A Hybrid Deep Learning Framework For Analyzing, Predicting, and Forecasting the Severity Level of Air Pollution in India

[1,*]P. Aruna Rani, [2]Dr.V. Sampathkumar

[1] Research Scholar, Department of Civil Engineering, Sathyabama Institute of Science and Technology, Chennai-600119.

[2] Professor, Department of Civil Engineering, Sathyabama Institute of Science and Technology, Chennai-600119.

Email-ID: [1]arunaranip76@gmail.com, [1]arunanates@gmail.com, [2]svsjpr@gmail.com

Abstract

A significant problem occurs with natural resources, such as air pollution caused by various environmental factors and climate change. Air pollution poses a major threat to human health and sustainability. The varying levels of air pollutants mix dynamically, increasing air pollution and impacting human health proportionally to their basic health conditions. For example, a severity level of the air pollution immediately affects an old person or someone with breathing issues and can lead to sudden death. To save people, it is essential to develop an accurate and timely forecasting system to mitigate its adverse effects and take immediate action. Conventional forecasting systems use statistical and basic AI methods, often struggle to process complex and large amounts of continuous data generated from the air. Also, spatiotemporal dependencies from the air quality data were not extracted. Thus, this paper proposed a hybrid DL model, integrating a CNN with LSTM to analyse and accurately forecast the severity levels of air pollution. Basically, CNN model helps to extracts the spatial features from the air quality data while the LSTM model used to extract the temporal dependencies. The proposed CNN-LSTM can provide a robust prediction model for air pollution. The CNN-LSTM model is evaluated by implementing it in Python and experimenting with real-world datasets from various surveillance monitoring stations. The overall performance of the proposed CNN-LSTM is compared with the standalone LSTM, CNN and traditional ML models such as RF and SVM. The final result indicate that proposed DL-based hybrid CNN-LSTM model performs healthier than the others and obtains the highest forecasting accuracy.

Keywords: Air Pollution, Deep Learning Model, CNN-LSTM, Pollutant Level Estimation, Air Pollution Vs. Human Health.

Introduction

Air pollution is the pollution of the air by harmful substances such as particulates, gases, and biological molecules. It may cause allergies, diseases, and human death [1]. It may also cause damage to other living organisms such as food crops, animals, and the natural environment. It can be both man-made and natural. Man-made air pollution contains emissions from power generation, motor vehicles, industrial processes, and agricultural activities. Air pollution is an environmental issue affecting millions of people's health. It occurs only when harmful substances are introduced into the Earth's atmosphere. Some human activities that affect the quality of the air by making it pollutants like vehicle emission and burning fossil fuels like coal, oil, and natural gas which are the major reason for the air pollution, particularly in urban areas [2]. Industrial activities also play a vital role in air pollution. It emits extensive pollutants like heavy metals, volatile organic compounds, and other toxic substances. The

agricultural industry also contributes to it. Pesticides and fertilizers release VOCs, PM, and ammonia (NH3) into the air. Farming produces a potent greenhouse gas, NH3, and methane (CH4), combined with other pollutants in order to create fine particulate matter (PM2.5) [3]. PM2.5 can pass through the lungs and get into the blood, having an effect on such health-related issues like heart attack, strokes, an asthma attack, chronic obstructive pulmonary disease (COPD) and lung cancer [4].

Air pollution detection is crucial for environmental monitoring and people health protection. Various methods are used for detecting, such as sensors, air quality monitoring systems, remote monitoring, analytical methods, and particulate air sampling techniques [5]. Using sensors to detect some pollutants like nitrogen dioxide (NO2), sulphur dioxide (SO2), particulate matter (PM), carbon monoxide (CO) and ozone (O3) [6]. Remote monitoring techniques contain ground and satellite-based remote sensing, allowing large-scale air quality assessments across broad geographic areas. Satellite remote sensing uses some instruments like Ozone Monitoring Instrument (OMI) and the Moderate Resolution Imaging Spectroradiometer (MODIS) which helps to measure pollutants like ozone and aerosols [7]. Ground-based remote sensing uses Differential Optical Absorption Spectroscopy (DOAS), Light Detection and Ranging (LiDAR) to detect and track pollutants [8]. Mobile monitoring includes vehicles equipped with air quality sensors to measure pollution levels while traveling to various places. It is mainly useful for urban areas where high pollution levels vary over short distances. Advanced analytical methods like gas chromatography-mass spectrometry (GC-MS) and high-performance liquid chromatography (HPLC) used to determine the pollution level. HPLC is mainly used to analyse air pollutants for organic and inorganic compounds. GC-MS is especially effective for analysing volatile organic compounds like VOCs [9].

The efficiency of DL models, especially, LSTM networks, in detecting and predicting air pollution. It is a type of recurrent neural network (RNN) especially designed to capture long-term dependencies and trends in sequential data, making it highly effective for air pollution prediction [10]. It is also valuable for real-time monitoring, assessing the effect of interventions, and anomaly detection. It also has some challenges, such as computational resources and data quality. LSTM and other deep learning models will be crucial in air pollution management and mitigation. The following sections discuss the literature survey, the proposed approach, the results of the model, and the conclusion.

Contribution of the paper

This research makes several important contributions to the context of air pollution analysis and detection, specifically within the Indian environmental region.

1. This paper presents an advanced hybrid model that integrates the work of a Convolutional Neural Network (CNN) and a Long-Short-Term Memory (LSTM) model to analyse the temporal and spatial dependencies in the air quality data. This hybrid approach helps to enable more robust, accurate, and context-aware forecasting of air pollution levels.

2. The CNN portion is applied to recognize intricate spatial connections of several air pollutant variables, and the LSTM layer is used to obtain long-term tendencies and fluctuations in time. Such a combined framework effectively overcomes the shortcomings of conventional models that do not take into consideration dynamic interactions inherent in the air quality data over time and across regions.

3. This model not only helps to forecast the future air pollutant result but also classifies the levels and types of air pollution. This enables alerts for vulnerable pollution and timely interventions, which helps to enhance the public health response mechanisms.

4. Further experiments were conducted using real-time air pollution data taken from multiple monitoring stations across various Indian regions. This ensures that the models' performance impacts the real-time challenges and conditions of the environment, including meteorological influences, different types of pollutants, and differences in various regions.

5. To evaluate the overall performance of the proposed hybrid CNN-LSTM model, it was compared with some existing ML and DL models like CNN, SVM, Random Forest (RF), and LSTM. The experimental result stated that this proposed CNN-LSTM model provides more robustness and accuracy in analysing and forecasting the severity levels of air pollution.

6. The execution of the CNN-LSTM model in Python and the model's ability to adapt to large-scale data support environmental surveillance systems and smart city infrastructures.

Literature survey

Due to advancements in deep learning models, DL-based models have recently gained widespread use in air pollution detection. Among various air pollutants, NO2 and SO2 are the primary pollutants that cause several health issues. For accurate prediction of these pollutant particles in the air, A. Heydari et al. (2022) have proposed a hybrid DL model (LSTM-MVO), for pollution detection. The result of the LSTM-MVO model is compared with the other models. The comparison outcome depicts that the LSTM-MVO model predicts the presence of pollutant particles with high accuracy and low RMSE. Due to various environmental factors, the pollutant level increases in smart cities. This creates multiple types of health problems, especially respiratory diseases. Thus, a hybrid CNN-LSTM-based approach is designed to analyse and forecast air pollution levels in Beijing, China (A. Bekkar et al. (2021). The model's performance is compared with other models like standalone CNN, LSTM, GRU, Bi-GRU, and Bi-LSTM. The overall comparison result proves that the CNN-LSTM model performs better in predicting air pollutant levels with lower RMSE and MAE rates of 23,921 and 6,742, respectively. L. Zhu et al. (2023) have proposed a deep learning-based model, CNN, to detect water quality and an LSTM for predicting air quality in urban areas. The model's performance is evaluated using F1-score, accuracy, etc, and compared to the existing models. Compared to other methods, the CNN-LSTM has achieved 92% and 91% accuracy on predicting water and air pollution levels in urban areas. Xing. J. et al. (2020) demonstrated a novel model that integrates chemical indicator data with a DL model to forecast pollutant levels. Using the chemical transport simulator, the AQI ratio is estimated. That can be classified using the deep learning model. The simulation output is compared with the earlier machine learning methods and found that the CTM-deep learning outperforms the others. However, its computational complexity is high and takes more time to simulate and forecast.

Periyanan and Palanivel Rajan (2024) have proposed a modified Gated Recurrent Unit architecture for forecasting air pollution. It uses a Dual-Slope Leaky ReLU activation function to activate the internal layers and filter functions to process the data efficiently. The activation function fine-tunes the parameters with the help of the female SWO algorithm. By combining the robustness of SWO, GRU improves its predictive efficacy in air pollution forecasting. Yoo and Oh (2020) demonstrated that deep learning models have powerful data learning abilities and provide more efficiency in time-series data analysis for forecasting. The LSTM model outperforms time-series data prediction, and thus, it has been used for air quality analysis in Madrid, Navares, and Aznarte (2020). However, LSTM fails to process seasonal data effectively. Hence, a seasonal-LSTM (SLSTM) was proposed to solve the issues in processing seasonal air quality data (Skarlatos et al., 2023). From the experiment, it is identified that LSTM performs better in analyzing and predicting time-series data. Zhao et al. (2023) have used Gated Recurrent Units and stated that they are similar to LSTM, but their training time is

high. Some parameters are adjusted and tuned to reduce the model's training time (Chen et al., 2019). Several research fields have widely used the GRU model (Qin et al., 2022). Following the seasonal data processing, the GRU model is extended to address the existing challenges and proposed seasonal-GRU(SGRU) (Groenen, 2018).

In 2020, ShuWang et al. applied a GRNN for AQI prediction, comparing it to MLP and SVR. Because sensor drifts are less invariant, the gas recurrent neural network performs well, but it is also more susceptible to atmospheric variability and humidity. In 2020, Pasupuleti et al. compared decision trees, linear regression, and random forests. Significant air pollutants and meteorological conditions are obtained through the application of Arduino. Due to its overfitting ability, Random Forest provides accurate outcomes that minimize errors. The main limitation of Random Forest is that it requires more memory and incurs higher costs. In 2019, Desislava Ivanova and Angel Elenkov applied the Raspberry Pi platform along with MLP algorithms from ML for accurate air pollutant predictions. The multilayer perceptron surpasses the classification problem applied to discrete values and the regression used for continuous values. Due to the use of discrete values, multilayer perceptrons with backpropagation result in inputs that, when not passing the activation function, yield outputs of 0 or 1.The attainment of the coefficient of determination (R2) is better when the need for incremental feeding is enhanced. Fan et al. (2018) presented a study that defines the impacts of air pollution and solar radiation. Using an SVM model, six air pollutants, PM2.5, PM10, SO2, NO2, CO, and O3, are analyzed and predicted. The result of the model shows that this model has achieved better results with a lower RMSE value. In polluted regions, enhancing the accuracy of Rs and Rd predictions depends on selecting suitable air pollution inputs.

Bhuvaneshwari et al. (2022) have proposed a Gaussian SVM model for Air Pollution Prediction. Monitoring air pollution in dynamic real-time environments is not accurate; despite using advanced WSN technology, there are limitations, such as insufficient coverage of wide regions. To overcome this barrier, the paper primarily focuses on a region-based air pollution system for monitoring real environments in smart cities. The system consists of two phases for predicting heavy and light traffic areas, utilizing the Gaussian SVM model to forecast air pollutants like PM10, CO, NO2, PM2.5 and O3. Meta-heuristic algorithms are employed to select the predicted areas, where sensor nodes are subsequently placed. For the cross-validation process, the dataset is divided into training and testing sets. As a result, a Mean Error prediction value of 9.83 is achieved, which is lower than that of traditional model solutions, and this SVM-based model attains 95% accuracy. Farooq et al. (2024) have presented a paper on an enhanced approach for predicting air pollution using quantum support vector machines. In machine learning-based models, SVM is commonly used for classification and proves to be more effective. The increased dataset complicates the selection of suitable features, which must outperform the classification process. This proposed model utilizes the SVM for feature map selection and employs a standard dataset for air quality prediction. In the experiment, by utilizing the quantum lab and IBM quantum computing cloud, the accuracy of the quantum SVM outperforms the classical SVM model in air quality prediction. As a result, using the same dataset for both classical SVM and quantum-based SVM, the accuracy attained by the classical SVM model ranges from 87% to 91%, while the quantum SVM model's accuracy ranges from 94% to 97%. This result indicates that optimal feature map selection is key for accurately predicting air pollution.

Limitations of the Existing model

The traditional approaches have several limitations for predicting air quality. The convolutional model primarily addresses temporal trends by utilizing time-series data or capturing spatial correlations, but it is not able to process both simultaneously. Meteorological variables like

wind speed and temperature are not considered, yet these are important for pollutant accumulation and dispersion. These limitations affect prediction accuracy and generalizability. The existing models lack the capability to capture spatiotemporal dependencies and dynamic environmental factors, which leads to failures in handling high-resolution time series data and regional generalization. Current ML models and some single DL models, such as CNN or LSTM, demonstrate limited adaptability across different geographical areas, resulting in decreased accuracy for new geographic locations or extreme scenarios.

Motivation for the proposed model

To address these limitations, a new hybrid CNN-LSTM architecture is proposed for both spatial patterns and temporal sequences. The proposed CNN-LSTM model integrates convolutional layers and LSTM units for spatial feature extraction and captures long-term dependencies while combining meteorological variables. This study makes a novel contribution to air pollution forecasting by initially combining CEEMDAN-based feature extraction with a PSO-optimized CNN-LSTM. The proposed approach incorporates hyperparameter tuning, deep learning in a separate framework, and single decomposition, whereas the traditional approach applies CNN-LSTM or optimization individually. As evidenced by experimental results with a real-time air pollution dataset, the proposed model enhances both trustworthiness and prediction accuracy.

Problem statement

One of India's emerging and most pressing environmental challenges is air pollution. It increases the death rate among elderly individuals and severely impacts those with respiratory illnesses. Accurately assessing and predicting air quality is complex because the concentration of pollutants in the air is highly dynamic. Earlier methods utilized machine learning and other conventional AI algorithms that performed adequately. However, they exhibited several limitations in accurately analyzing and extracting spatiotemporal feature patterns from air quality and pollution data. Moreover, earlier systems often failed to promptly provide precise prediction outputs, which are essential for warning the public to take preventive actions.

To overcome these challenges, it is essential to develop an advanced data analytics and forecasting framework capable of managing high-dimensional air quality data while maintaining spatial and temporal features. This paper seeks to bridge the gap by establishing a hybrid deep learning framework that combines a convolutional neural network and a long short-term memory network to effectively handle high-dimensional data, analyze, predict, and accurately forecast the severity level of air pollution. The CNN model addresses spatial dependencies, while the LSTM model addresses temporal dependencies, allowing the hybrid CNN-LSTM to manage spatiotemporal dependencies for precise predictions of air pollution severity levels. The aim of designing the hybrid deep learning framework model is to enhance the predictive accuracy and perform better than conventional methods, supporting proactive decision-making regarding public health management.

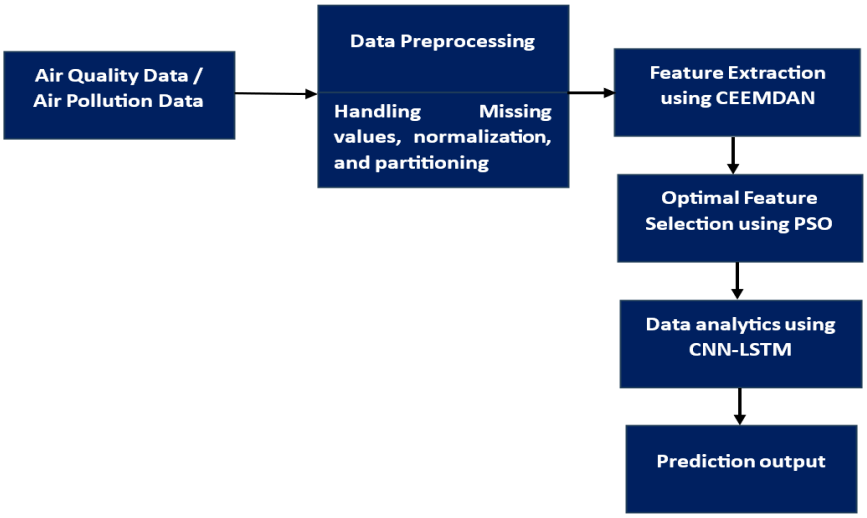
Existing Method

Traditionally, various research works have been performed to forecast air quality. For example, the authors C.H. Cordova et al. (2021) have proposed an MLP and LSTM recurrent ANN model to predict the air pollutant level in metropolitan Lima, Peru. The air pollutant level is observed based on the values observed from five stations. The final result of the model indicates that the LSTM combined with recurrent ANN model performed better and had a high precision value. Though this model performed better, it required additional features and a self-identification technique for future development and model identification. To overcome these

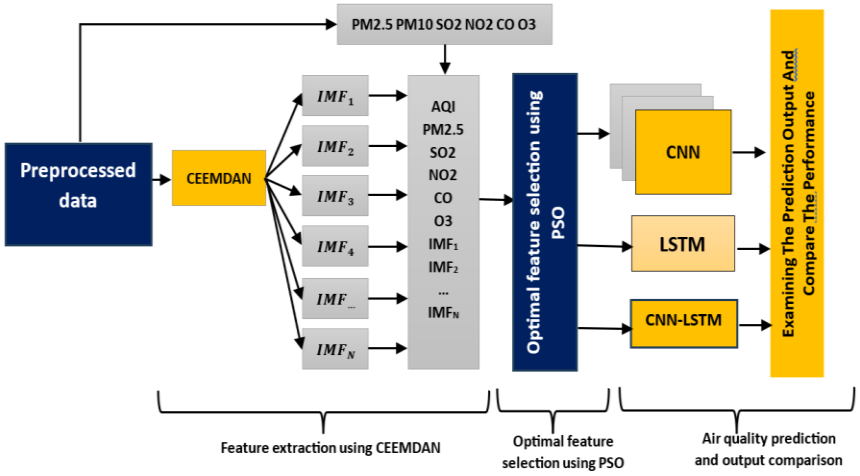
issues, this paper proposes an LSTM-based Knowledge discovery extraction system to predict and identify the air pollutant level accurately.

Proposed methodology

The proposed methodology introduces a **DL-based hybrid (CNN-LSTM)** for accurate and robust prediction of air pollution by leveraging spatial and temporal air quality data features. The framework consists of several key components, as indicates in Figure-1.



(a) Overall Workflow



(b) Core Data Processing

Figure-1: Overall Workflow Of The Proposed Model

Data Pre-Processing

The air quality data accumulated from openly accessible platforms like Kaggle, IoT-based sensor networks, and Indian Pollution Control Boards is followed by the data collection and preprocessing stage. This dataset contains various features consisting of pollutant concentrations like NO2, SO2, PM2.5, CO, and O3, as well as atmospheric conditions like humidity, temperature, atmospheric pressure, weather, and equivalent temporal codes. Applying the combined mean imputation, interpolation, and forward filling approaches

addresses the missing value to maintain data completeness. The performance and the overlap are improved in the neural network models, particularly in the DL technique, where the feature generalisation outperforms the Min-Max Scalling. By this approach, it converts the given input features into a certain range of 0 and 1, and it is mathematically formulated as,

$$x_{normalization} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Here, the original feature value is shown as x, the minimum value of the feature is represented as x_{min} and maximum values of the features are presented as x_{max}. LSTM network was used to prepare sequence modelling, this dataset was modified to time series windows, allowing the proposed model to learn time links and complex patterns throughout continuous past monitoring.

Feature extraction and pollutant concentration prediction

The main goal is to enhance the precision of predicting air pollution concentration by applying an advanced model called Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). This method is used for feature extraction, allowing the model to effectively analyze and break down complex environmental signals into simpler components. Doing so can identify and utilize the most relevant features contributing to pollution levels, leading to more accurate and reliable predictions.

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

It is the process that used to optimise the feature extraction process by altering the complex non-stationary and nonlinear time series $x(t)$ into a perfect set of simpler oscillatory components is known as Intrinsic Mode Functions (IMFs). Mathematically, CEEMDAN was evaluated using the following equation:

$$x(t) = \sum_{i=1}^{n} IMF_i\,(t) + r_n(t)$$

In the above equation, $r_n(t)$ Denoted the final residual components once all the IMFs are extracted and $IMF_i(t)$ denoted the $i-th$ intrinsic mode functions. Unlike the traditional EEMD or EMD approach, the CEEMDAN approach performs ensemble averaging, which helps to improve reconstruction stability, reduce mode fixing, and introduce adaptive white noise in the decomposition process. This multiscale analysis helps capture important temporal frequency patterns and filter the frequency noise, making this model more suitable for analysing environmental data such as air pollutant levels. The IMFs obtained by decomposition, each representing a specific band, serve as an input to DL models like CNN-LSTM, which helps to improve the learning process by providing more relevant and cleaner feature sets. The parameter settings are commonly the number of realisations (e.g., 100). For instance, (0 2) is the noise standard deviation and the stopping criterion of IMF extraction. This CEEMDAN-based decomposition, as illustrated in the document, enhances the overall performance of the CNN-LSTM network by identifying complex trends, which are modeled temporally using the LSTM model, and spatial features are analyzed using the CNN model, which leads to reducing the errors while forecasting and also enhances the prediction accuracy in predicting pollutant concentration. The important parameters that are used for the implementation of CEEMDAN are provided in table-1

Table 1. Parameters of CEEMDAN

Parameters	Description	Value Used
Ensemble size (N)	Number of realizations with different noise instances added	250
Noise amplitude	Standard deviation of the added white Gaussian noise	0.2 × std(x(t))
Max IMF number	Maximum number of IMFs to be extracted	10
Stopping criterion	Threshold on the mean of the standard deviation for residual	0.05
Shifting iterations	Number of shifting iterations for each IMF computation	50
Noise type	Nature of added noise during ensemble generation	Gaussian white noise

CEEDAN-based feature Extraction

The original pollutant time series x(t) is converted into Intrinsic Model Functions (IMF), which are represented by $IMF_1, IMF_2, IMF_3, \ldots, IMF_n$. This can be mathematically represented by:

$$x(t) = \sum_{i=1}^{n} IMF_i(t) + r_n(t)$$

Where the i[th] intrinsic mode function is represented by $IMF_i(t)$, and the final residual is represented by $r_n(t)$The information features offer decomposed IMFs that separate the specific frequency components from the pollutant data. The extracted features are stored in the hybrid CNN–LSTM framework. CNN is used to extract features from the input of the pollutant matrix, and the LSTM model captures the temporal dependencies across time series data. The performance predictions are improved by applying PSO (Particle swarm optimization) which helps to optimize hyperparameters from the LSTM model. The PSO simulated the swarm of particles used to explore an ideal solution by upgrading the position and velocity based on personal and global best performances. These velocity and position-enhanced metrics from PSO are formulated by,

$$v_i^{(t+1)} = wv_i^{(t)} + c_1 r_1\left(p_i^{best} - x_i^{(t)}\right) + c_2 r_2\left(g^{best} - x_i^{(t)}\right)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)}$$

Where, x_i represents the position and v_i represents the velocity of the i^{th} particle, p_i^{best} is the personal best position of a particle i, g^{best} represents the best position in the global among all particles, w represents the inertia weight, c_1 and c_2 represents the acceleration constants, and r_1 and r_2 represents the random numbers from the range between [0,1]. The intersection of CEEMDAN-CNN-LSTM-PSO techniques majorly improves the model's capability to leverage meaningful patterns, noise minimization, and enhanced pollutant concentration assumption through conventional methods.

CNN-LSTM Architecture for Air Quality Examination

The CNN and LSTM network model is a powerful hybrid deep learning technique for studying earthly and space-related dependencies in air quality data. This technique is particularly effective for environmental datasets, which are typically multivariate, non-linear, and time-dependent.

Convolutional Neural Network (CNN)

A CNN is one of the DL-based models which is mainly developed to learn spatial patterns automatically and adaptively from the input data, particularly CNNs used in tasks like images and spatial data. CNNs are specifically effective in tasks like image classification, time-series prediction, and object detection because they can capture the spatial dependencies and local patterns within the data. This model initiates with a Convolutional layer, where small filters, like kernels, extract local features like textures, patterns, or edges. These features develop a feature map, which is passed through non-linear activation functions such as ReLU activation function to present a non-linearity to the model. Then pooling layers helps to minimize the dimensionality and computational cost, maintaining the more vital data from every region. The resultant feature maps are in the format of a 1D vector and passed into fully connected layers, then it is used to extract the features with high-level reasoning. Finally, the output layer generates the final predictions.

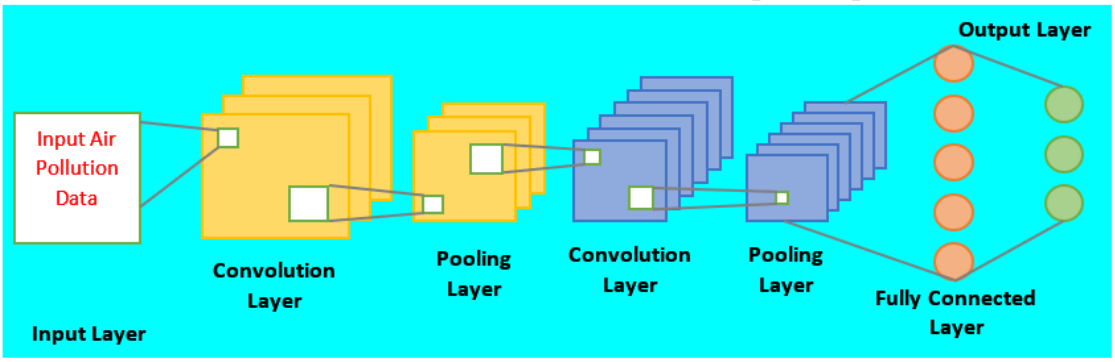


Figure-2: Structure of CNN

The core operation in a Convolutional Neural Network (CNN) can be mathematically represented as:

$$Z_{i,j}^{(l)} = f\left(\sum_{m=1}^{M}\sum_{p=1}^{P}\sum_{q=1}^{q} W_{p,q}^{(m,l)} . X_{i+p-1,j+q-1}^{(m)} + b^{(l)}\right)$$

Where the final output feature of the model is represented as $Z_{i,j}^{(l)}$, the activation function is represented as f, the weight of the kernel filter at position (p, q) is represented as $W_{p,q}^{(m,l)}$, Input feature map value from the mth channel at position (i+p−1, j+q−1) is represented by $X_{i+p-1,j+q-1}^{(m)}$The bias time added to the output term is denoted as $b^{(l)}$The total number of input channels is represented by M and the dimension of the filter is represented by $P \times Q$.

Long Short-Term Memory (LSTM)

LSTM model was also used in the integration with other networks such as CNN, mainly to analyze data like images and videos. The LSTM architecture includes three main gates that manage its memory cell which inlcudes input gate, the forget gate, and the output gate. They control which pieces of information get into and leave the memory cell at any time.

Specifically, the input gate plays a key role in determining how much new data should be stored within the memory, helping the model manage and retain important information over time. It also considers the present input and last hidden state input and output values, which range from 0 to 1 for each data point present in the memory cell. The data should be rejected when the value is 0, and the data should be stored when the value is 1. The garbage gate decides which data needs to be eliminated from the memory cell. The hidden data of the memory cell is analyzed through data controlled by the output gate. The system selectively stores, updates, and retrieves the information over the long-term data by using these gates. Using the following equations, the output of each gate is evaluated and detected.

$$forgot\ gate\ (f_t) = \sigma\left(W_f * [h_t - 1] + b_f\right) \quad (1)$$

Where $w_f, b_f, h_t - 1, f_t$ and σ, represents the weight value, bias value, hidden state value, input data point, forget gate, and sigmoid function.

$$input\ gate\ (I_t) = \sigma\left(W_i * [h_t - 1, x_t] + b_I\right) \quad (2)$$

Where, $W_i, and\ W_C$, represent the weighted value and $b_I\ and\ b_C$ represent the bias value. Now, by multiplying the forget gatef_t with old cell state and $I_t * C_t$The updated element chosen by the input gate is updated to the cell state.

$$\widetilde{C_t} = \tanh\left((W_C * [h_t - 1, x_t] + b_C\right) \quad (3)$$

$$C_t = f_t \odot C_{t-1} + I_t \odot \widetilde{C_t} \quad (4)$$

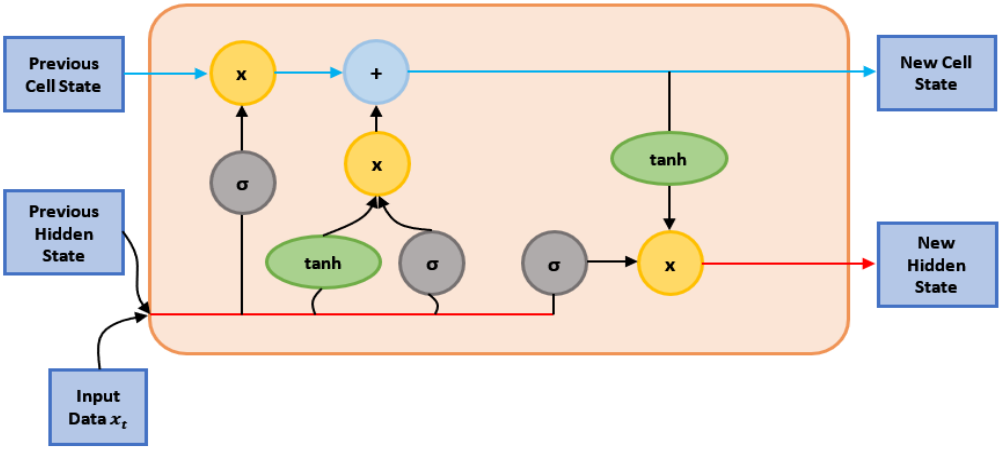


Figure-3: Structure of LSTM

And the $\odot$ denotes the element-wise multiplication and C_t denotes the updated element in the cell state.

$$hidden\ aget\ (h_{g)} = f_g * \tanh\left(C_g\right) \quad (5)$$

$$Output\ Gate\ (O_g) = \sigma(W_O * \left[h_g - 1, x_g\right] + b_O \quad (6)$$

Convolutional Neural Network (CNN) Component

Basically, CNN model helps to extracts features from images and predicts the air quality index. Structured tabular data, such as air pollution, is noted to predict the future. Pollutants across locations and time windows are an example of CNN. CNN can automatically detect and learn features within the data through convolution.

For example:

- Patterns between PM2.5 and NO2 levels in a specific region
- Spatial variations across multiple monitoring stations or locations

CNN helps detect beyond images, like natural language processing, essential for understanding how air components mix or influence one another.

Long Short-Term Memory (LSTM) Component

LSTMs are used to learn complex pattern in the time-series data to predict the air quality. This helps understand how air pollution levels change over time and captures long-term dependencies, enabling more accurate pollution levels. After the CNN extracts features, the LSTM layer processes this data to learn the temporal evolution of pollutant concentrations and how they change over time. LSTM is a RNN type that can predict air pollution levels, helping the model make more accurate long-term pollution forecasts.

Combined CNN-LSTM Workflow for Air Quality

A suitable format model is created, and the Raw air quality data and other relevant data (e.g., PM2.5, NO2, O3) are transformed into it. CNN identifies patterns within data and pollutant interaction features and extracts spatial relationships. The process involves extracting features from the CNN across time to capture and analyze sequential trends. The LSTM network predicts the air quality metric (e.g., PM2.5 to SO2 ratio) based on the learned patterns and temporal relationships.

Table 2. Advantages of CNN-LSTM in Air Quality Applications

Feature	Benefit
Spatial Feature Learning (CNN)	Understands inter-pollutant and inter-location relationships
Temporal Modeling (LSTM)	Captures time-based pollution trends and patterns
Multivariate Capability	Handles multiple pollutants simultaneously
Scalability	Suitable for integration with real-time IoT sensor data
Accuracy	Outperforms many traditional ML models in RMSE, MAE, etc..

In India, where pollution levels vary by region and time (due to the traffic, climate, festivals, crop burning, etc.), the CNN-LSTM model is particularly effective because it adapts to regional spatial differences, urban vs rural. It captures seasonal and event-based spikes like Diwali and winter fog. It can estimate ratios and interactions.

Performance Evaluation

To evaluate the air pollution forecasting model (CNN-LSTM) by using some performance metrices like RMSE (Root Mean Square Error), MEA (Mean Absolute Error) and accuracy. The evaluation techniques and specific relevance are not explained in detail. The brief explanation with mathematical models is provided below:

1. Root Mean Square Error (RMSE)

MSE calculates the average magnitude of prediction error. In air pollution forecasting, lower RSME values suggest that the predicted pollution levels, such as PM2.5 and NO2, closely match the actual values. The obtained RMSE value reflects the accuracy and trustworthiness

of the model. The RMSE heavily punishes the greater errors and makes it capable when the large deviations are particularly undesirable, for instance, the pollution spikes fail to predict.

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

Where the true pollutant concentration at a time t is denoted as y_i, the model predates the concentration at time t is denoted as $\hat{y}_i$ and the total number of data points are denoted as n.

2. Mean Absolute Error (MAE)

The MAE provides an average absolute difference between the actual and predicted values. The MAE is easier to explain than RMSE, and it is less sensitive to outliers. The MAE value is used to determine how much, on average, the model deviates from the true pollutant concentration.

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

Where the true pollutant concentration at a time t is denoted as y_i, the model predates the concentration at time t is denoted as $\hat{y}_i$ and the total number of data points are denoted as n.

3. Accuracy

The severity of air pollution is categorized as safe versus unsafe air; accuracy metrics are used to evaluate how frequently the model correctly predicts pollution categories. These accuracy metrics are vital for innovating a new emergency rule or warning.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ predictions} \times 100$$

Where accuracy is utilized for regression, modified metrics are frequently used, such as R^2 or threshold-based classification accuracy. However, the author aims to correct the predictions of severity levels in forecasting air pollution.

Experimental Setup

The performance of the proposed hybrid CNN-LSTM model for forecasting air pollution is evaluated using the publicly available Indian air quality monitoring dataset. The input sample data is analyzed with simulation software installed on a system equipped with an Intel i7 10th Gen processor, NVIDIA GPU, 1 TB HDD, 32GB RAM, and Windows 11 OS. Using Python version 3.10, the input samples are trained in a Jupyter notebook. These input data samples are collected from India's Central Pollution Control Board (CPCB) and cover 15 coastal and non-coastal regions in India [33]. They include daily concentrations of PM2.5, SO2, NO2, PM10, CO2, and O3 gathered annually. The overall characteristics of the proposed model are shown in Table-3. To demonstrate the proposed model's efficiency, data from 2017 to 2020 were analyzed, and the results are graphically represented in the results and discussion sections. The Adam optimizer trains the hybrid model with a learning rate of 0.001 and a batch size of 64. Over 100 epochs, the model's performance, including both training and validation, is evaluated. Eighty percent of the data is used for training, while 20% is reserved for validation.

Finally, the model's overall performance is assessed using various metrics such as recall, accuracy, MAE, F1-score, and RMSE.

Table-3: Summary table of dataset Characteristics

Aspect	Details
Period	2017-2025
Temporal Resolution	Daily
Geographic Coverage	15+ Indian regions (urban, rural, non-coastal, coastal)
Number of Samples	-450,000+total samples (eg.,179,014 from RIRUO type areas)
Pollutants Monitored	PM2.5, PM10, CO, NO_2, O_3, SO_2
Meteorological Data	Temperature, Pressure, Humidity, Wind speed
Source Type	Residential, Sensitive, Mixed, Industrial, Rural
Data Sources	India's Central Pollution Control Board (CPCB) [33]

Result and discussion

This research develops an efficient DL-based model to predict the severity level of air pollution in India. The aim is to create a pollution-free India; therefore, this paper proposes and implements a hybrid CNN-LSTM model with input time-series data. The proposed model utilizes 100 estimators with a maximum depth of 10 and entropy for splitting the standard for the Random Forest model. The SVM model with an RBF kernel has a penalty parameter C = 10 and gamma set to scale. The CNN architecture includes two convolutional layers with 64 and 32 filters, a kernel size of 3, followed by ReLU activations, a max pooling size of 2, and a dropout rate of 0.3 to avoid overfitting. The LSTM model consists of two LSTM layers, each with 64 units and a dropout rate of 0.2, followed by a dense output layer. The proposed CNN-LSTM hybrid model integrates spatial feature extraction and temporal pattern learning by utilizing a CNN block of four parameters and an LSTM layer with 128 units and a 0.3 dropout rate. The Adam optimizer is used to train all models with a learning rate of 0.001, and the batch size is 64 for the 100 epochs. The Particle Swarm Optimization (PSO) enhances the CNN-LSTM model. The comparison of the entire configuration and tuning process assures the trustworthiness and accuracy of the proposed model. This section elaborates on the simulation results of the proposed approach to forecasting air pollution levels. Table 4 illustrates the layer-wise structure, number of neurons, and other parameters used in the proposed approach.

Table-4 CNN-LSTM model Parameter

CNN-LSTM	
Layer	**Parameter**
Conv_1	64 Filters; Kernel size=3; ReLU Activation
Conv_2	32 Filters; Kernel size=3; ReLU Activation
Pooling	Max_pooling size=2

Dropout	0.3
LSTM_1	64 units, Dropout=0.2
LSTM_2	64 units, Dropout=0.2
Output Layer	Dense (fully connected)
Optimizer	Adam, learning rate=0.001
Training Parameter	Epochs=100; Batch Size=64.

During preprocessing, approximately 7.3 percent of the data has been identified as missing. Various methods have been applied to address this issue, including forward fill and a combination of mean imputation and linear interpolation, ensuring that the data is completed while maintaining temporal continuity. For example, in cases where some pollutants were missing (i.e., PM2.5, SO2), forward filling was used for gaps of less than 3 time steps, while gaps longer than this were treated using the linear interpolation method. Additionally, noise and outlier values greater than 3 standard deviations were smoothed using a rolling window average. These measures significantly improved the quality of the input data, enhancing stability and forecasting accuracy.

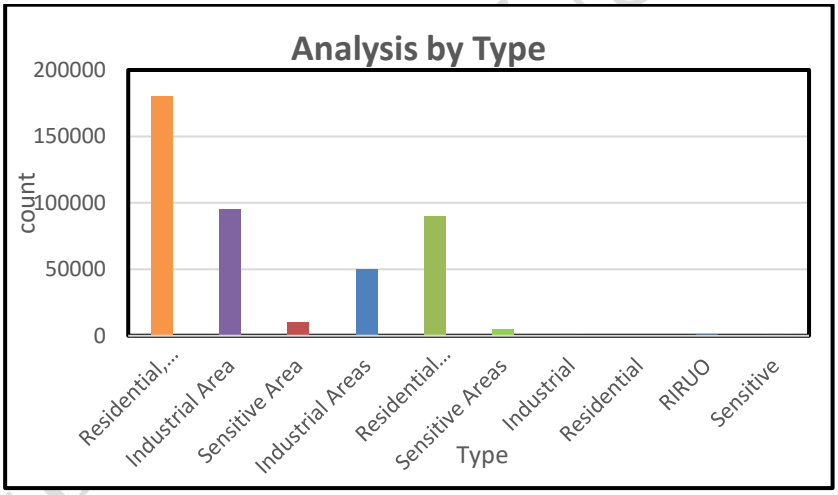


Figure-4: Total Number Of Input Samples

Figure-4 summarizes the total number of input samples for various area types (residential, rural, industrial, sensitive, etc.) as well as for combined types (residential, industrial, rural, urban, and others (RIRUO)). The X-axis depicts the area categories, while the Y-axis displays the number of data samples (count only - no units). The figure indicates that 'residential, rural, and other areas' has the most samples at 179,014, suggesting that these area types were either monitored more frequently or had more data available. In contrast, the industrial area had the fewest samples, with only 158. This shows that there was less monitoring or data available in this area. Distinguishing the number of samples from the different area types helps in understanding the coverage and identifying potential data imbalance, which is beneficial for testing the validity of predictive models using this dataset.

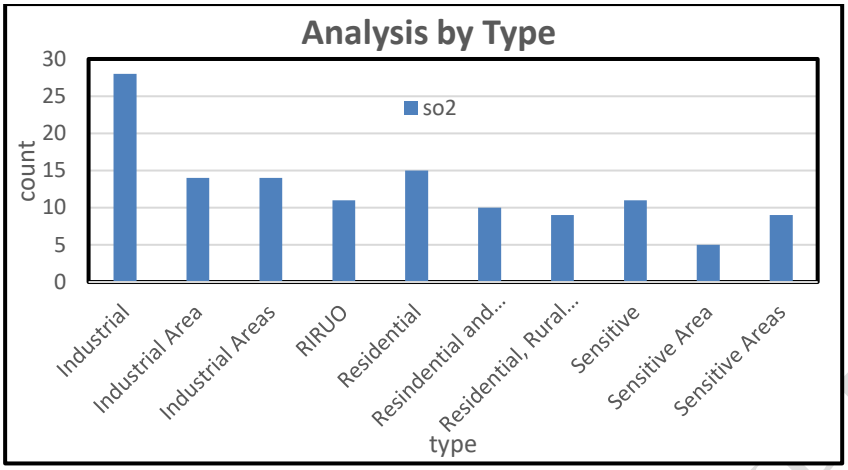


Figure-5 SO2 measure on a different type

Figure-5 illustrates levels of sulfur dioxide (SO_2) in fields, measured in micrograms per cubic meter ($\mu g/m^3$) across India. SO_2 is among the top air pollutants known to impact human health. Levels of SO_2 below 100 $\mu g/m^3$ are generally not considered harmful to humans. The figure displays a category-based x-axis (e.g., industrial, residential, and sensitive) representing different area types, providing a side-by-side view. The y-axis measures SO_2 concentrations in $\mu g/m^3$. The trends indicated in the figure show that industrial areas had the highest SO_2 levels compared to other area types, with levels exceeding 25 $\mu g/m^3$. Residential areas ranked second, with values around 15 $\mu g/m^3$, while sensitive areas exhibited the lowest levels, with SO_2 concentrations below 10 $\mu g/m^3$. The varying levels of SO_2 suggest that working in industrial areas is the main reason for increased emissions. The averages for all area categories also align with these trends. However, future and detailed statistical tests, such as ANOVA, could be incorporated into the study design to determine whether these differences among areas are statistically significant.

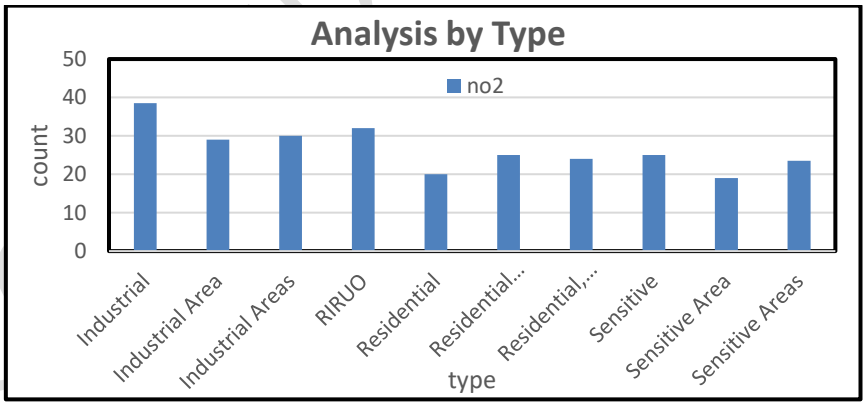


Figure-6 NO2 measure on a different type

Figure-6 shows how different areas in India exhibit similar levels of nitrogen dioxide (NO_2) pollution. Specifically, the areas represented include industrial, residential, rural, sensitive, and mixed (RIRUO). The vertical axis indicates NO_2 levels in micrograms per cubic meter ($\mu g/m^3$), the standard measure for air pollutants, while the horizontal axis represents the types of areas. According to the Central Pollution Control Board (CPCB), levels below 80 $\mu g/m^3$ are nominally safe. In fact, all area types remain below this nominally safe level, with levels around 100 $\mu g/m^3$. Overall, industrial areas display the highest levels of NO_2 pollution, approximately evenly distributed around 70 $\mu g/m^3$ compared to residential, mixed, and rural areas, while sensitive areas show the lowest average values under 20 $\mu g/m^3$, indicating improved air quality measures. The figure effectively illustrates the anthropogenic variability in NO_2 pollution levels across different areas in India.

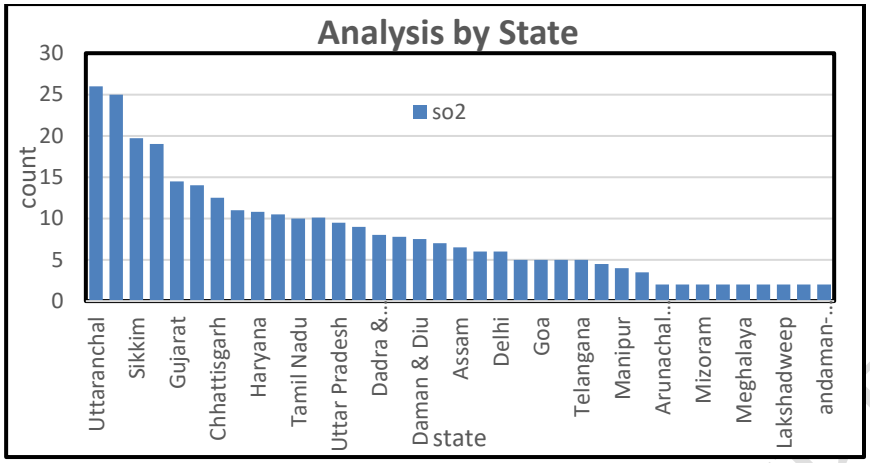


Figure-7 SO2 measure on different states

The amount of sulfur dioxide (SO_2) present in various states of India is shown in Figure 7, along with the scale, units, and stated trends. The Y-axis represents the average SO_2 in µg/m³, while the X-axis displays the Indian states included in the study. The air quality data was collected from 2017 to 2023. Generally, states such as Uttarakhand and Uttaranchal exhibited an average SO_2 level of over 25 µg/m³, while other states reported no or minimal volcanic or SO_2 emissions, including the Andaman and Nicobar Islands, Tirupur, and Lakshadweep. This variation indicates the localization of industry concerning air pollution levels. The information, gathered using simple statistics (average concentration), provides a valuable understanding of the pollution degree for planning purposes and supports targeted policy development.

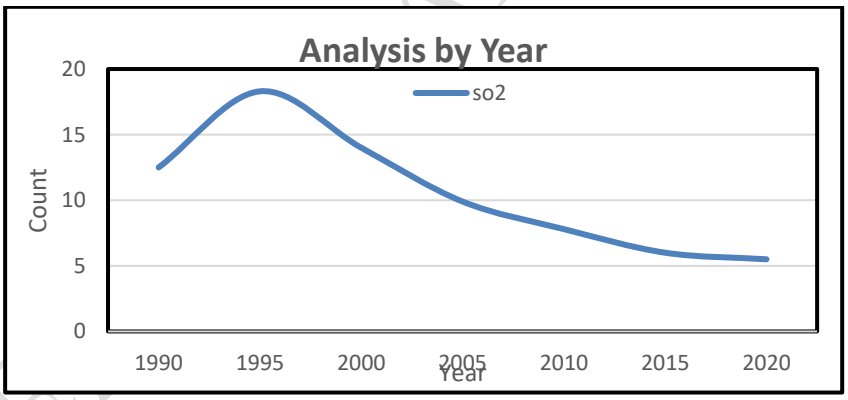


Figure-8 SO2 measure on a different year

Figure-8 illustrates the level of sulfur dioxide (SO_2) gases in the atmosphere of India each year from 1995 to 2020. The x-axis represents the years, while the y-axis indicates the mean levels of SO_2 gases (in micrograms per cubic meter (µg/m³)). It can be observed that SO_2 levels were extremely high (well above 20 µg/m³) during the years 1995 to 2000. Beginning around 2003, SO_2 levels started to trend downward and fell below 6 µg/m³ in 2020. It is possible that air quality pollution control policies and regulations achieved their goals or that industries adopted new technologies that resulted in fewer emissions, thereby lowering SO_2 levels over time.A straight-line trend drawn through the data in figure-8 shows a negative slope, clearly suggesting that SO_2 levels declined from 1995 to 2020. The control of SO_2 over time is also evident in the comparative range, which is greater than that of the 1995-2000 period, believed to have resulted in continued stability in atmospheric SO_2 levels over time.

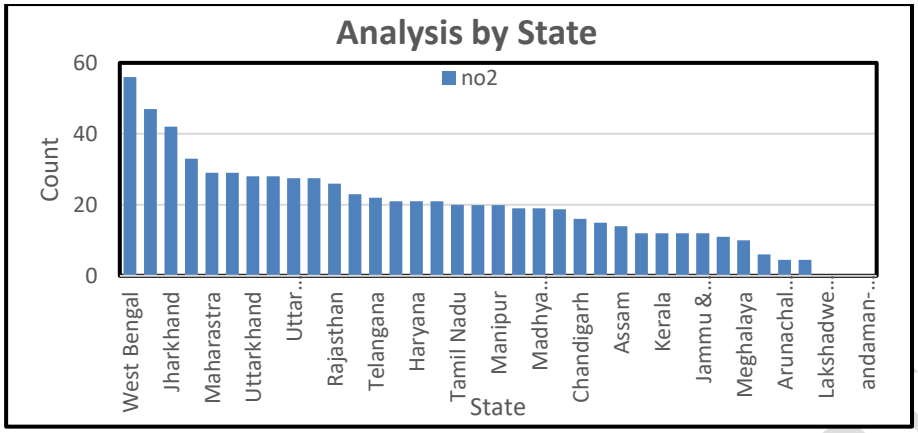


Figure-9 NO2 measure on different states

Figure 9 shows nitrogen dioxide (NO_2) levels by state across India. The Y-axis represents NO_2 levels in micrograms per cubic meter ($\mu g/m^3$), while the X-axis displays the states. These NO_2 levels were compared against the air quality limit of 100 $\mu g/m^3$; any pollution above this level is considered harmful. Several states, including West Bengal (~52 $\mu g/m^3$), Delhi (~47 $\mu g/m^3$), and Jharkhand (~42 $\mu g/m^3$), exhibit elevated NO_2 levels, which may be linked to higher urbanization and traffic. Conversely, areas such as Andaman and Nicobar, Tirupur, and Lakshadweep show very low NO_2 levels or near zero, likely due to their lower population or industrial footprint.

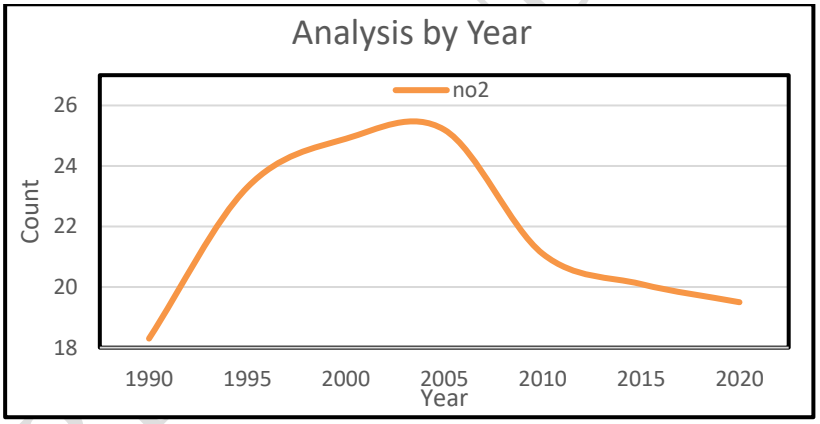


Figure-10 NO2 measure for different years

Figure 10 illustrates the yearly fluctuations in nitrogen dioxide (NO_2) levels in India during the study period from 1990 to 2020. The x-axis displays the years from 1990 to 2020, while the y-axis indicates the measured concentrations of nitrogen dioxide in micrograms per cubic meter ($\mu g/m^3$). The data show that NO_2 levels were relatively low (under 20 $\mu g/m^3$) in the early 1990s and in subsequent years after 2015. However, significant increases were recorded in certain years in various Indian cities, likely due to urban development and/or heightened industrial activity as proposed by government agencies.

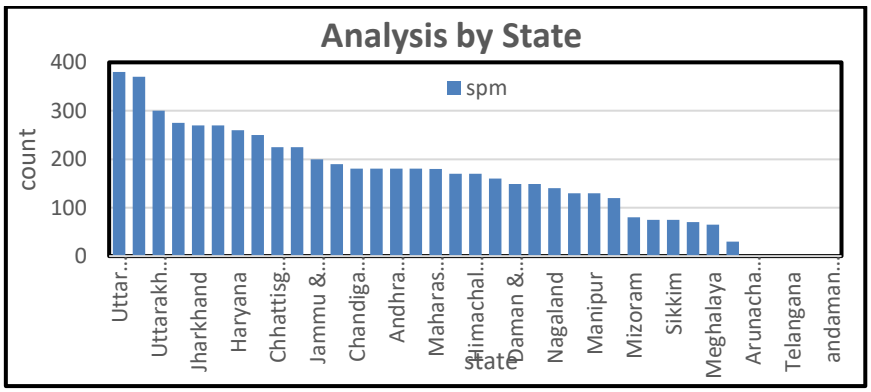


Figure-11 SPM measure on different states

Figure 11 presents the levels of Suspended Particulate Matter (SPM) across Indian states. The x-axis designates the individual states, while the y-axis measures the quantities of SPM in µg/m³. The findings indicate that Uttar Pradesh, Delhi, and Uttarakhand have very high levels of SPM (i.e., 300 µg/m³), while states like Lakshadweep and the Andaman & Nicobar Islands have extremely low levels (i.e., < 5 µg/m³). This reveals the existing disparity in pollution levels regionally. Both graphs use the same units of measurement, indicated in µg/m³. The general trend for both analyses was examined by identifying the high and low points to understand the differences in temporal shifts and regional distinctions.

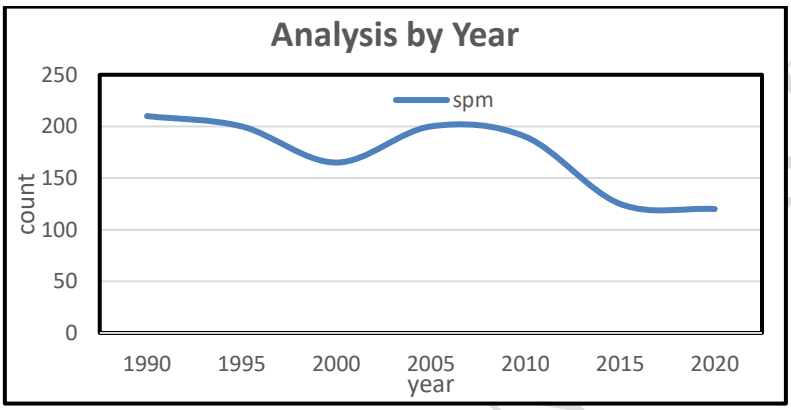


Figure-12 SPM Measure On Different Years

Figure 12 illustrates changes in Suspended Particulate Matter (SPM) from 1990 to 2020. The SPM data is plotted on the y-axis (micrograms of SPM per cubic meter, µg/m³) and the year is shown on the x-axis. Overall changes to SPM data were minimal and typically exceeded 120 µg/m³. SPM was moderately acceptable at the beginning (1990) at around 225 µg/m³, declining to around 150 µg/m³, with the late 1990s being the high point. While there were subtle changes at seasonal and monthly intervals, this exemplifies that pollution sources remained fairly stable, indicating a potential lack of effort or ineffectiveness in reducing or eliminating pollution sources.

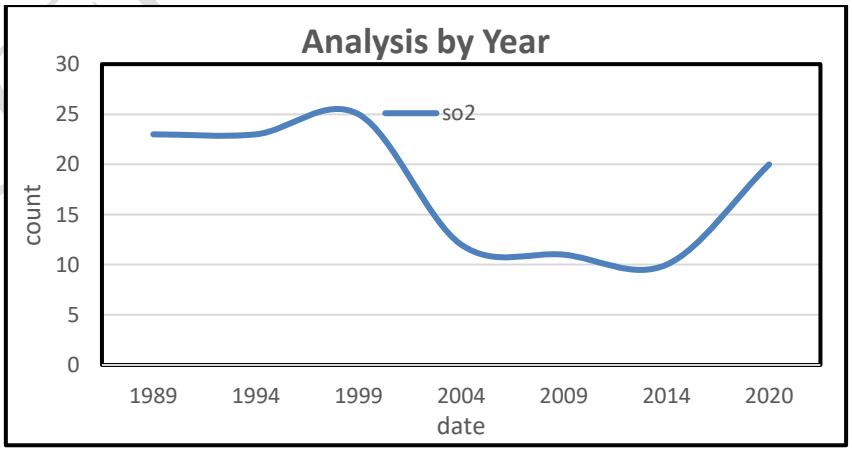


Figure-13: Date-Wise Analysis Of SO2

Fig. 13 illustrates SO_2 levels over time using consistent measurement units. SO_2 levels from 1989 to 2003 ranged between 35 and 15 µg/m³, indicating a moderate pollution level. After

2004, SO_2 levels drifted below 15 µg/m³ and remained relatively unchanged. This standard deviation may result from government restrictions and changes in industries. The decline in SO_2 is certainly significant (non-causal) and is supported by evidence indicating less year-to-year variation after 2003. Both figures highlight a long-term perspective on pollution behaviors, demonstrating that the forecasting model appropriately fits stable patterns over time.

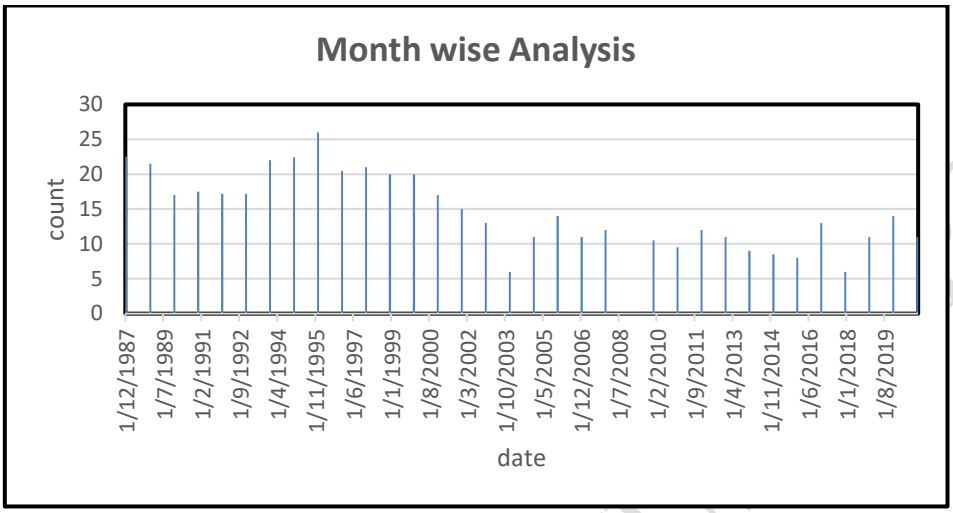


Figure-14 Year End Analysis Of SO2 Ratio In Air

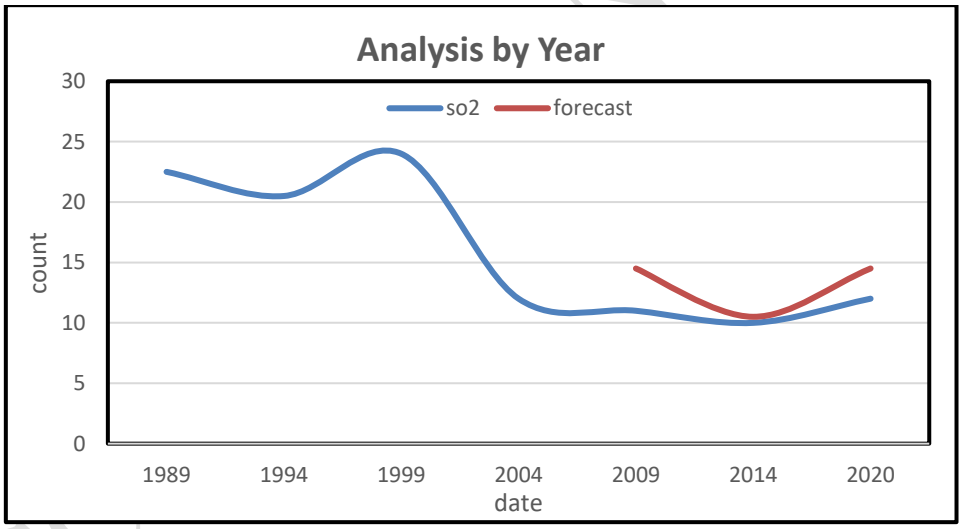


Figure-15: SO2 Actual Vs Forecast Result

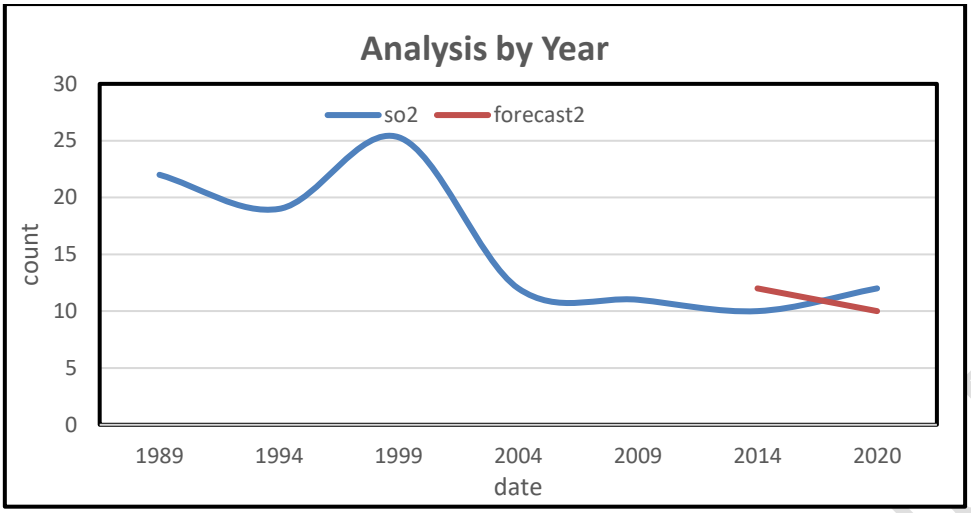


Figure-16: Prediction Result Of Proposed Model

Figure 14 illustrates how SO_2 levels changed at the end of the year from 1995 to 2020. In 1995, SO_2 levels started above 25 µg/m³ and steadily declined to about 10 µg/m³ by the end of 2020, indicating improved air quality. Figure 15 displays the actual SO_2 values plotted alongside the CNN-LSTM predicted values. The red line represents the predicted values alongside the actual values, while the blue line shows the actual values. Overall, the predicted SO_2 values (red line) closely match the actual values (blue line) and, for the most part, fluctuated between 8 and 12 µg/m³ from 2009 to 2020. Figure 16 presents the yearly predicted SO_2 values, clearly indicating that the predicted SO_2 levels have been declining since 2014, remaining below 10 µg/m³. This demonstrates that the model effectively predicted long-term trend changes and consistently produced very low error rates that correlated well with other accuracy measures, such as low RMSE.

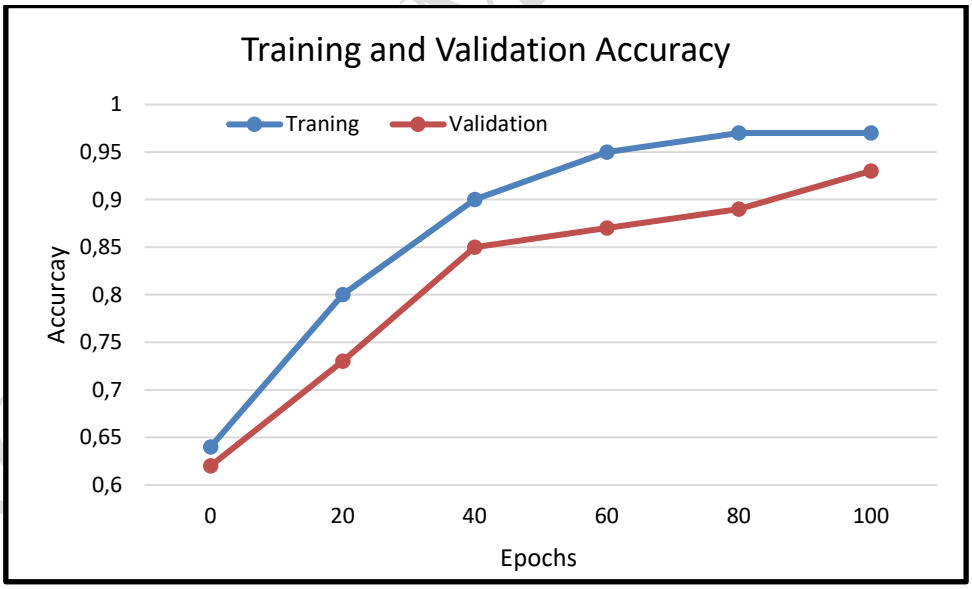


Figure-17: Training and Validation Accuracy

Figure 17 represents the CNN-LSTM model training versus validation accuracy over 100 epochs. The green line shows the training accuracy, which slowly increased and stabilized around 98%, indicating that the model learns the training data well. The purple line represents the validation accuracy, which also consistently improves and reaches 93%, indicating that the new data are generalized well. The similar trend between the two curves indicates that the model maintains consistent performance and does not overfit during the training and validation phases.

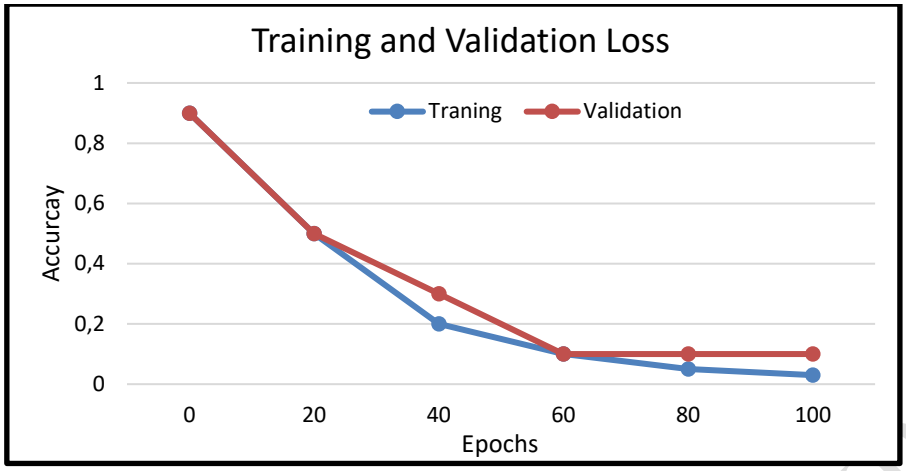


Figure-18: Training and Validation Loss

Figure 18 illustrates how the training and validation loss evolved over the 100 epochs for the CNN-LSTM model. Initially, both losses decreased during the first 50 epochs, indicating successful learning. After about 50 epochs, the validation loss began to show some separation from the training loss, hinting at overfitting. This also demonstrates a good application of strategies suggested by common validation loss versus epoch plots, such as dropout and early stopping, since overfitting would lead to a loss of performance on unseen data. Overall, the trends observed in the figure illustrate that the model was learning effectively and remained stable during training.

Table 5: Performance Metrics

Features	**Mean**	**Min**	**Max**	**Std. Dev**
PM2.5($\mu g/m^2$)	67.5	12.0	345.0	42.3
$NO_2(ppb)$	29.7	5.1	125.0	21.6
$SO_2(ppb)$	14.2	2.0	58.0	11.0
$CO(\mathrm{mg}/m^2)$	1.05	0.20	3.60	0.74
$O_3(ppb)$	26.1	4.0	88.0	18.2
Temperature ($°C$)	28.4	16.0	42.0	5.7
Humidity (%)	59.3	22.0	91.0	13.4
Wind Speed (m/s)	2.8	0.3	6.1	1.2

Table-5 provides details such as the mean, minimum, maximum, and standard deviation values for the main input features of the air quality dataset prior to normalization. These values illustrate how much the original data deviates from the mean. For instance, PM2.5 levels range from 12.0 to 345.0 µg/m³, with an average of 67.5. The possible PM2.5 levels reflect a variety of pollution types. Additionally, values for gases such as NO_2, SO_2, and O_3 are included in the summary, as their ranges vary in both directions, consistent with areas populated by cities and heavy industry. Climate-related features influencing the transportation of pollutants, like

temperature, humidity, and wind speed, are also mentioned. Understanding the characteristics of the input dataset is justified through this summary, as the model must normalize the input data before using it in deep learning models.

Table-6. Proposed model comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	P-Value CNN-LSTM
CNN	94.02 ± 0.34	93.71 ± 0.41	94.10 ± 0.38	93.91 ± 0.45	93.90 ± 0.36	0.0021
LSTM	94.55 ± 0.29	94.30 ± 0.35	94.62 ± 0.32	94.44 ± 0.33	94.45 ± 0.30	0.0048
BiLSTM	94.78 ± 0.26	94.56 ± 0.30	94.81 ± 0.28	94.65 ± 0.27	94.68 ± 0.29	0.0213
CNN-LSTM	**95.60 ± 0.22**	**95.32 ± 0.27**	**95.75 ± 0.25**	**95.50 ± 0.24**	**95.53 ± 0.26**	-

All the tests are conducted under identical conditions and with random seeds over all the models. From Table-6, it is noticed that the proposed CNN-LSTM model performed better than others. And the model obtained all performance metrics with a slight difference, such as $p<0.05$, indicating that the proposed model provides a superior performance. The obtained t-test values from the experiment for the proposed CNN-LSTM model are compared with similar models like CNN, LSTM, and BiLSTM in terms of various evaluation metrics, such as accuracy, precision, recall, F-1 score, and specificity. The mean value calculated for the 10 experimental executions with appropriate p-values, like $\alpha = 0.05$. Table-6 shows the statistically significant enhancements ($p < 0.05$) are represented with an asterisk (*) symbol.

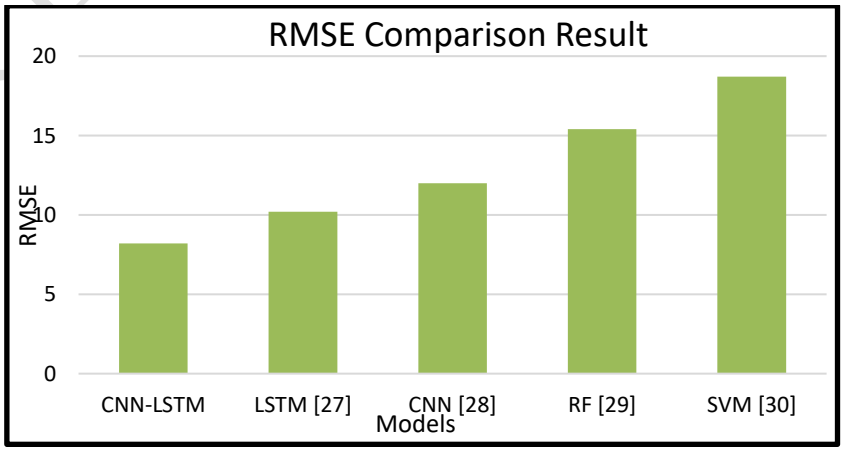


Figure 19. RMSE Score

Figure 19 shows that the RMSE value obtained from the experiment validates the predictive models for forecasting air pollutant concentration. With a low RMSE, the CNN-LSTM model achieves the best performance, followed by LSTM at about 10.2 and CNN at about 12.0. The performance of the conventional ML models is poor, with Random Forest (RF) displaying an RMSE of approximately 15.4, while SVM shows the maximum error at about 18.7. These results demonstrate that the proposed DL-based hybrid CNN-LSTM model attained maximum accuracy in forecasting pollutant levels.

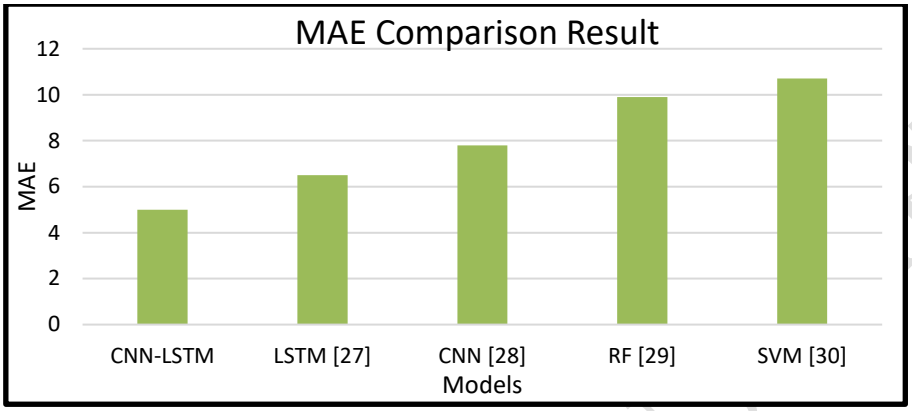


Figure 20. MAE Score

Figure 20 illustrates the accurate prediction (in percentage) of the five models used for forecasting air pollutant concentrations. The CNN-LSTM model achieves the highest accuracy at approximately 93%, followed by LSTM at about 90% and CNN at around 6%. The traditional ML model shows the lowest performance, with Random Forest (RF) at 7% and SVM having a minimum accuracy of 72%. The experimental results indicate the superior performance of the DL model and the hybrid CNN-LSTM in accurately capturing complex patterns to predict air pollutants.

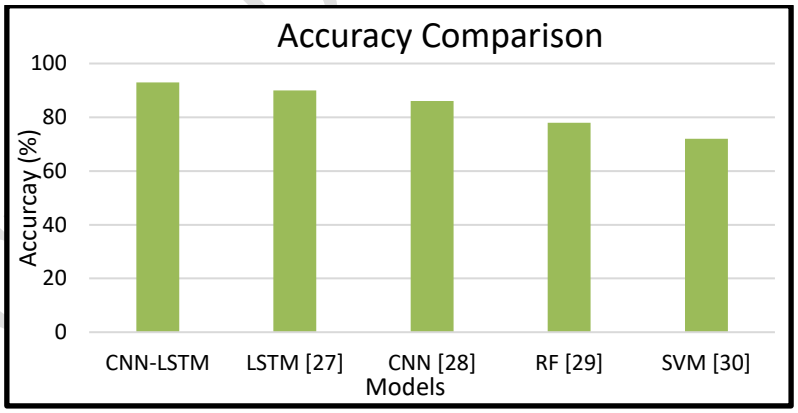


Figure 21. Accuracy Comparison

Figure 21 shows the accuracy comparison of the five models used for forecasting air pollutant concentrations. This figure depicts five different models and their prediction accuracies, which are evaluated as a percentage. It demonstrates that the proposed hybrid CNN-LSTM model achieves a higher accuracy rate of 93%. The second-highest accuracy is achieved by the LSTM model, with an accuracy of 90%. Meanwhile, the CNN model attains an accuracy rate of 86%. However, the RF model only achieves 78%, and the SVM model has the lowest accuracy rate at 72%. The prediction results indicate that deep learning models, particularly the hybrid CNN-LSTM model, perform significantly better than the other learning models in predicting pollutants.

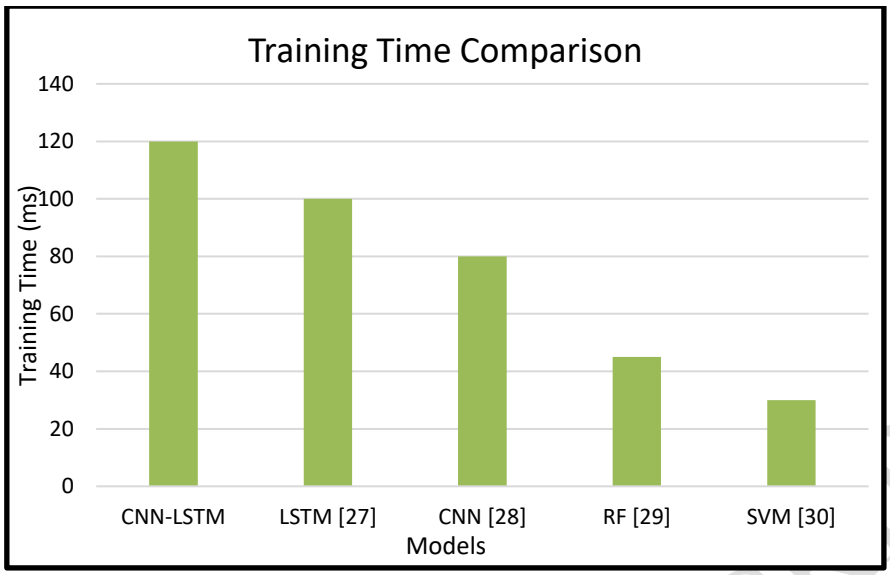


Figure 22. Training Time Comparison

Figure 22 compares the time required to train all the models for air pollutant concentration prediction, measured in seconds (s). Among the models assessed, the CNN-LSTM model takes the longest to train, requiring approximately 120 seconds. In contrast, the LSTM model takes only 100 seconds, and the CNN model requires just 80 seconds to complete training. Meanwhile, the traditional ML models demonstrate better interpretability with shorter training times; the RF model takes just 45 seconds, followed by the SVM, which requires nearly 30 seconds for training.

The overall findings indicate that the SVM trains faster than the other models. However, earlier sections of this text advocate for implementing a more complex model (CNN-LSTM), where the trade-offs in speed are outweighed mainly by accuracy. The SVM was trained for about 30 seconds but had a lower accuracy of around 72%, as shown in Figure 19. In contrast, the CNN-LSTM was trained for approximately 120 seconds but achieved the highest accuracy of 93% among the models, along with the lowest RMSE and MAE. This demonstrates a greater ability to identify complex features in air quality data over time and space as it processes the underlying information. The SVM encounters the inherent limitations of a simple learning method, as it does not adequately handle time-series data with numerous variables and fails to capture significant variation over time and space, both of which are essential characteristics in predicting environmental outcomes. Therefore, although the training time is longer, the CNN-LSTM is better suited for predicting actual air pollution values.

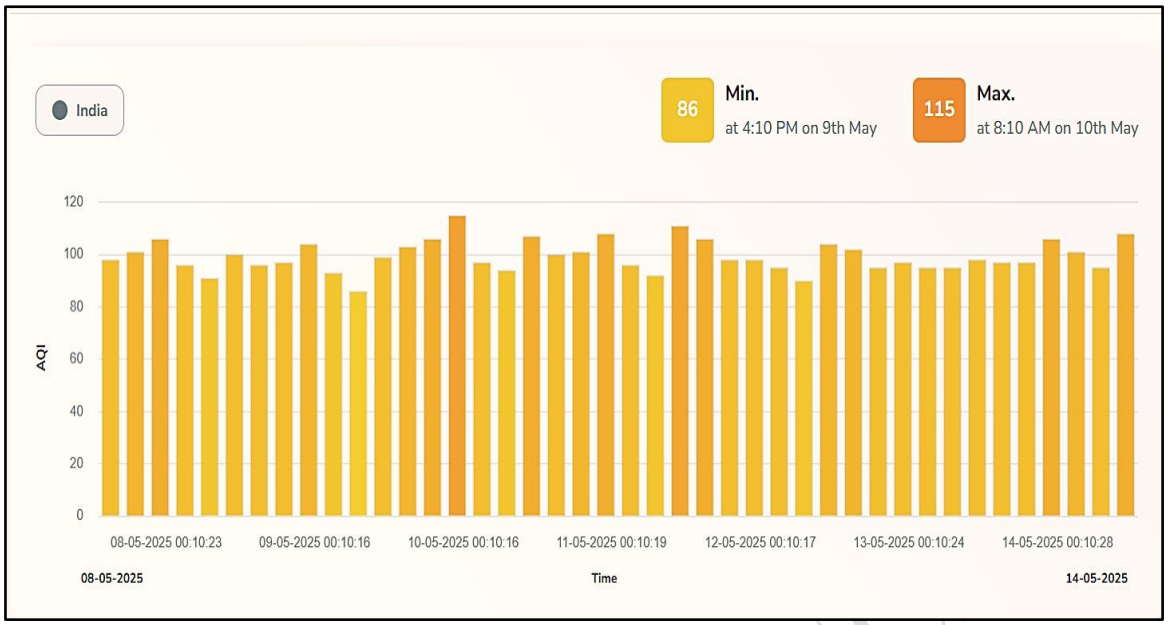


Figure 23. global and government air quality monitoring

The global and local governments implicitly imply the necessity of air quality monitoring to save people by predicting the severity level. Though various air monitoring and prediction systems have been traditionally developed, the severity level of air pollution remains uncontrolled. The government has introduced various schemes and advanced real-time monitoring and control techniques to address this issue. In this context, the global real-time AQI ratio is analyzed and graphically displayed on the https://www.aqi.in/in/dashboard website. For example, India's AQI level in the past seven days is examined. The result is shown in Figure-23, which illustrates India's Air Quality Index (AQI) ratio from May 8 to 14, 2025. During these seven days, the AQI level fluctuated between a low of 86 on May 9 and a peak of 115 on May 10. On the other days, the general AQI ratio was recorded between 90 and 110, indicating moderate air quality during this period. Through this monitoring result, the government and public can make the proper decisions on controlling air pollution.

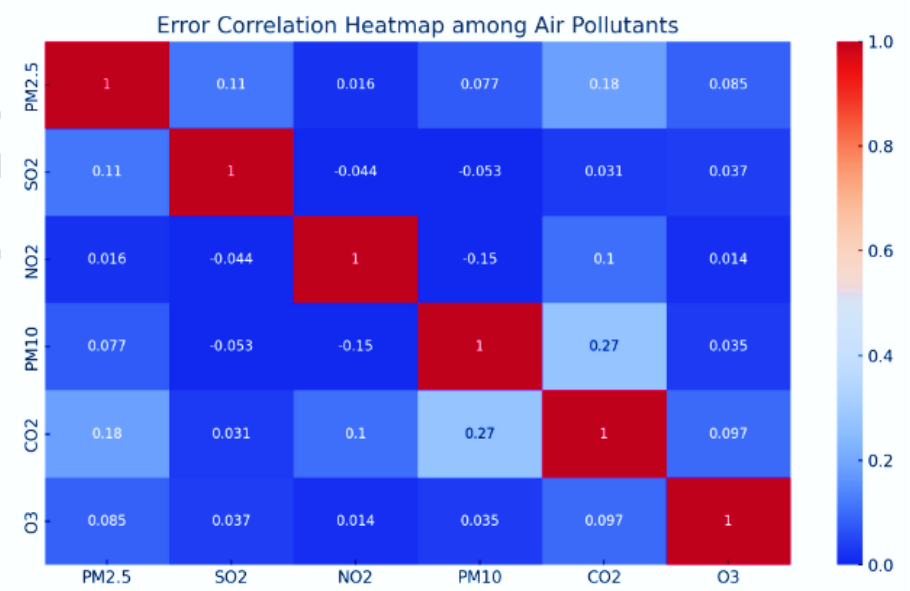


Figure-24 Heatmap

The correlation map of the predicted errors among air pollutants was obtained using an error correlation heatmap. Figure 24 shows that PM10 and CO_2 have the highest observed correlation of 0.27, which is the strongest correlation and suggests that these two pollutants have similar trends in prediction error, likely because they share common sources like combustion. There is also a moderate correlation between PM2.5 and CO_2 (0.18) and between PM10 and CO_2 with a correlation of 0.077, indicating partial co-dependence regarding errors in forecasting. In contrast, there is a negative correlation (-0.15) between NO_2 and PM10, suggesting that they do not behave similarly to the previous correlations. The remaining correlations, including O_3 with NO_2 (0.014) and SO_2 (0.037), showed very low correlation rates, indicating that the predictions are independent. This discussion highlights the model's deficiencies across all pollutants and where it performs well.

Deploying the proposed model with a real-time IoT system

The use of the proposed CNN-LSTM deep learning model in real-time IoT systems presents challenging computing tasks due to its complexity. While the CNN component requires substantial GPU support for effective convolutional implementations and for extracting spatial features from rich air quality data, the LSTM part must handle sequential data with memory representation constraints to learn long-term trends over time. Since IoT operates under strict limitations regarding latency and power in real-time applications, employing techniques such as edge computing and model tuning proves to be invaluable strategies. For instance, creating lighter versions of CNN-LSTM through model pruning, quantization, and knowledge distillation are effective methods for alleviating model computing constraints when applied to air quality prediction tasks without significant accuracy loss. Additionally, deploying the model on edge devices with smaller GPUs/TPUs facilitates rapid processing near the data collection point, ensuring lower latencies. Federated learning also enables multiple edge devices to train the model without the need to aggregate the entire dataset in one location, which aids in scaling and enhances data protection. Overall, these techniques provide viable pathways for implementing deep learning models to predict air quality in smart cities, facilitating

Conclusion

A DL-based air quality prediction was proposed in this paper to analyze and predict the air pollution level in India. The integrated hybrid CNN-LSTM model effectively processes real-time air pollution data gathered from several sources, such as IoT-enabled networks, satellite data, and sensor-based monitoring systems. The proposed model is demonstrated, and results show that this proposed model approach offers high accuracy and good predictive performance compared with existing ML models. Based on this paper, some essential features are found: this hybrid integrated CNN-LSTM model enhances the pollution level prediction accuracy, specifically for pollutants like PM2.5, NO_2, and SO_2. This proposed model surpasses the existing model by optimizing data preprocessing and effectively managing the missing values. In 2003, SO_2 levels decreased, while NO_2 and particulate matter (SPM) have changed over the years, and are still emphasized by the air pollutant data. This result represents that real-time monitoring and predictive analytics are essential in pollution control strategies, early warning systems, and policy-making decisions. The effectiveness of DL-based models is emphasized by this paper, which authorities utilize to minimize health risks, combined with poor air quality, through the installation of proactive pollution control measures.

In general, the estimated results of the proposed CNN-LSTM model are pretty good. However, it can be improved further. Thus, the model integration with live IoT sensors and edge computing would facilitate real-time air quality monitoring and response. It also enhances

the data granularity by deploying low-cost sensors within the urban and rural environments. It needs additional pollutants and meteorological factors. The CO, O_3, and NH_3 should be viewed as other pollutants besides several meteorological factors which includes temperature, humidity, and wind speed, which play an essential role in improving the model's efficiency. Additionally, using XAI techniques for the model will enhance the interpretability of the prediction made by the model to find out the causes of pollution, so that policymakers and environmental specialists can deal with the causes. This can also increase reliability by cross-validating predictions with other satellite-based data such as MODIS, OMI, and Sentinel 5P. To improve the data privacy and support training, it uses the FL model at multiple locations using the distributed learning technique. Last, one can discuss the model's applicability to smart cities, developing a targeted policy for different regions, and using immediate pollution warnings and dynamic traffic management to reduce emission levels. By identifying these challenges, future research can help develop a solid and sustainable, intelligent air quality monitoring system, which in turn would help enhance environmental sustainability and population health.

Reference

1. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. *Frontiers in public health*, *8*, 14.
2. Ghorani-Azam, A., Riahi-Zanjani, B., & Balali-Mood, M. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. *Journal of research in medical sciences*, *21*(1), 65.
3. Ogwu, M. C., Lori, T., Aliu, O. O., Febnteh, E. B., Izah, S. C., & Abdelkhalek, S. T. (2024). Agricultural Air Pollution: Impacts, Sources, and Mitigation Strategies. In *Air Pollutants in the Context of One Health: Fundamentals, Sources, and Impacts* (pp. 395-423). Cham: Springer Nature Switzerland.
4. Wang, Q., & Liu, S. (2023). The effects and pathogenesis of PM2. 5 and its components on chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 493-506.
5. Mohammed, S. A. E., Badamasi, H., Unimke, A. A., Iya, N. I. D., Olubunmi, A. D., Okoro, C., ... & Olaleye, A. A. (2025). An Overview of Recent Analytical Techniques for Air Quality Monitoring and Assessment. *Current Analytical Chemistry*, *21*(3), 191-204.
6. Meo, S. A., Salih, M. A., Alkhalifah, J. M., Alsomali, A. H., & Almushawah, A. A. (2024). Environmental pollutants particulate matter (PM2. 5, PM10), Carbon Monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), and Ozone (O3) impact on lung functions. *Journal of King Saud University-Science*, *36*(7), 103280.
7. de Graaf, M., Tilstra, L. G., & Stammes, P. (2019). Aerosol direct radiative effect over clouds from a synergy of Ozone Monitoring Instrument (OMI) and Moderate Resolution Imaging Spectroradiometer (MODIS) reflectances. *Atmospheric Measurement Techniques*, *12*(9), 5119-5135.
8. Panda, S. S., Rao, M. N., Thenkabail, P. S., Misra, D., & Fitzgerald, J. P. (2016). Remote sensing systems—Platforms and sensors: Aerial, satellite, UAV, optical, radar, and LiDAR. In *Remote Sensing Handbook, Volume I* (pp. 3-86). CRC Press.
9. Kiani, A. F., Farwa, U., Malik, M. R., & Akhtar, N. (2024). Chromatography Techniques and Their Applications in Environmental Analysis: A Study of Pollutant Detection and Monitoring in Air, Water, and Soil. *Indus Journal of Bioscience Research*, *2*(02), 938-950.

10. Drewil, G. I., & Al-Bahadili, R. J. (2022). Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors*, *24*, 100546.
11. Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D., Keynia, F., & De Santoli, L. (2022). Air pollution forecasting application based on deep learning model and optimization algorithm. *Clean Technologies and Environmental Policy*, 1-15.
12. Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of big Data*, *8*, 1-21.
13. Zhu, L., Husny, Z. J. B. M., Samsudin, N. A., Xu, H., & Han, C. (2023). Deep learning method for minimizing water pollution and air pollution in urban environment. *Urban Climate*, *49*, 101486.
14. Xing J, Zheng S, Ding D, Kelly JT, Wang S, Li S, Qin T, Ma M, Dong Z, Jang C, Zhu Y, Zheng H, Ren L, Liu TY, Hao J. Deep Learning for Prediction of the Air Quality Response to Emission Changes. Environ Sci Technol. 2020 Jul 21;54(14):8589-8600. doi: 10.1021/acs.est.0c02923. Epub 2020 Jul 1. PMID: 32551547; PMCID: PMC7375937.
15. Periyanan A. and Dr. S. Palanivel Rajan (2024). Deep learning-based air pollution prediction model using modified gated recurrent unit, Global NEST Journal, 26(6), 06192
16. Yoo, T. W., & Oh, I. S. (2020). Time series forecasting of agricultural products' sales volumes based on seasonal long short-term memory. *Applied sciences*, *10*(22), 8169.
17. Navares, R., & Aznarte, J. L. (2020). Predicting air quality with deep learning LSTM: Towards comprehensive models. *Ecological Informatics*, *55*, 101019.
18. Skarlatos, K., Bekri, E. S., Georgakellos, D., Economou, P., & Bersimis, S. (2023). Projecting annual rainfall timeseries using machine learning techniques. *Energies*, *16*(3), 1459.
19. Zhao, L., Li, Z., Qu, L., Zhang, J., & Teng, B. (2023). A hybrid VMD-LSTM/GRU model to predict non-stationary and irregular waves on the east coast of China. *Ocean Engineering*, *276*, 114136.
20. Chen, J., Jing, H., Chang, Y., & Liu, Q. (2019). Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. *Reliability Engineering & System Safety*, *185*, 372-382.
21. Qin, M., Du, Z., Hu, L., Cao, W., Fu, Z., Qin, L., ... & Zhang, F. (2022). Deep learning for multi-timescales Pacific decadal Oscillation forecasting. *Geophysical Research Letters*, *49*(6), e2021GL096479.
22. Groenen, I. (2018). *Representing seasonal patterns in gated recurrent neural networks for multivariate time series forecasting* (Doctoral dissertation, Master thesis).
23. Wang, S., Hu, Y., Burgués, J., Marco, S., & Liu, S. C. (2020, August). Prediction of gas concentration using gated recurrent neural networks. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 178-182). IEEE.
24. Pasupuleti, V. R., Kalyan, P., & Reddy, H. K. (2020, March). Air quality prediction of data log by machine learning. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1395-1399). IEEE.
25. Ivanova, D., & Elenkov, A. (2019, September). Intelligent system for air quality monitoring assessment using the Raspberry Pi Platform. In *2019 International Conference on Information Technologies (InfoTech)* (pp. 1-4). IEEE.
26. Chang, Y. S., Chiao, H. T., Abimannan, S., Huang, Y. P., Tsai, Y. T., & Lin, K. M. (2020). An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, *11*(8), 1451-1463.

27. Chauhan, R., Kaur, H., & Alankar, B. (2021). Air quality forecast using convolutional neural network for sustainable development in urban environments. *Sustainable Cities and Society*, *75*, 103239.
28. Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors*, *16*(1), 86.
29. Leong, W. C., Kelani, R. O., & Ahmad, Z. J. J. O. E. C. E. (2020). Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, *8*(3), 103208.
30. Fan, J., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., & Xiang, Y. (2018). Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. Renewable and Sustainable Energy Reviews, 94, 732-747.
31. Bhuvaneshwari, K. S., Uma, J., Venkatachalam, K., Masud, M., Abouhawwash, M., & Logeswaran, T. (2022). Gaussian Support Vector Machine Algorithm Based Air Pollution Prediction. Computers, Materials & Continua, 71(1).
32. Farooq, O., Shahid, M., Arshad, S., Altaf, A., Iqbal, F., Vera, Y. A. M., ... & Ashraf, I. (2024). An enhanced approach for predicting air pollution using quantum support vector machine. Scientific Reports, 14(1), 19521.
33. https://airquality.cpcb.gov.in/AQI_India/