. J. and Zhang, R. (2025). A Self-Attention Fusion-Based Bi-LSTM Framework for Occupant-Centric Prediction of Indoor Environmental Factors. In Proceedings of the 6th International Conference on Civil and Building Engineering Informatics, volume 8, page 11.

Long, H., Luo, J., Zhang, Y., Li, S., Xie, S., Ma, H., and Zhang, H. (2023). Revealing Long-Term Indoor Air Quality Prediction: An Intelligent Informer-Based Approach. Sensors, 23(18):8003. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

Minassian, R., Mihăiţă, A.-S., and Shirazi, A. (2025). Optimizing indoor environmental prediction in smart buildings: A comparative analysis of deep learning models. Energy and Buildings, 327:115086.

Opoku, D.-G. J., Perera, S., Osei-Kyei, R., Rashidi, M., Bamdad, K., and Famakinwa, T. (2024). Digital twin for indoor condition monitoring in living labs: University library case study. Automation in Construction, 157:105188.

Putri, T. H., Caraka, R. E., Toharudin, T., Kim, Y., Chen, R. C., Gio, P. U., Sakti, A. D., Pontoh, R. S., Pratiwi, I. R., Nugraha, F. A. L., Azzahra, T. S., Cerelia, J. J., Darmawan, G., Faidah, D. Y., and Pardamean, B. (2024). Fine-tuning of predictive models cnn-lstm and conv-lstm for nowcasting pm2.5 level. IEEE Access, 12:28988–29003.

Shi, Y., Du, Z., Zhang, J., Han, F., Chen, F., Wang, D., Liu, M., Zhang, H., Dong, C., and Sui, S. (2023). Construction and evaluation of hourly average indoor pm2.5 concentration prediction models based on multiple types of places. Frontiers in Public Health, 11.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. CoRR, abs/1409.3215.

Syafei, A. D. and Kurnianto, N. P. (2023). Study on peak hours, ventilation, and resident activities towards indoor air quality on pm2.5 in surabaya. IOP Conference Series: Earth and Environmental Science, 1263.

U.S. Environmental Protection Agency (2021). How is the aqi calculated? Accessed: 2023-01-20.

Xu, C., Xu, D., Liu, Z., Li, Y., Li, N., Chartier, R., Chang, J., Wang, Q., Wu, Y., and Li, N. (2020). Estimating hourly average indoor PM2.5 using the random forest approach in two megacities, China. Building and Environment, 180:107025.

Zheng, Z., Castaño Molina, J., Mengiste, E., Prieto, S., and García de Soto, B. (2025). Digital twins for healthier spaces: A scalable framework for monitoring indoor environmental quality. In ISPRS Geospatial Week 2025, Dubai, UAE.