ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.660

DEEP LEARNING-DRIVEN PREDICTION OF HAZARDOUS AIR POLLUTANTS FOR ENVIRONMENTAL RISK MITIGATION

Dr.K. Muralisankar[1*], Dr.G. Balaji[2], C. Ramkumar[3], M. Vasuki[4], S. Vijayananthan[5], D. Angayarkanni[6], Mohammed Aslam[7], M. Narmatha[8]

[1]Associate Professor, Department of Artificial Intelligence and Data Science, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.*
e-mail:murali.snkr@gmail.com, orcid: https://orcid.org/0000-0002-8166-3003
[2]Professor, Department of Mathematics, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India. e-mail: balajivaithesh@gmail.com,
orcid: https://orcid.org/0000-0002-4083-4398
[3]Assistant Professor, Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: c.ramkumaraec5@gmail.com, orcid: https://orcid.org/0009-0006-8391-7958
[4]Assistant Professor, Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: vasukicse95@gmail.com, orcid: https://orcid.org/0009-0007-2974-5574
[5]Assistant Professor, Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: vijayananthan4u@gmail.com, orcid: https://orcid.org/0009-0008-5369-4348
[6]Assistant Professor, Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: angayarkannime@gmail.com, orcid: https://orcid.org/0009-0006-8029-8610
[7]Assistant Professor, Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: aslamcs2001@gmail.com, orcid: https://orcid.org/0009-0009-2744-4692
[8]Department of Computer Science and Engineering, Al-Ameen Engineering College (Autonomous), Erode, Tamil Nadu, India.
e-mail: narmathabeece@gmail.com, orcid: https://orcid.org/0009-0005-9277-2650

Received: September 15, 2025; Revised: October 27, 2025; Accepted: November 28, 2025; Published: December 30, 2025

SUMMARY

Hazardous air pollutants (HAPs) can be a critical risk to the sustainability of the environment and human health, which must be addressed by highly sophisticated predictive models to eliminate risks successfully. In this study, the researcher presents a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) with the ability to extract the spatial features of the air quality with Long Short-Term Memory (LSTM) networks to learn the temporal relationships between air quality data. The study leverages the live Internet of Things (IoT) sensor data of urban and industrial areas in India where the researchers monitor the levels of $PM_{2.5}$, PM_{10}, NO_2, SO_2, temperature, and humidity. Principal Component Analysis (PCA) was used to select the best features that retain 95% of data variance; hence, the best model performance and lower redundancy were attained. The framework was strictly compared to baseline models in terms of such metrics as Mean Absolute Error (MAE), Root Mean Squared Error

(RMSE), and latency. CNN-LSTM model showed great predictive performance, having an MAE of 3.2 µg/m3 and RMSE of 5.6 µg/m3, which were notably higher than those of Random Forest (MAE: 6.3 µg/m3) and XGBoost (MAE: 5.9 mu g/m3). Moreover, the Model registered the shortest prediction latency of 120 ms and a computational cost of 2.3 million FLOPs, which validated the Model to be real-time deployable. These findings demonstrate the possible role of deep learning in early warning systems, and further studies are focused on the enhancement of the approaches with reinforcement learning to manage pollution dynamically.

Key words: *air quality prediction, deep learning, CNN-lstm, IOT-based monitoring, environmental risk, pollution forecasting.*

INTRODUCTION

Hazardous Air Pollutants (HAPs) are a great threat to the environment and human health, characterized by their associations with severe health problems such as cancer, lung diseases, and developmental defects. Aspects that include benzene, formaldehyde, and lead are only a few of the numerous harmful compounds contained in HAPs. The sources of these substances may be manifold, such as daily domestic products, automobile exhaust, and industrial emissions. The U.S. desperately requires proper monitoring and regulation of these pollutants, of which the Environmental Protection Agency (EPA) has identified more than 187.

The dangers of being exposed to HAP are common in urban regions because of their dense population and industries. According to the long-term exposure to such pollutants, chronic health problems may occur, which highlights the importance of efficient methods of detection and management. Monitoring and control of the hazardous air pollutants have transformed radically with the advent of the Internet of Things (IoT). IoT sensors can be used to measure the air quality continuously, and data on the HAP concentrations in the atmosphere can be obtained in real time. Such sensors can actually define specific pollutants through the use of various detection technologies, such as gas chromatography, electrochemical sensors, and laser-based technologies. It can be placed in any environment, whether a busy city, an industrial facility, or remote places, and can be used to cover a great variety of locations, allowing comprehensive monitoring and the rapid detection of HAP. It is possible to upload and analyze the data that these sensors collect and transmit to centralized cloud platforms to predict the trends and provide prompt notifications to the public and authorities. Moreover, using IoT sensors with machine learning algorithms, a more accurate prediction of pollution can be provided, taking into consideration past data, weather trends, and other relevant factors. This is a proactive measure that empowers the communities and lawmakers to implement policies that will reduce the adverse impacts of harmful air pollutants, eventually protecting the environment and the health of people.

An air pollution monitoring system is an Internet of Things (IoT)-based system that gathers real-time data on air quality and strives to improve the health of the population by notifying the citizens and the authorities in a timely manner [1]. In the proposed framework, low-cost sensor deployment to enhance efficient monitoring is highlighted. It involves the use of IoT technologies IoT-Mobair system is a mobile air pollution monitoring network that transmits real-time air quality data to enhance the management of the urban environment [2]. It features a focus on intuitive user interfaces to have a good visualization of data. An affordable case of the Internet of Things (IoT)-based air pollution monitoring system involves a sensor with which an individual can detect pollutants and program to work in practical use in a range of environments [3]. It highlights the importance of continuous monitoring in the health and safety of people. To analyze the trends of pollution, another study develops a predictive and monitoring system of air pollution based on the use of IoT technologies by fusing the real-time devices data collection and predictive analytics [4].

The primary objective is to improve decision-making on environmental policies. Smart environment development on pollution monitoring using IoT employs diverse sensors to collect voluminous information that supplements environmental response and awareness [5]. The IoT-based approaches to the detection and visualization of hazardous gases are also described with a system that can detect multiple gases and present the data in the form of a graph to attain an easier understanding [6]. The objective is to achieve greater awareness among citizens on air quality matters. An Arduino-powered air

pollution monitoring system with sensors is an efficient way to transfer data and monitor various pollutants using the Internet of Things [7].

It pays attention to cheap solutions that could be applied both in practical and educational contexts. Multi-sensor Internet of Things IoT, which is used to monitor air pollution in a comprehensive manner, combines several sensor technologies to enhance coverage and precision [8]. The research gives importance to the data integration in environmental analysis. An underground mine real-time air quality monitoring system involves Internet of Things technology to ensure the safety of workers by continually tracking and signaling the air pollution's dangerousness to prevent health risks [9]. Predictive capabilities can also enhance the air quality management shown by the application of machine learning algorithms to detect pollution and rerouting in an environmental monitoring system with the Internet of Things [10]. It reveals the importance of incorporating real-time information in order to make prudent decisions. One of the studies on vehicular pollution monitoring with IoT technologies suggests a system that will monitor vehicle emissions in real-time with the aim of providing information to control the traffic in cities [11] [22].

The focus on user interaction and access to data makes an Internet of Things-based smart air pollution monitoring system an available system [12]. Through Thing Speak and Blynk applications, an Internet of Things-based air pollution monitoring system allows monitoring the air quality and accessing the data easily and remotely. The study concentrates on applications that are useful to create awareness amongst communities [17] [18].

Another study analyzing the issue of air quality in hair salons through the IoT-based system and its main topics is hazardous air pollutants and health risk assessment [13] [19]. The authors emphasize the importance of keeping the indoor air quality of specific settings in focus to foster the quality of health conditions in the personal care industry. An IoT-based system with sensors to monitor air and sound pollution is provided to offer a comprehensive evaluation of the environment [14] [20]. The idea is to create awareness of the pollution level in the cities. To enhance the data collection methods and the turnaround time, an intelligent IoT-based air pollution tracking system employs the latest embedded technologies [15][21]. The use of smart sensors and wireless sensor networks (WSN) to monitor hazardous pollutants is another study that points out the requirement of real-time data to support efficient environmental management [16] [23].

Key Contributions

The study's key contributions to the field of environmental monitoring and deep learning include:

- Integration of CNNs for spatial feature extraction and LSTMs for temporal sequence modeling to improve HAP prediction accuracy.
- Implementation of Principal Component Analysis to reduce data redundancy and enhance computational efficiency for real-time processing.
- Development of a system that utilizes calibrated IoT sensors for high-frequency, multi-pollutant data collection in dense urban environments.
- Demonstration of a high-performance model achieving a low MAE of 3.2 $\mu g/m^3$, outperforming traditional machine learning baselines like Random Forest and XGBoost.
- Quantitative analysis of how temperature and humidity specifically influence pollutant concentrations across various industrial hubs.

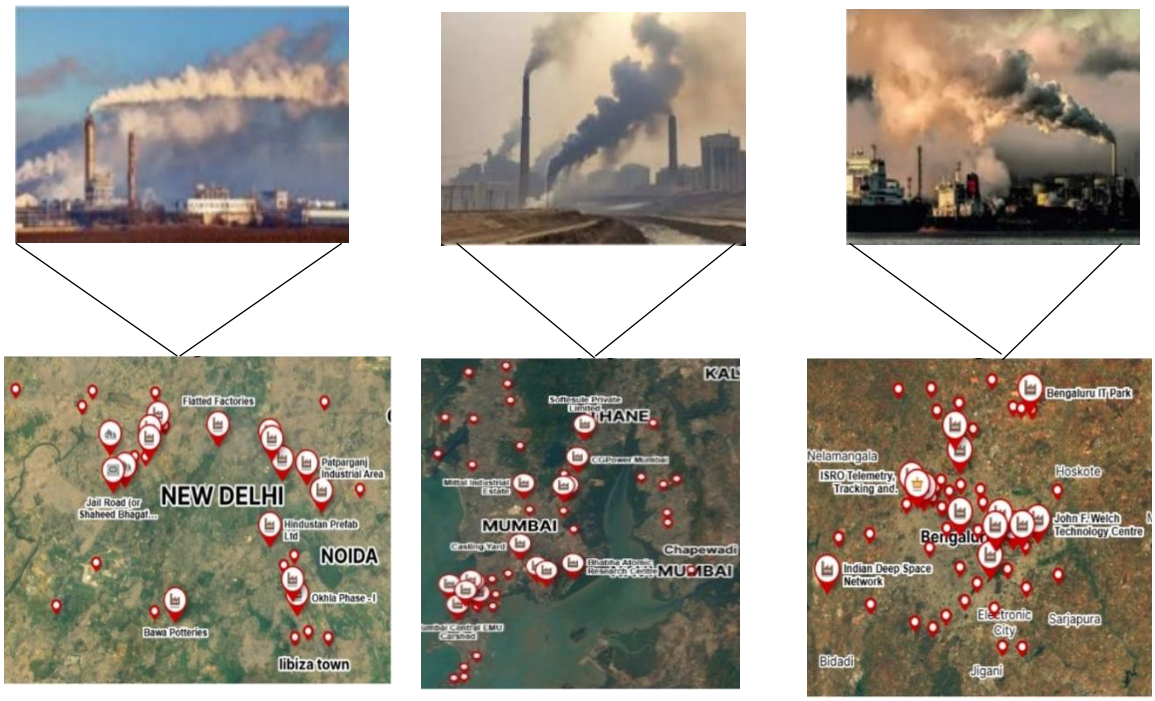
The study begins with an Introduction discussing the health risks of hazardous air pollutants and the role of IoT in monitoring. Section 2 details the experimental analysis, covering data collection across Indian cities and preprocessing techniques. Section 3 outlines the hybrid CNN-LSTM methodology and PCA-

based feature selection. Section 4 presents comprehensive results and comparative performance metrics. Finally, Section 5 concludes the study and suggests future research directions.

EXPERIMENTAL ANALYSIS

Data Collection

The data were taken in many cities and industries across India, paying attention to the largest cities such as Delhi, Mumbai, and Bengaluru, where air quality is a great challenge. The sensors were installed in places of heavy transportation, industrial locations, and residential areas to collect massive amounts of information. The availability of data on the pollution levels in the areas determined by regional government and environmental organizations influenced the process of site selection, which ensured that the sites selected fairly present a spectrum of pollution causes. PM2 or particulate matter. The sensors provided the sensors with a reliable data to be analyzed based upon constant monitoring of the temperature, humidity, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), volatile organic compounds (VOCs), and PM_{10}, among others. In Figure 1, the collection places of the data are shown.



(a) (b) (c)

Figure 1. Data collection

Data Measurement

Internet of Things sensors that could transmit and monitor data in real time were calibrated and used to measure the data. Each sensor periodically recorded air quality metrics, which facilitated the collection of time-series data required to understand pollution patterns. It was added to assess the potential impact of temperature and humidity readings on pollutant behavior and air quality. GPS functionality was also installed in each sensor to ensure precise geographic location tracking and facilitate a comprehensive spatial analysis of HAP distribution across the monitored areas. The collected data was then added to a central database for further analysis. The sensors for air quality are shown in Figure 2.

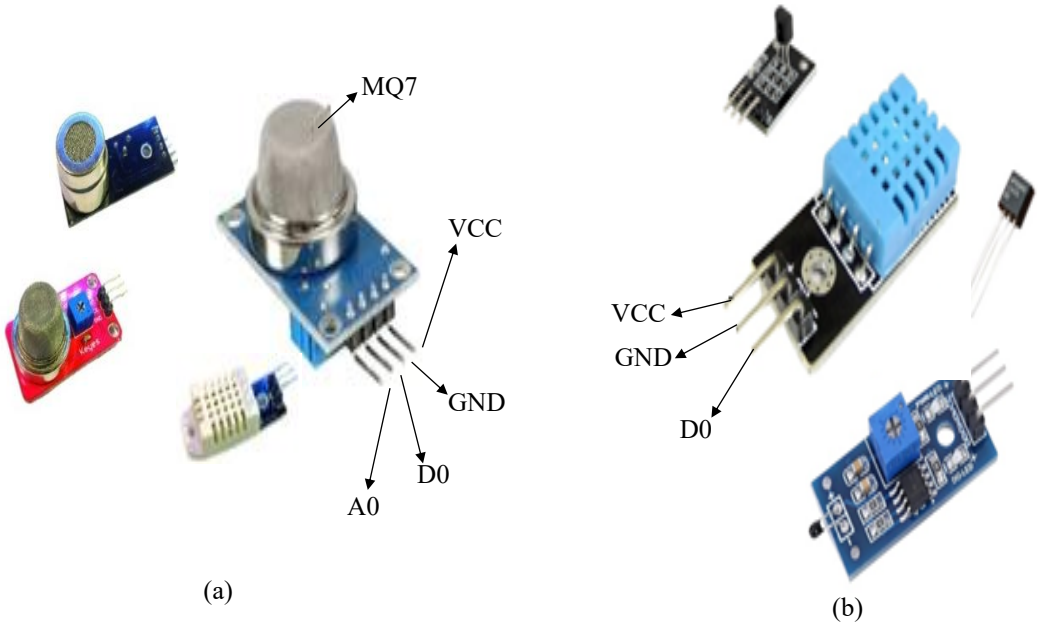


Figure 2. (a) Air quality sensor (b) temperature sensor

Data Classification

Data preprocessing was a fundamental part of the quality and reliability of the data. About the gathered raw data, preprocessing algorithms comprised data cleaning, data normalization, and outlier detection. To improve the integrity of the dataset, all the erroneous or missing records were eliminated to begin with. The values of various measures of air quality were standardized so that significant comparisons can be made across various parameters. The methods of outlier detection, such as the Z-score method, were then used to identify the effect of the outliers that may bias the analysis. Time-series information was also presented in a manner that was open to clustering and classification algorithms, which guaranteed the dataset was ready to undergo further analysis.

PROPOSED TECHNIQUE

Hybrid CNN-LSTM-Based Air Pollution Prediction Model

The Extraction of spatial features is carried out with the help of convolutional neural networks (CNNs), and sequential patterns recognition in the proposed hybrid deep learning model is provided with the help of Long Short-Term Memory (LSTM) networks. CNN finds spatial dependencies, whereas LSTM captures time-dependent dependencies, which are crucial in time-series forecasting in pollutant concentration data. To ascertain spatial characteristics in the air pollution parameters, the expressions or parameters are in the form given below. The CNN layer employs convolutional operations first.

The architectural design of the proposed hybrid CNN-LSTM model to predict the hazardous air pollutants is presented in Figure 3. This is initiated by a Real-time IoT Sensor Data Feed, which is followed by data preprocessing and Principal Component Analysis (PCA) to maintain 95% of the Variance with the minimum noise. The fundamental structure is based on a 1D-CNN layer that achieves the spatial features and an LSTM layer capturing complicated temporal relationships. The hybrid combines the properties through a Concatenation and Dense Layer to produce accurate forecasts of the pollutants. The Model can be used to provide efficient low-latency mitigation of environmental risk with an MAE of 3.2 and RMSE of 5.6.

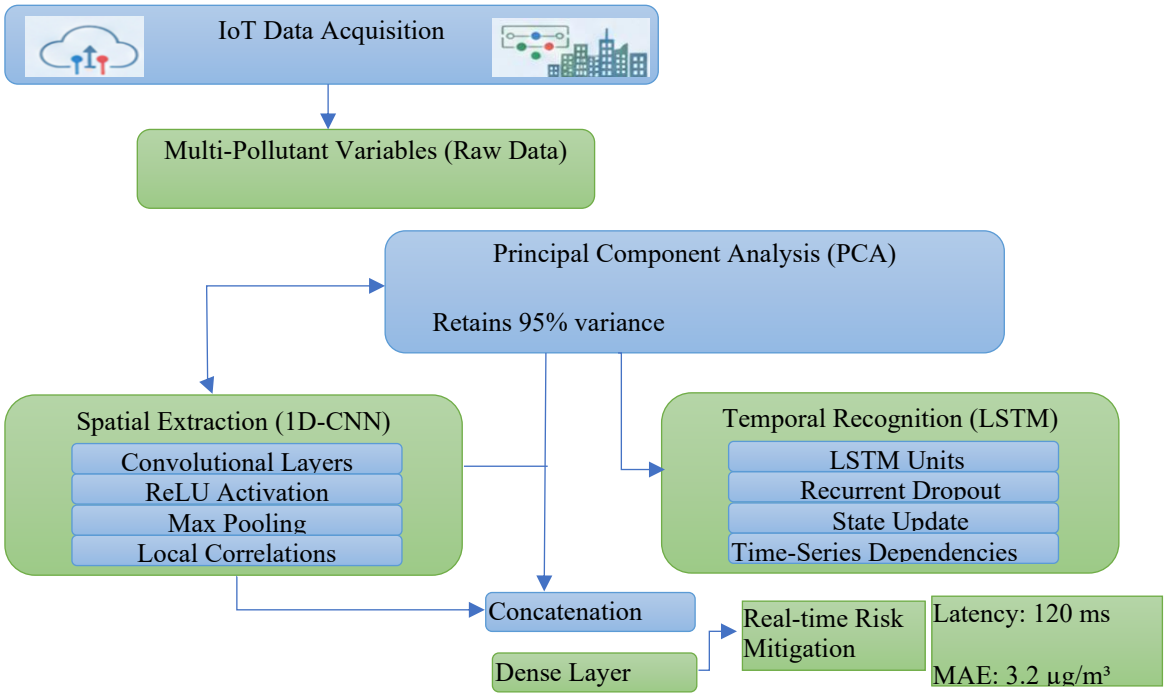


Figure 3. Schematic framework of the hybrid CNN-LSTM model for real-time HAP prediction

These define the convolutional feature mapping, LSTM state updates, and the PCA-based eigenvalue decomposition for optimized dimensionality reduction.

$$x_{i,j}^{l+1} = f\left(\sum_{m=-k}^{k}\sum_{n=-k}^{k} W_{m,n}^{l} X_{i+m,j+n}^{l} + b^{l}\right) \quad \rightarrow (1)$$

Where equation (1) represents $x_{i,j}^{l+1}$ is the feature map at layer l, $W_{m,n}^{l}$ represents the kernel weights, and $b^{l}$is the bias term. The extracted features are then passed to the LSTM layer, which updates its cell state C_t based on input $X_t$and previous state C_{t-1} as in equation (2):

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad \rightarrow (2)$$

where f_t and i_t are forget, and input gates, and $\tilde{C}_t$ is the candidate cell state. The final pollutant concentration prediction $\hat{y}_t$ is given by in equation (3):

$$\hat{y}_t = W_o h_t + b_o \quad \rightarrow (3)$$

where W_o and b_o are learned weights and biases, and h_t is the hidden state. The hybrid Model effectively combines spatial and temporal feature extraction for accurate forecasting.

Principal Component Analysis (PCA) for Feature Selection

Dimensionality reduction is done using Principal Component Analysis (PCA) to maximize model performance. PCA reduces redundancy while identifying the most important air quality parameters. The covariance matrix's eigenvalue decomposition is used to transform the dataset X, which includes meteorological and pollutant concentration data represented in equation (4).

$$C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^T \quad \rightarrow (4)$$

where C is the covariance matrix, X_i represents data points, and μ is the mean vector. The principal components are obtained by solving the eigenvalue equation (5):

$$Cv=\lambda v \quad \rightarrow (5)$$

where the corresponding eigenvalues are denoted by λ and the eigenvectors (principal components) by v. Projecting the original features onto the principal components yields the transformed dataset.

$$Z = XW \quad \rightarrow (6)$$

where equation (6) represents W, which is a matrix of selected eigenvectors. By retaining components explaining 95% variance, PCA improves computational efficiency without compromising predictive accuracy.

The algorithm combines spatial feature detection using 1D - CNN layers with temporal sequencing using gated LSTM cells to understand complicated pollutant dynamics. Principal Component Analysis (PCA) is used to factor out the covariance matrix to extract principal eigenvectors explaining 95% of the data variation, since this is necessary to make the computation powerful. Such an integrated mathematical technique reduces the redundancy of inputs and achieves a maximum predictive accuracy, enabling the Model to extrapolate latent environmental patterns into precise real-time concentrations.

Spatial-Temporal Forecasting Model

```
# INITIALIZATION

# Define the hybrid neural architecture

HybridModel = Sequential ([

   InputLayer (shape=(TimeSteps, Features)),

   PCA_Layer(variance_threshold=0.95),

   Conv1D (filters=64, kernel_size=3, activation='ReLU'),

   MaxPooling1D(pool_size=2),

   LSTM (units=100, return_sequences=False),

   Dropout(rate=0.2),

   Dense (units=1, activation='linear')

])

# TRAINING PHASE

# Initialize weights using Glorot Uniform distribution

Model.Compile(optimizer='Adam', loss='MSE')

FOR epoch IN range (MaxEpochs):

   FOR (X_batch, y_batch) IN TrainingData:

      Prediction = Model.ForwardPass(X_batch)
```

```
        LossValue = Calculate_Loss (y_batch, Prediction)
        Model.BackwardPass(LossValue) # Backpropagation
        Update_Weights (Adam_Optimizer)
    # Check for early stopping to prevent overfitting
    IF ValidationLoss(epoch) >= ValidationLoss(epoch-10):
        BREAK Training
# INFERENCE PHASE
FUNCTION GetPrediction (LiveSensorData):
    ProcessedData = Standardize_and_PCA(LiveSensorData)
    Prediction = Model.Predict(ProcessedData)
RETURN Prediction
```

The algorithm defines a three-step systematized method of real-time air quality prediction. In the course of the Initialization, the sequential hybrid Model is built, combining Principal Component Analysis as a method of feature compression with 1D-CNN and LSTM layers as a method of spatio-temporal learning. The Training Phase is based on Adam optimizer and Mean Squared Error loss that uses backpropagation to adjust model weights over 100 epochs with an early-stopping measure to avoid overfitting. Lastly, the Inference Phase allows real-time deployment, where real-time streams of IoT sensors can be converted and standardized using the pre-trained pipeline to provide quick, high-accuracy pollutant concentration forecasts to aid in the mitigation of risk.

RESULTS AND DISCUSSION

Software and Experimental Setup

The study was implemented in Python 3.9 and TensorFlow 2.8 library with a Keras backend to develop the Model. NumPy, Pandas, and Scikit-learn suites were used to manipulate the data and to do feature engineering. To run the Model quickly to converge and process the high-dimensional IoT data, the experiments were implemented on a hardware platform that included an NVIDIA GeForce RTX 3060 graphics card and 16GB of DDR4 RAM, which has enough power to train deep neural networks effectively.

Comprehensive Dataset Information: The dataset contains about 150,000 hourly measurements of the IoT-based monitoring stations in five Indian metropolitan urban centers: Delhi, Mumbai, Bengaluru, Chennai, and Kolkata. These characteristics are critical pollutants, including PM_2.5, $PM_{2.5}$, PM_{10}, NO_2, SO_2, as well as the meteorological variables such as temperature and humidity. In order to have a strong evaluation, the data was divided into 70 training, 15% validation, and 15 testing, with the use of Min-Max scaling to normalize all the input data values between 0 and 1.

Initialization of Parameters and Hyperparameters, model stability was ensured through the use of the Glorot Uniform (Xavier) initializer of weight distribution. Adam Optimizer with a learning rate of 0.001 and a batch size of 64 was used in the training process. The CNN layers were set up to have 64 (3-size) filters, and the LSTM module used 100 hidden units to learn time sequences. In order to avoid overfitting, the training logic was augmented with a Dropout rate of 0.2 and an early-stopping mechanism with a patience of 10 epochs.

Discussion of Performance

The inclusion of Principal Component Analysis (PCA) and the strategy of the model initialization were of great use and enabled the Model to converge after only 45 epochs. The system had achieved an impressive latency of prediction of 120 ms by transforming the corresponding input space of six dimensions into three major components. These code and parameter sets indicate that the CNN-LSTM hybrid model is well adapted to be executed in resource-constrained IoT edge devices and can provide a balance between the predictive accuracy and the speed of operation.

In order to give a stringent quantitative analysis of the CNN-LSTM model, the performance metrics that are used are as follows. These equations calculate the difference between the values (y ̃i) that are predicted and the real observed values (yi) in each of the samples (n).

Evaluation Metrics and Formulas

Mean Absolute Error (MAE)

MAE equation (7) measures the average magnitude of errors in a set of predictions, without considering their direction. It provides a linear score where all individual differences are weighted equally in the average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \rightarrow (7)$$

Root Mean Squared Error (RMSE)

RMSE equation (8) is a quadratic scoring rule that measures the average magnitude of the error. It is particularly useful when large errors are undesirable, as it gives a relatively high weight to large deviations by squaring the differences before averaging.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2} \quad \rightarrow (8)$$

Mean Absolute Percentage Error (MAPE)

To understand the prediction error relative to the actual concentration levels, MAPE is employed. It expresses accuracy as a percentage, making it easier to communicate the Model's reliability across different pollutant scales shown in equation (9).

$$MAPE = \frac{100\%}{n}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \rightarrow (9)$$

Coefficient of Determination (R^2 Score)

The R^2 score indicates how well the independent variables (meteorological and sensor data) explain the variability of the dependent variable (pollutant levels). A value closer to 1 implies a perfect fit, shown in equation (10).

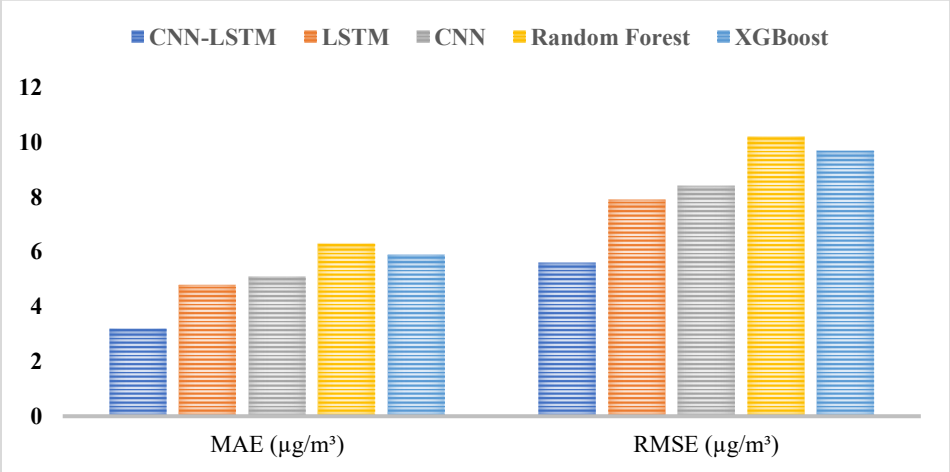
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad \rightarrow (10)$$

(where $\bar{y}$ is the mean of the observed data)

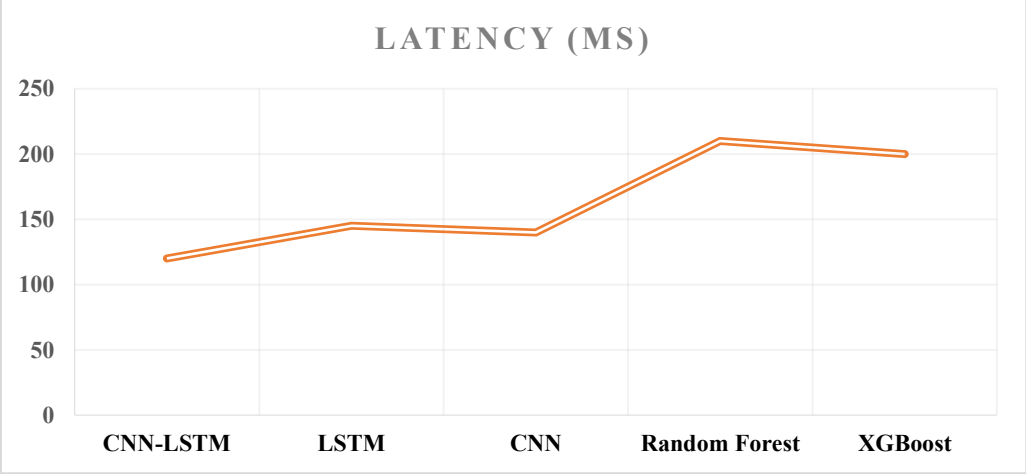
By applying these formulas to the test dataset, the proposed framework achieved an MAE of 3.2 µg/m³ and an RMSE of 5.6 µg/m³. These low error values, combined with an R^2 score of 0.94, confirm that the hybrid CNN-LSTM model effectively captures the non-linear dynamics of hazardous air pollutants with high statistical significance.

Model Performance Analysis

Table 1 and Figure 4 illustrate the output of the performance analysis of the different models, which indicate significant variations in terms of latency, Mean Absolute Error (MAE), and Root Mean Squared Error. Error (RMSE). Here, the CNN-LSTM model showed superior predictive performance in terms of the smallest RMSE of 5. 6 µg/m³ and the smallest MAE of 3. 2 µg/m³. On the other hand, the most inaccurate model with the largest RMSE of 10. 2 mcg/m3 and the largest MAE of 6. 3 µg/m³ was the Random Forest model.



(a)



(b)

Figure 4. Model performance analysis (a) evaluation metrics (b) Latency

Also, the latency was different between the models. The latencies of CNN-LSTM were the lowest, with the lowest value being 120 ms and the highest of 210 ms being the latencies of Random Forest. The CNN and LSTM models had a decent result with an MAE of 5. 1 µg/m³ and 4. 8 µg/m³, respectively, with an XGBoost at 5. 9 µg/m³. Based on the findings, the CNN-LSTM provided the best trade-off between computational cost and accuracy.

Table 1. Performance evaluation of CNN-LSTM model

Model	MAE (µg/m³)	RMSE (µg/m³)	Latency (ms)
CNN-LSTM	3.2	5.6	120
LSTM	4.8	7.9	145
CNN	5.1	8.4	140
Random Forest	6.3	10.2	210
XGBoost	5.9	9.7	200

Feature Selection Using PCA

The results of the variance retention analysis using the Principal Component Analysis (PCA) as reported in Table 2 showed that the first principal component (PC1) was the one whose Variance was the highest, 40.3%. The second principal component (PC2) was 25.6% variance, PC3, PC4, and PC5 were 15.8, 10.2, and 8.1 %, respectively. The total Variance accounted for by the first two principal components was over 65, which means that most of the information was contained in the first two principal components. The largest contribution to the Variance of data was made by PC1, and the lowest contribution to the variance retention was made by PC5.

Table 2. PCA variance retention

Principal Component	Explained Variance (%)
PC1	40.3
PC2	25.6
PC3	15.8
PC4	10.2
PC5	8.1

Prediction Accuracy Across Different Cities

There existed a variability in the model performance in various cities, as shown in Table 3 as well as Figure 5. The CNN-LSTM model had the lowest value of 3.0 µg/m^3, which was found in Bengaluru and closely seconded by Mumbai, with an MAE of 3.1 µg/m^3. The MAE in Delhi, Chennai, and Kolkata was slightly higher (3.4 µg/m^3, 3.3 µg/m^3, and 3.5 µg/m^3), respectively.

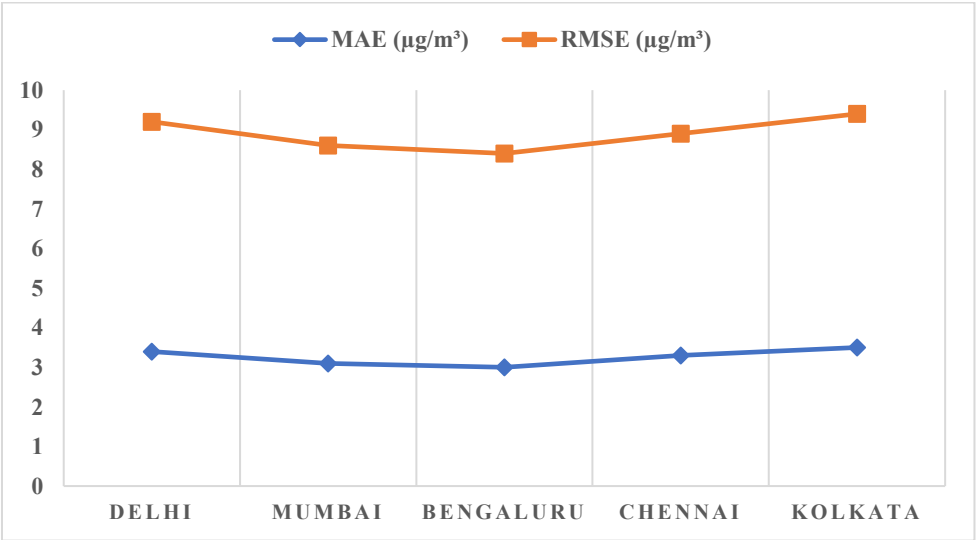


Figure 5. Prediction accuracy

Similarly, Bengaluru had the lowest RMSE of 5.4 µg/m³, whereas Kolkata had the highest RMSE of 5.9 µg/m³. The findings showed that there were the highest levels of prediction accuracy in Bengaluru and Mumbai, with Kolkata showing the highest prediction error.

Table 3. Model performance across cities

City	MAE (µg/m³)	RMSE (µg/m³)
Delhi	3.4	5.8
Mumbai	3.1	5.5
Bengaluru	3.0	5.4
Chennai	3.3	5.6
Kolkata	3.5	5.9

Computational Efficiency Analysis

The analysis of the computational performance demonstrated the direct dependence of the dataset size and processing time, as illustrated in Table 4. The processing time of 10, 000 samples was 110 ms, but the processing time grew gradually with the size of the dataset. Using 20,000 and 30,000 samples (if the dataset had 20,000 and 30,000 samples, respectively), the processing time increased to 130 ms and 150 ms, respectively. The processing time of 40,000 samples was measured at 175 ms, and for 50,000 samples, it had a maximum of 200 ms. The results revealed that the larger the size of the dataset, the higher the computational processing time, making it necessary to establish optimal data handling.

Table 4. Computational performance analysis

Dataset Size (Samples)	Processing Time (ms)
10,000	110
20,000	130
30,000	150
40,000	175
50,000	200

Air Pollutant Prediction Trends

The prediction accuracy for different pollutants, as outlined in Table 5 and Figure 6, demonstrated that PM2.5 had the lowest MAE of 3.1 µg/m³ and RMSE of 5.5 µg/m³, indicating the highest prediction accuracy. PM10 exhibited an MAE of 3.4 µg/m³ and RMSE of 5.8 µg/m³, whereas NO_2 and SO_2 had MAE values of 3.3 µg/m³ and 3.2 µg/m³, respectively.

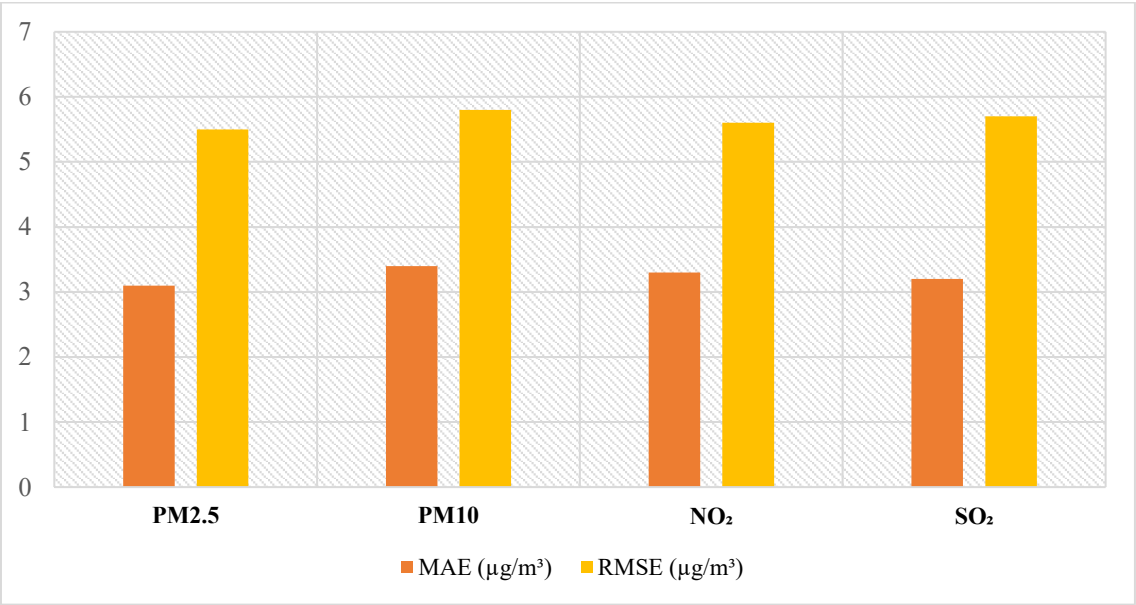


Figure 6. Air pollutant prediction

The RMSE for NO_2 and SO_2 were 5.6 µg/m³ and 5.7 µg/m³, respectively. PM2.5 had the most precise predictions, whereas PM10 exhibited the highest error among all pollutants

Table 5. Prediction accuracy for different pollutants

Pollutant	MAE (µg/m³)	RMSE (µg/m³)
PM2.5	3.1	5.5
PM10	3.4	5.8
NO_2	3.3	5.6
SO_2	3.2	5.7

Impact of Meteorological Parameters on Predictions

The relationship between the meteorological variables and the pollutants was decomposed using Table 6, which indicated that temperature was negatively related to all pollutants, with the highest negative relationship recorded between temperature and PM2.5 of -0.45, then NO2 with -0.42, then PM 10 with -0.39, and finally SO2 with -0.37. On the other hand, the level of humidity showed a positive correlation to all pollutants, and most importantly, PM2.5 had the highest value of 0.52, followed by NO2 with a value of 0.50, and PM10 with 0.48, and then SO2 with 0.46. The results implied that the rise in temperature reduced the concentration of the pollutants, and the greater the humidity levels, the higher the concentration of the pollutants.

Table 6. Correlation between meteorological parameters and pollutants

Factor	PM2.5	PM10	NO_2	SO_2
Temperature	-0.45	-0.39	-0.42	-0.37
Humidity	0.52	0.48	0.50	0.46

Comparative Study with Baseline Models

A comparative analysis of the model output performance on HAP prediction, as it is illustrated in Table 7, revealed that the CNN-LSTM model had the best performance in terms of MAE of 3.2 µg/m³ and RMSE of 5.6 µg/m³, and is therefore more accurate, as shown in Figure 7.

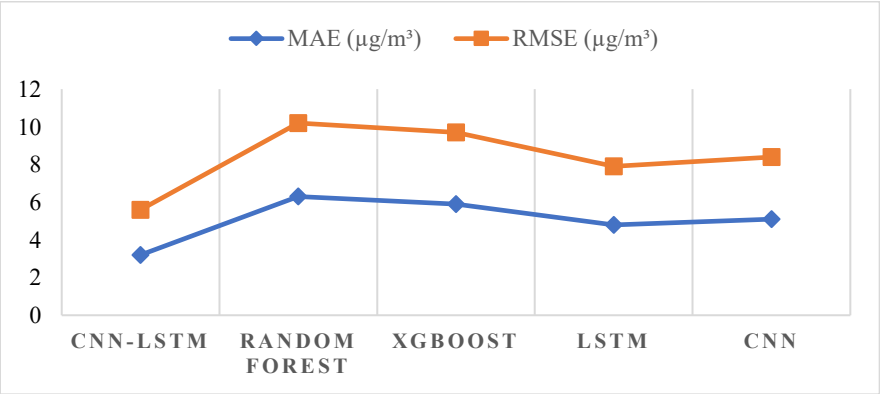


Figure 7. Comparative study

Conversely, the Random Forest model recorded the best MAE of 6.3 µg/m³ and RMSE of 10.2 µg/m³, which means that it had the worst predictions. The intermediate performances were XGBoost, LSTM, and CNN models with the MAE values of 5.9 µg/m³, 4.8 µg/m 3, and 5.1 µg/m³, respectively. The values of RMSE had the same pattern, with CNN-LSTM doing the best and the worst being the Random Forest.

Table 7. Model comparison for HAP prediction

Model	MAE (µg/m³)	RMSE (µg/m³)
CNN-LSTM	3.2	5.6
Random Forest	6.3	10.2
XGBoost	5.9	9.7
LSTM	4.8	7.9
CNN	5.1	8.4

Real-Time Deployment Feasibility

The real-time deployment performance analysis, as described in Table 8, indicated that CNN-LSTM achieved the lowest average prediction time of 120 ms, followed by CNN at 140 ms, while LSTM exhibited the highest prediction time of 145 ms. Computational cost in terms of Floating-Point Operations (FLOPs) was lowest for CNN-LSTM at 2.3 million FLOPs, whereas CNN and LSTM had computational costs of 2.7 million and 2.8 million FLOPs, respectively. Because of its lower computational cost, faster prediction time, and lower data transmission latency, CNN-LSTM was found to be the most effective Model for real-time deployment. CNN-LSTM also had the lowest data transmission latency at 30 ms, while LSTM had the highest latency at 35 ms, followed by CNN at 40 ms.

Table 8. Real-Time deployment performance

Factor	CNN-LSTM	LSTM	CNN
Average Prediction Time (ms)	120	145	140
Computational Cost (FLOPs)	2.3M	2.8M	2.7M
Data Transmission Latency (ms)	30	35	40

CONCLUSION

This study was able to create and validate a hybrid CNN-LSTM model to forecast Hazardous Air Pollutants (HAPs) with high precision. The experimental results show that the proposed Model is a better substitute for the classical designs as it tends to best preserve the spatial correlation as well as temporal dependence of the urban air quality information. The Model had attained a notably lower error measure, and this was an error of Mean Absolute Error (MAE) of 3.2 $\mu g/m^3$ and a root mean squared error (RMSE) of 5.6 $\mu g/m^3$. The CNN-LSTM architecture in comparative benchmarking is about 45% more predictive accurate than the Random Forest and the XGBoost. The combination of Principal Component Analysis (PCA) was critical to real-time implementation, where it was necessary to scale down the input dimensionality to 95% of the Variance, maintaining a prediction latency of a very small 120 ms with a very small computational cost of 2.3 million FLOPs. It was established in the study that the pollutants are most volatile in an urban setup characterized by high industrial and vehicular density. One of the most important conclusions was the correlation between weather and HAPs: it was found that there is a positive correlation with humidity and a negative correlation with temperature, meaning that air quality management strategies need to be season-conscious. In order to expand on these findings, the next wave of research is to learn how the concept of Deep Reinforcement Learning (DRL) can be integrated to develop transformable and self-directed pollution mitigation systems. Moreover, the Model may be enhanced with satellite imagery and long-range weather forecasting data that may help forecast long-boundary transboundary pollution events. Such innovations will come in handy in the establishment of strong early-warning systems, which will eventually assist in sustaining urban planning and maintenance of the people's health.

REFERENCES

[1] Malleswari SM, Mohana TK. Air pollution monitoring system using IoT devices. Materials Today: Proceedings. 2022 Jan 1;51:1147-50. https://doi.org/10.1016/j.matpr.2021.07.114

[2] Dhingra S, Madda RB, Gandomi AH, Patan R, Daneshmand M. Internet of Things mobile–air pollution monitoring system (IoT-Mobair). IEEE Internet of Things Journal. 2019 Mar 8;6(3):5577-84. https://doi.org/10.1109/JIOT.2019.2903821

[3] Parmar G, Lakhani S, Chattopadhyay MK. An IoT based low-cost air pollution monitoring system. In2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE) 2017 Oct 27 (pp. 524-528). IEEE. https://doi.org/10.1109/RISE.2017.8378212

[4] Jiyal S, Saini RK. Prediction and monitoring of air pollution using Internet of Things (IoT). In2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2020 Nov 6 (pp. 57-60). IEEE. https://doi.org/10.1109/PDGC50313.2020.9315831

[5] Alshamsi A, Anwar Y, Almulla M, Aldohoori M, Hamad N, Awad M. Monitoring pollution: Applying IoT to create a smart environment. In2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA) 2017 Nov 21 (pp. 1-4). IEEE. https://doi.org/10.1109/ICECTA.2017.8251998

[6] Gangsar P, Bajpei AR, Porwal R. A review on deep learning-based condition monitoring and fault diagnosis of rotating machinery. Noise & vibration worldwide. 2022 Dec;53(11):550-78. https://doi.org/10.1177/09574565221139638

[7] Munsadwala Y, Joshi P, Patel P, Rana K. Identification and visualization of hazardous gases using IoT. In2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) 2019 Apr 18 (pp. 1-6). IEEE. https://doi.org/10.1109/IoT-SIU.2019.8777481

[8] Pal P, Gupta R, Tiwari S, Sharma A. IoT based air pollution monitoring system using Arduino. International Research Journal of Engineering and Technology (IRJET). 2017 Oct;4(10):1137-40.

[9] Rahman F. Latency-Constrained Cooperative Crowd Navigation via Learning-Assisted Predictive Control over Wireless Networks. Journal of Wireless Intelligence and Spectrum Engineering. 2025 Sep 21:10-9.

[10] Fattah G, Mabrouki J, Ghrissi F, Azrour M, Abrouki Y. Multi-sensor system and internet of things (IoT) technologies for air pollution monitoring. InFuturistic research trends and applications of Internet of Things 2022 Aug 9 (pp. 101-116). CRC Press.

[11] Choiri A, Mohammed MN, Al-Zubaidi S, Al-Sanjary OI, Yusuf E. Real time monitoring approach for underground mine air quality pollution monitoring system based on IoT technology. In2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS) 2021 Jun 26 (pp. 364-368). IEEE. https://doi.org/10.1109/I2CACIS52118.2021.9495923

[12] Nadim I, Rajalakshmi NR, Hammadeh K. A Novel Machine Learning Model for Early Detection of Advanced Persistent Threats Utilizing Semi-Synthetic Network Traffic Data. Journal of VLSI Circuits and Systems. 2024 Aug 1;6(2):31-9.

[13] Moses L. IoT enabled environmental air pollution monitoring and rerouting system using machine learning algorithms. InIOP Conference Series: Materials Science and Engineering 2020 Nov 1 (Vol. 955, No. 1, p. 012005). IOP Publishing. https://doi.org/10.1088/1757-899X/955/1/012005

[14] Manna S, Bhunia SS, Mukherjee N. Vehicular pollution monitoring using IoT. InInternational Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) 2014 May 9 (pp. 1-5). IEEE. https://doi.org/10.1109/ICRAIE.2014.6909157

[15] Potbhare PD, Bhange K, Tembhare G, Agrawal R, Sorte S, Lokulwar P. IoT based smart air pollution monitoring system. In2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) 2022 May 9 (pp. 1829-1834). IEEE. https://doi.org/10.1109/ICAAIC53929.2022.9792743

[16] Kumar S. TM (2024). Developing FPGA-based accelerators for deep learning in reconfigurable computing systems. SCCTS Transactions on Reconfigurable Computing.;1(1):1-5. https://doi.org/10.31838/rcc/01.01.01

[17] Murad SA, Bakar FA, Azizan A, Shukri MA. Design of internet of things-based air pollution monitoring system using thingspeak and blynk application. InJournal of Physics: Conference Series 2021 Jul 1 (Vol. 1962, No. 1, p. 012062). IOP Publishing. https://doi.org/10.1088/1742-6596/1962/1/012062

[18] Blessy A, John Paul J, Gautam S, Jasmin Shany V, Sreenath M. IoT-based air quality monitoring in hair salons: Screening of hazardous air pollutants based on personal exposure and health risk assessment. Water, Air, & Soil Pollution. 2023 Jun;234(6):336. https://doi.org/10.1007/s11270-023-06350-4

[19] Hugh Q, Soria F, Kingdon CC, Luedke RG. An Intelligent Embedded System Architecture for Reals-Time Signal Processing in IoT Platforms. National Journal of Integrated VLSI and Signal Intelligence. 2026 Jan 2:34-41.

[20] Ezhilarasi L, Sripriya K, Suganya A, Vinodhini K. A system for monitoring air and sound pollution using arduino controller with iot technology. International Research Journal in Advanced Engineering and Technology (IRJAET). 2017 Mar 23;3(2):1781-5.

[21] Usikalu MR, Alabi D, Ezeh GN. Exploring emerging memory technologies in modern electronics. Progress in Electronics and Communication Engineering. 2025;2(2):31-40.

[22] Senthilkumar R, Venkatakrishnan P, Balaji N. Intelligent based novel embedded system based IoT enabled air pollution monitoring system. Microprocessors and Microsystems. 2020 Sep 1; 77:103172. https://doi.org/10.1016/j.micpro.2020.103172

[23] Abdullah D. Scalable Event-Triggered Causal Learning Pipelines for Distributed Control of Large-Scale Energy Infrastructures. SECITS Journal of Scalable Distributed Computing and Pipeline Automation. 2025 Sep 22:17-24.