Archives of Environmental Protection
Vol. 50 no. 4 pp. 9–21

PL ISSN 2083-4772
DOI 10.24425/aep.2024.152891

Air quality prediction using stacked bi- long short-term memory and convolutional neural network in India

Karkuzhali S[1*], Thendral Puyalnithi[2], Nirmalan R[2]

[1]Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India
[2]Department of Artificial Intelligence and Data Science,Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India

* Corresponding author`s e-mail: karkuzhali@mepcoeng.ac.in

Keywords: Air Quality, Bi-Long Short-Term Memory, Convolutional Neural Network, Adam Optimizer, training process, hybrid analysis, pollution

Abstract: Air quality is a critical aspect of environmental health, and its assessment and prediction serve as pivotal components in mitigating the adverse effects of air pollution. This study focuses on advancing air quality prediction in India through the application of cutting-edge deep learning techniques, specifically the Stacked Bidirectional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN) architecture. Through meticulous preprocessing - encompassing missing value handling, normalization, and temporal sequencing - the dataset is prepared for the Stacked Bi-LSTM and CNN hybrid model. The model architecture leverages the temporal sequence-capturing capabilities of Stacked Bi-LSTM layers, enhancing it with the spatial feature extraction prowess of CNN layers. This integrated approach aims to address the intricate and non-linear dependencies present in air quality time series data. During the training phase, the Adam optimizer is used to fine-tune the model's hyperparameters, with Mean Squared Error (MSE) serving as the loss function. Important assessment metrics, including as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and MSE, are used to evaluate the performance of the model. Furthermore, this study conducts a detailed temporal analysis, unraveling diurnal, seasonal, and long-term trends in air quality fluctuations.

The study aims to offer valuable insights into the temporal and spatial patterns of air quality in India, thereby aiding environmental policymakers, urban planners, and researchers in formulating effective strategies for air quality management. The application of Stacked Bi-LSTM and CNN architectures in this research holds promise for enhancing real-time forecasting accuracy and facilitating informed decision-making towards sustainable environmental practices.

Introduction

Air quality is a critical component of environmental health, and its assessment and forecasting are vital for public welfare. Rapid urbanization, industrialization, and other anthropogenic activities have significantly contributed to air pollution, making it a major global concern. India, being one of the world's most populous and economically dynamic countries, faces substantial challenges in managing and mitigating air quality issues. The period from 2010 to 2023 have witnessed significant changes in India's socio-economic landscape, accompanied by shifts in industrial activities, energy consumption patterns, and urban development. Understanding the spatiotemporal dynamics of air quality is crucial for developing effective policies, urban planning, and public health initiatives.

Traditional methods of air quality prediction often fall short in capturing the intricate patterns and non-linear dependencies present in time series data. In response to these challenges, this study explores advanced deep learning techniques, specifically employing Stacked Bidirectional Long Short-Term Memory (Bi-LSTM) and CNN architectures, to conduct a comprehensive Time Series Analysis of Air Quality in India. Traditional air quality forecasting models typically rely on statistical or shallow machine learning techniques. While effective to some extent, these methods often struggle with the complex, non-linear, and dynamic nature of air quality data. The motivation behind this research stems from the need for more sophisticated models that can unravel intricate patterns in the time series data, providing accurate and adaptive air quality predictions. Stacked Bi-LSTM and CNN architectures are integrated to provide a comprehensive solution to air quality forecasting by capturing both geographical features and temporal dependencies.

The following are the main goals of this study: (1) to perform an extensive time series analysis of India's air quality over the years 2010–2023; (2) to develop and implement a Stacked

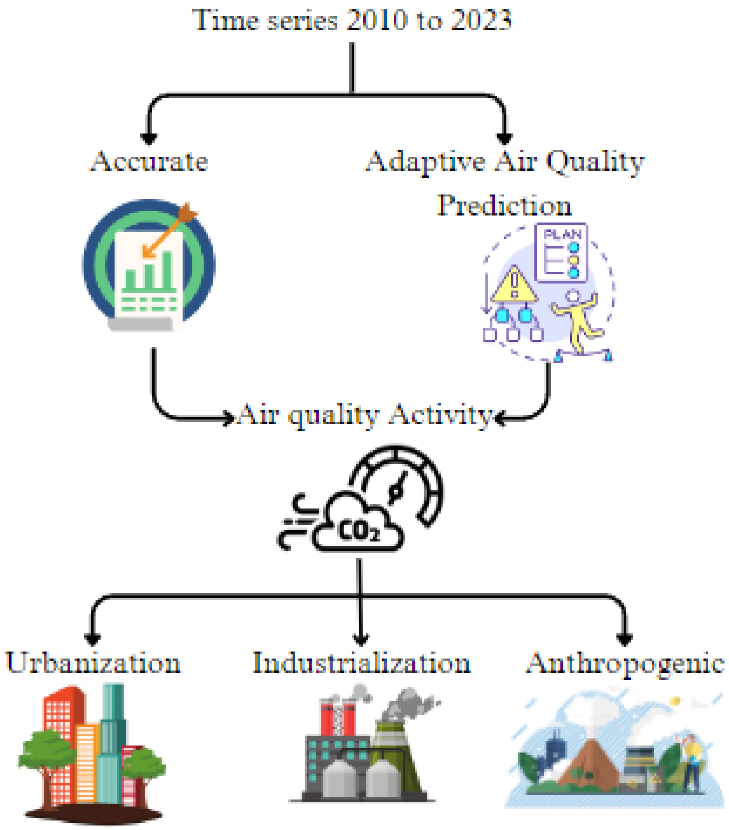

Figure 1. Diagram of air quality activities

Bi-LSTM and CNN hybrid model for air quality forecasting, considering parameters such as PM2.5, PM10 concentrations, and the Air Quality Index (AQI) as demonstrated in Figure 1; (3) to evaluate the performance of the proposed model using standard metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

This research focuses on the spatial and temporal dynamics of air quality in India, acknowledging the complexity of the factors influencing air pollution. By adopting deep learning techniques, this work aims to contribute valuable insights into the intricacies of air quality changes, laying a foundation for more accurate and adaptable forecasting models. This investigation is significant as it can help environmental scientists, urban planners, and policymakers make well-informed judgments and strategies to address India's air quality problems.

Related Works:

Lu et al.(2023) propose a ground-breaking deep learning model called Inception-ResNet, which integrates residual connections with Gate Recurrent Units (GRU) and Inception architectures. This model processes environmental component sequences relevant to dam deformation and extracts multi-scale features using modified Inception-ResNet blocks with spatial attention and channel attention modules.

Zhang et al.(2023) investigate GATBL-Learning, a deep learning-based model for predicting air pollution diffusion trends. This model uses a graph attention network to aggregate and extract spatial data from cities, while also mining time series characteristics and seasonality discipline to enhance adaptability.

Matthaios et al.(2023) outline a novel approach to assessing the exposure of cars to air pollution, utilizing measurements of particulate matter, nitrogen dioxide, NOx, aerosol lung surface deposited area, and ultrafine particles, alongside ambient air quality. This approach estimates pollutant concentrations and collects fine-grained road data using machine learning algorithms combined with mass-balance concepts. This study evaluates the framework's ability to gather pollutant data, scale-up storage, and translate high-level user queries into SQL. The results provide foundational data for future research aimed at improving forecasting models for roadside and highway air quality.

Janarthanan et al.(2021) performed preprocessing on the Chennai City AQI dataset to eliminate duplicate entries and missing values. The AQI values were then classified using a deep learning model that combined SVR and LSTM. The improved prediction accuracy facilitated by this model contributed to sustainable development planning for metropolitan area by providing timely information on AQI levels.

Yu et al.(2023) developed a model architecture to anticipate the sensor-based interior temperature in the Australian urban environment using a variety of machine learning (ML) and deep learning (DL) techniques. The study aims to provide an improved deep ensemble machine learning framework (DEML). Data was collected from 96 devices between August 2019 and November 2022, encompassing low-cost sensor-based indoor environmental metrics, satellite-derived outdoor climate features, and ambient station-based temperature readings.

Kanmani et al.(2022) provide a deformable active contour model that is energy-efficient and can be used to track the growth of Cryptogams, a bioindicator known for its sensitivity to pollution levels. The model facilitates precise pollution monitoring by tracking the vegetative development of Cryptogams over two weeks. Filter stacks for the model are constructed using the deep convolutional neural network architecture known as the VGG 16.

Wood et al.(2022) assessed the predictive performance of eleven climatic variables as part of the Combined Local Area Benchmark (CLAB). Discrepancies were found between the 2019 and 2020 projections when evaluating the predictive capabilities of nine machine learning algorithms and three deep learning algorithms. A thorough analysis of prediction outliers revealed that consistent CLAB predictions are often challenging when ground-level meteorological data is sparse or unreliable.

Prado-Rujas et al.(2024) propose an artificial intelligence framework that uses convolutional long short-term memory networks to estimate the levels of eleven pollutants in a Region of Interest (ROI). The framework is capable of simultaneously managing multiple pollutants, handling sensor inputs, recovering missing data, and using different modules for external input data. An hourly dataset spanning ten years was used to train and validate the system.

Fu et al. (2023) highlight the difficulties of environmental preservation in real-world air quality monitoring. Jurado et al. (2022) evaluate the dispersion of airborne pollutants in metropolitan settings, noting that sensors and models perform well. Although computationally expensive and unsuitable for real-time application, computational fluid dynamics (CFD) models are widely used for such analyses. Deep learning methods have been developed to address these limitations, with the MultiResUnet architecture being the most successful model.

Shin et al. (2023) addressed the limitations of deep neural network topologies and conventional computational methods by introducing a fully convolutional network (FCN)-based deep learning regression model. This technique enables rapid prediction of the mean age of air (MAA) without compromising geographic information. The model reduces the prediction error by 43.14% and 34.77%, respectively.

Wang et al. (2024) captured spatial-temporal visual characteristics, such as PM2.5, PM10, and AQI, by integrating a convolutional neural network (CNN) with long short-term memory (LSTM). The model performs better in predicting air quality at night, but its predictions during the day are less accurate. Yadav et al. (2023) developed a globally scalable approach for monitoring air quality (AQ) in low- and middle-income countries (LMICs). In high-income countries (HICs) with sufficient ground data, the approach employs transfer learning to adapt a deep learning model that maps satellite images to AQ. The trained model can explain up to 54% of the AQ distribution. Iskandaryan (2023) proposed machine learning approaches for air quality prediction based on traffic, meteorological, and air quality data from Madrid. The study employs both graph-based and grid-based algorithms, showing promise for future forecasting.

Cao et al. (2023) propose a hybrid approach to simultaneously predict seven air quality indices from different monitoring sites. The model predicts a matrix series of various indicators using extended ARIMA, EMD, and SVG techniques. According to experimental results, the model outperforms more sophisticated models in terms of both accuracy and time complexity.

Shao et al. (2023) presents a novel coupled air quality optimization prediction model based on the Dung Beetle Optimization (DBO) technique, Extreme Gradient Boosting (XGBoost), the Informer time series technique, and Variational Mode Decomposition (VMD). The coupling methodology decomposes time series data, classifies different feature data based on approximation entropy, optimizes VMD with DBO, and uses the Spearman coefficient method to screen important features. To generate predicitons, the Informer algorithm and DBO-optimized XGBoost process various feature data independently, then combine and reconstruct the projected values. The new model performs better when applied to Nanjing air quality prediction (R-squared=0.961, RMSE=1.988, MAE=1.624).

Zhang et al. (2019) examine large-scale data and forecast data to evaluate air quality using the LightGBM model. They compare data from 35 Beijing air quality monitoring stations and extract temporal features using a sliding window technique. The experimental results demonstrate the effectiveness of this approach.

Liu et al. (2019) discussed that although previous studies have supported the use of seq2seq for air quality prediction, two issues remain unresolved. To address these issues, the n-step recurrent prediction approach was suggested. Liu et al. (2023) developed an enhanced extreme learning machine (GA-KELM) prediction approach based on genetic algorithms to successfully address this challenge. The kernel matrix replaces the hidden layer output matrix, and the number of hidden nodes and layers is optimized using a genetic algorithm to mitigate the issue of deteriorating learning capacity in conventional limit learning machines.

Iskandaryan et al. (2023) presents a method that combines a Gated Recurrent Unit (GRU), a Graph Convolutional Network (GCN), and Attention, which they refer to as the Attention Temporal Graph Convolutional Network. This model

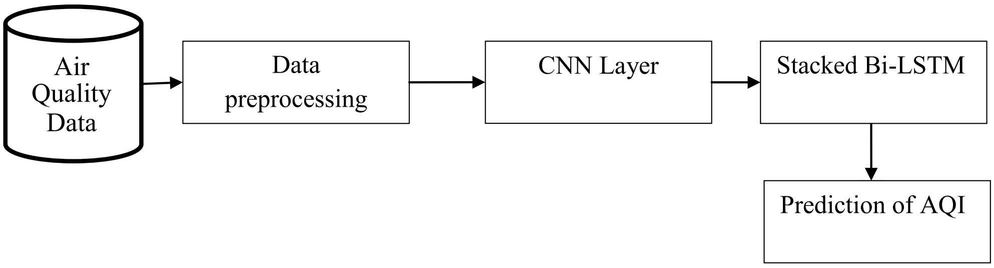

Figure 2. System Design diagram

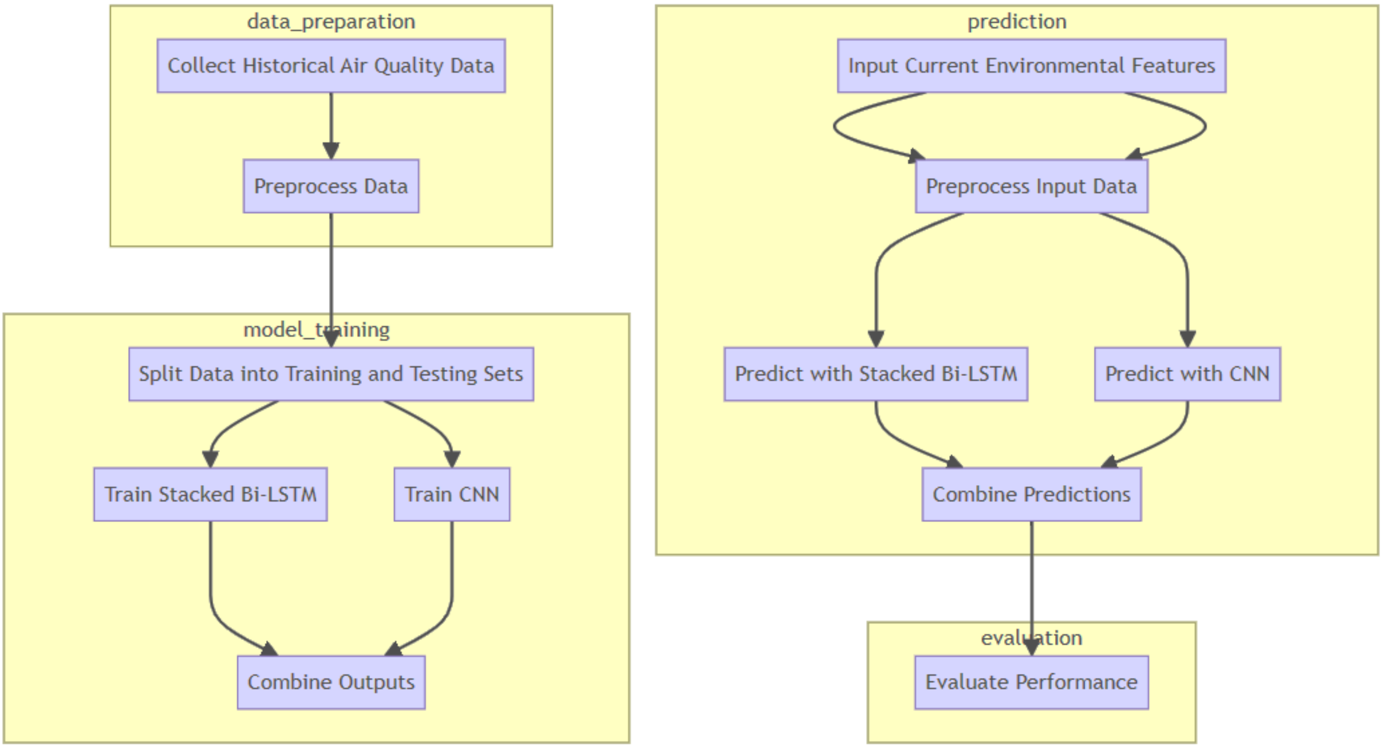

Figure 3. System Architecture

is proposed for predicting air quality. Metrics like the Pearson Correlation Coefficient and Mean Absolute Error were used to evaluate the proposed model against air quality data from Madrid for the periods of January to June 2022 and January to June 2019. The results showed that the model performed better during these periods. Yang et al. (2022) used source data derived from meteorological condition datasets that measure temperature, humidity, and atmospheric pressure, as well as air pollutant datasets that include PM2.5, PM10, and SO2 hourly concentrations. Air quality prediction models, namely the Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models, were developed for four distinct scenarios. The Shapley Additive Explanation (SHAP) approach was utilized to assess the explainability of these models.

This study found that considering only meteorological factors does not improve forecast accuracy. On the other hand, prediction accuracy is higher when meteorological conditions are combined with other air contaminants. Wardana et al. (2023) proposes an affordable air quality monitoring solution. The tool operates two small ML models on a single microcontroller, and the model imputer effectively predicts data when the missing rates are below 80%. By offering in-depth research utilizing state-of-the-art deep learning techniques over an extensive temporal period, this study seeks to fill a gap in the literature and address the shortcomings of previous studies that focus on specific contaminants or traditional forecasting methods.

Czech et al. (2020) introduced artificial neural networks, fuzzy logic, and evolutionary algorithms, alongside various mathematical and statistical tools. They identified exhaust gas emissions as the primary cause of environmental damage resulting from the rise of motorization.

Karthikeyan et al (2023) presents an air pollution prediction model for smart environment design planning based

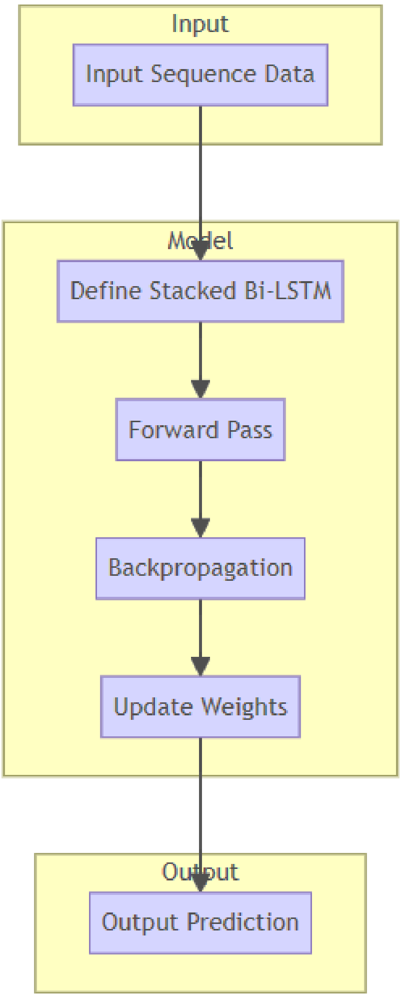

Figure 4. Stacked Bi-LSTM

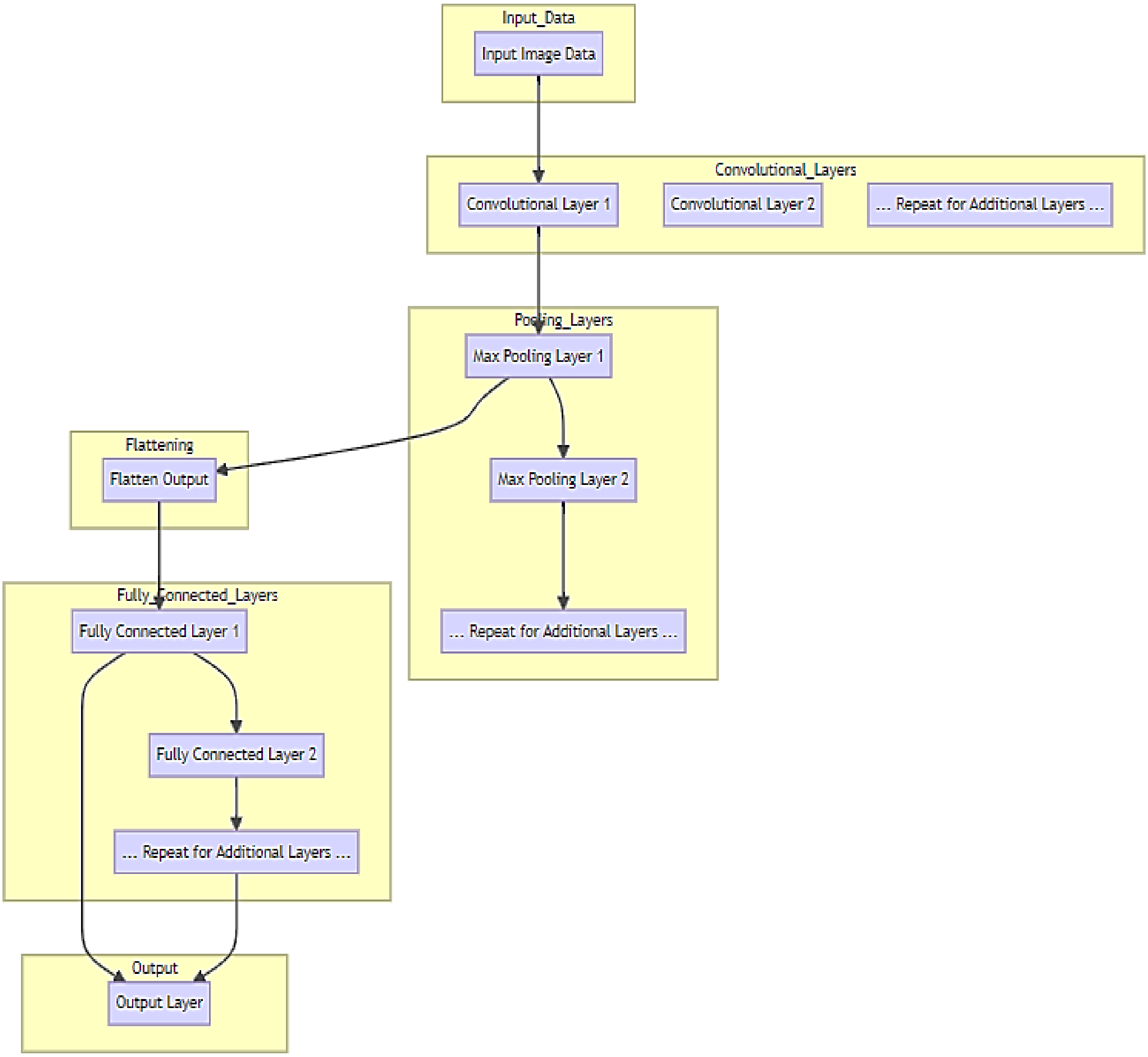

Figure 5. Convolution Neural Networks

on deep learning (DLAPPSEDP). This technique incorporates air pollution prediction, hyperparameter tweaking, and data pre-processing. For optimal tuning, it utilizes the atomic orbital search optimization (AOSO) algorithm and a graph convolutional network model. Comprehensive experimental analysis demonstrates the efficacy of this method and its improvements over other recent approaches.

Methodology

Collection of data:

The dataset contains air quality for Indian cities from 2010 to 2023, covering 453 cities across 31 states. The data provides detailed information on the air quality status of these locations. As it was collected by the Central Control Room for Air Quality Management, the dataset is considered reliable and trustworthy.

Data preprocessing

In data preprocessing, several crucial steps are undertaken to ensure the integrity and compatibility of the air quality dataset. First and foremost, missing values are handled using robust techniques such as imputation or removal, preventing potential distortions in the analysis. Subsequently, to facilitate optimal model convergence during training, the data undergoes normalization through Min-Max scaling, which scales the values to a consistent range, typically between 0 and 1. Finally, the temporal dynamics of the data are carefully considered by structuring it into sequences suitable for time series analysis. This involves segmenting the dataset into specific time windows, a crucial step for capturing temporal dependencies essential for accurate air quality predictions. These preprocessing steps collectively lay the foundation for a robust and reliable analysis of air quality trends using the

Table 1. Training Process

Sl.no	Training process	
	Framework	**Specialization**
1	Data Splitting	Temporal Order: Divide the historical air quality dataset into training, validation, and testing sets while respecting the temporal order. Ensure that the training set covers earlier time periods, and the testing set represents more recent data.
2	Loss Function and Optimizer	Mean Squared Error (MSE): Utilize the Mean Squared Error as the loss function. This measures the average squared difference between predicted and actual air quality values, emphasizing accurate prediction of both high and low values.
		Adam Optimizer: Implement the Adam optimizer, which adapts the learning rate during training, balancing the speed of convergence and avoiding overshooting.
3	Hyperparameter Tuning	Learning Rate: Experiment with different learning rates to find an optimal balance between rapid convergence and avoiding convergence issues.
		Batch Size: Adjust the batch size, considering computational resources. Smaller batches may provide a more accurate gradient, but larger batches can expedite training.
		Number of Epochs: Determine the appropriate number of epochs to prevent overfitting or underfitting. Utilize techniques like early stopping to halt training when the validation loss reaches an optimal point.
4	Model Architecture Fine-Tuning:	Iterative Process: Fine-tune the number of Stacked Bi-LSTM layers, CNN layers, and units within each layer through an iterative process. Experiment with different configurations to find the architecture that best captures the temporal and spatial patterns in the air quality data.
5	Regularization Techniques	Dropout: Implement dropout layers to prevent overfitting by randomly dropping out units during training. This enhances the model's generalization capabilities.
		L2 Regularization: Apply L2 regularization to penalize large weights in the network, promoting a more robust and generalized model.
6	Cross-Validation	Time Series Cross-Validation: Use time series-specific cross-validation techniques, such as TimeSeriesSplit, to evaluate the model's performance across different time periods. This ensures that the model generalizes well to unseen data.
7	Monitoring and Adjustments:	Validation Monitoring: Regularly monitor the model's performance on the validation set during training. Implement mechanisms like early stopping to prevent overfitting and ensure the model does not learn noise in the data.
		Adjustments: Make necessary adjustments to hyperparameters or model architecture based on the observed performance during training.

Stacked Bi-LSTM and CNN model. Figure 2 illustrates the block diagram of the proposed system.

Stacked LSTM & CNN

Recurrent neural network (RNN) architectures, such as stacked bidirectional long short-term memory networks (Bi-LSTM), are designed to identify and understand sequential patterns and dependencies in input data. The term “stacked” refers to the use of multiple layers of Bi-LSTM units, creating a deeper network that can learn increasingly complex representations (Figure 3).

The vanishing gradient issue is addressed in Figure 4, enabling the identification of long-term dependencies in sequential data. Bidirectional LSTMs enhance the model’s ability to capture information from both past and future time steps by processing input sequences in both forward and backward directions. The process of stacking multiple layers of bidirectional LSTM units is referred to as stacked Bi-LSTM. With increased network depth, each layer is able to learn hierarchical features and representations, facilitating the recognition of complex patterns and relationships in sequential data. Lower layers capture simpler and more localized patterns, while higher layers abstract more complex and intricate features. Due to its flexibility and capacity to automatically learn hierarchical representations, the stacked Bi-LSTM architecture is well-suited for tasks that require a nuanced understanding of sequential input. Its ability to capture both short- and long-term dependencies makes it particularly effective for applications such as natural language processing and time series prediction.

The depth of CNN network can be adjusted based on the complexity of the data and the patterns that need to be detected, as illustrated in Figure 5. In addition to analyzing air quality data, it also records sequential patterns and long-term dependencies. Mathematically, the hidden states ht for the forward and backward LSTMs at time t can be expressed as follows in Equation (1 and 2):

$$htf = LSTMforward\ (ht - 1f,\ xt) \tag{1}$$

$$htb = LSTMbackward\ (ht + 1b,\ xt) \tag{2}$$

Here, htf and htfb represent hidden states for the forward and backward LSTMs, respectively. The bidirectional LSTM then

Table 2. Temporal analysis

Sl.no	Temporal Analysis	
	Framework	**Specialization**
1	Diurnal Patterns	Identification: Analyse the model's ability to capture diurnal patterns, such as daily fluctuations in air quality.
		Visualizations: Utilize visualizations, such as line plots or heatmaps, to depict how well the model aligns with observed diurnal variations. Evaluate its accuracy in predicting peak and off-peak air quality levels throughout the day.
2	Seasonal Trends	Seasonal Analysis: Examine the model's performance in capturing seasonal variations in air quality over the course of the year.
		Comparison: Compare model predictions with observed seasonal trends, considering factors like temperature, humidity, and other environmental variables. Assess the model's adaptability to different seasons
3	Long-Term Dynamics	Trend Analysis: Investigate the model's ability to capture long-term trends in air quality data over multiple years.
		Correlation Studies: Conduct correlation studies between predicted and actual long-term trends. Assess the model's robustness in recognizing and adapting to gradual changes in air quality.
4	Event-Based Analysis:	Extreme Events: Analyse the model's response to extreme events, such as pollution spikes or unusual weather conditions.
		Performance During Events: Evaluate the model's performance during specific events and assess its reliability in predicting air quality deviations under non-standard conditions.
5	Statistical Metrics	Metrics Over Time: Calculate and analyze performance metrics (e.g., Mean Squared Error, RMSE) over different time periods.
		Temporal Changes in Accuracy: Assess how model accuracy changes over time, identifying periods of improved or degraded performance.
6	Dynamic Adaptability	Model Adaptation: Examine how well the model adapts to changes in air quality dynamics over time. Reactivity: Evaluate the model's reactivity to sudden shifts in air quality, reflecting its responsiveness to real-time changes.

combines these two hidden states to obtain the final hidden state ht at each time step, as shown in Equation (3):

$$ht = [htf, htb] \qquad (3)$$

This bidirectional flow facilitates the capturing of information from both past (t−1) and future (t+1) time steps, enhancing the model's ability to understand the temporal dynamics of the air quality data. For the CNN layers, the spatial features are extracted using convolutional operations. The convolutional operation can be mathematically represented as shown in Equation (4):

$$(X * W)i, j = \sum m = 0M - 1\sum n = 0N - 1Xi + m, j + \mathrm{n} \cdot Wm, n \quad (4)$$

Here, X represents the input data (multidimensional air quality data), and W denotes the convolutional kernel. The term (X∗W)i,j denotes the result of the convolution operation at the position (i,j). This operation is applied across the entire input, allowing the model to effectively capture spatial patterns and relationships among the various pollutants.

Integration of Stacked Bi-LSTM and CNN:

Merge layers: The outputs from the Stacked Bi-LSTM layers and the CNN layers are combined using merge layers to leverage both temporal and spatial information.

Dense layers: Fully connected dense layers are added to map the combined features to the final output layer for air qu ality predictions.

The Dense layers incorporate fully connected neurons to map the extracted features to the final output, predicting air quality parameters. The output Y of a dense layer, given an input X and weights W can be expressed by Equation (5):

$$Y = \sigma\,(X \cdot W + b) \qquad (5)$$

Here, σ represents the activation function (e.g., ReLU or sigmoid), b is the bias term, and X·W denotes the dot product between the input and weight matrices. This final stage in the model architecture ensures a seamless transformation of complex spatial and temporal information into accurate predictions for air quality parameters.

Training process

The training process for predicting air quality using Stacked Bi-LSTM and CNN model involves several key steps designed to ensure effective learning from historical data to enable the model to generalize well for accurate predictions. The steps involved in the training process are summarized in Table 1:

Temporal Analysis

Temporal analysis is a critical component in assessing the performance and capabilities of a model predicting air quality

Table 3. Spatial Analysis

Sl.no	Spatial Analysis	
	Framework	**Specialization**
1	Regional Variations:	Geographical Distribution: Assess how well the model captures regional variations in air quality across different geographical locations in India. Spatial Maps: Utilize spatial maps and heatmaps to visualize predicted vs. actual air quality parameters, highlighting areas of agreement and disparities.
2	Localized Patterns	Microscale Analysis: Investigate the model's ability to capture localized variations in air quality at a microscale level. hibit higher or lower accuracy in predictions.
3	Spatial Correlations	Inter-Pollutant Relationships: Examine the model's capability to recognize spatial correlations among different pollutants. Correlation Heatmaps: Generate correlation heatmaps to visualize spatial relationships and identify areas where the model may perform exceptionally well or encounter challenges.
4	Geospatial Visualization	GIS Integration: Integrate Geographic Information System (GIS) tools for geospatial visualization of air quality predictions. Overlay Analysis: Overlay predicted air quality values onto geographical maps, providing a visual representation of model performance across regions.
5	Temporal-Spatial Patterns	Time-Space Analysis: Investigate how well the model captures temporal-spatial patterns, considering both time-dependent and location-dependent variations. Animated Visualizations: Create animated visualizations to observe how air quality predictions evolve spatially over time.
6	Outlier Detection	Anomaly Detection: Implement techniques for detecting spatial anomalies or outliers in the model predictions. Identification of Deviations: Identify regions where the model may consistently overpredict or underpredict air quality parameters.
7	Sensitivity to Geographic Features	Land Use and Features: Assess the model's sensitivity to different land use and geographic features. Incorporate Spatial Covariates: Integrate additional spatial covariates that may influence air quality, such as land cover, traffic density, or industrial zones.

Table 4. Regression Model Performance Metrics

Algorithm	RMSE Training Data	RMSE Test Data	R-Squared value on train	R-Squared value on test
Linear Regression	13.583	13.672	0.9849	0.9847
Random Tree Regressor	0.4141	1.1508	0.9999	0.9998

Table 5. Classification Algorithm Performance Metrics

Algorithm	Model accuracy on train	Model accuracy on test	Kappa Score
Logistic Regression	0.7276	0.7271	0.5843
Decision Tree Classifier	1.0	0.9998	0.999
Random Forest Classifier	1.0	0.9998	0.9997
K-Nearest Neighbours	0.9981	0.9967	0.9951

using Stacked Bi-LSTM and CNN. The detailed steps of the temporal analysis process are summarized in Table 2:

Spatial Analysis

Spatial analysis is an essential component in evaluating the effectiveness of a model predicting air quality using Stacked Bi-LSTM and Convolutional Neural Network (CNN). The detailed steps of the spatial analysis process are outlined in Table 3:

Results and Discussion

R-squared and RMSE are commonly used metrics for evaluating regression models. It is recommended to use a combination of these metrics to comprehensively assess the model's performance. In this case, the RMSE metric was considered because it not only calculates the average distance between the predicted and actual values but also emphasizes the impact

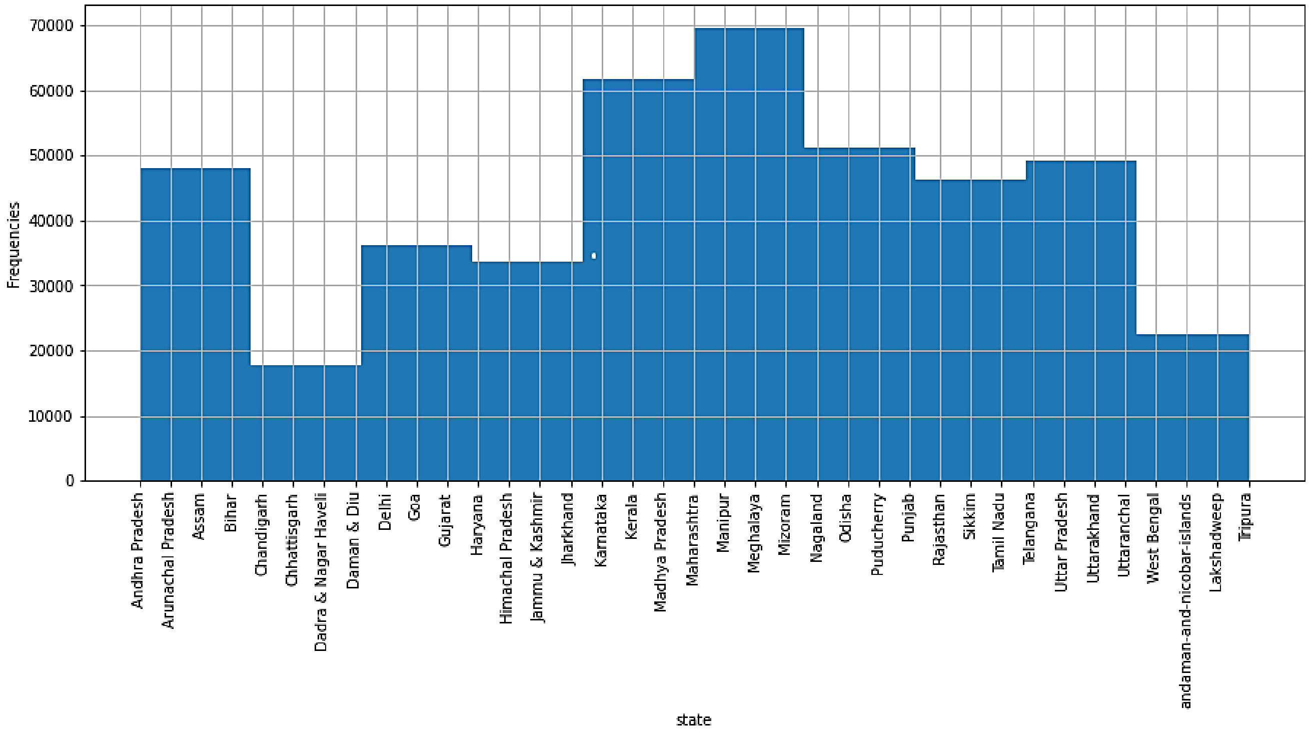

Figure 6. The visualization shows us the count of states present in the dataset.

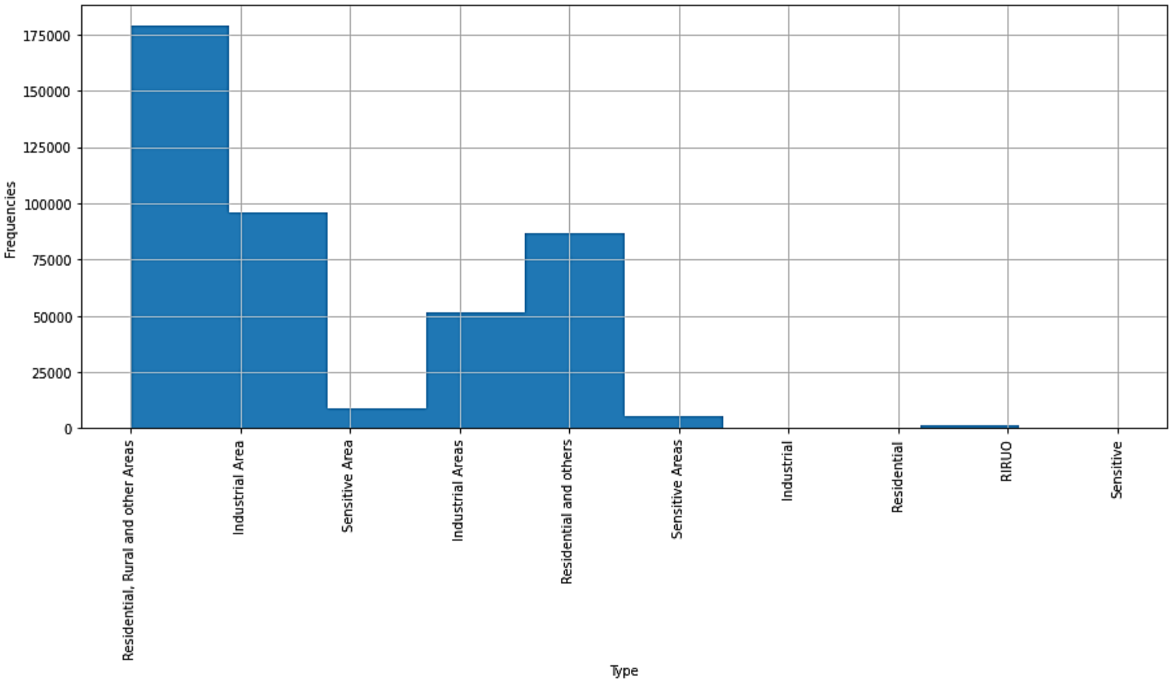

Figure 7. The visualization shows us the count of Types present in the dataset.

of significant errors, as larger errors have a greater influence on the RMSE value. To achieve better results, two regression algorithms were tested: Random Tree Regression and Linear Regression. The training and testing sets were evaluated using RMSE and R-squared metrics for both algorithms. Linear regression resulted in RMSE values of 13.583 for training and 13.672 for testing. In contrast, Random Tree Regression yielded higher Rsquared values of 0.9999 for training and 0.9998 for testing. Table 4 provides a detailed explanation of these findings.

Convolutional Neural Network (CNN) techniques were used to analyze the air quality frequency level in 37 states for the purpose of testing. The states of Maharashtra, Manipur, Meghalaya, and Mizoram recorded the highest frequency levels, reaching 69,000 MHz. In contrast, Chandigarh, Chattisgarh, Dadra & Nagar Haveli, and Daman & Diu exhibited the lowest frequency levels, ranging between 10,000 MHz and 20,000 MHz. Among the 37 states analyzed, 11 states had frequency levels between 40,000 and 52,000 MHz. Figure 6 provides a detailed visualization of the dataset for the selected 37 states.

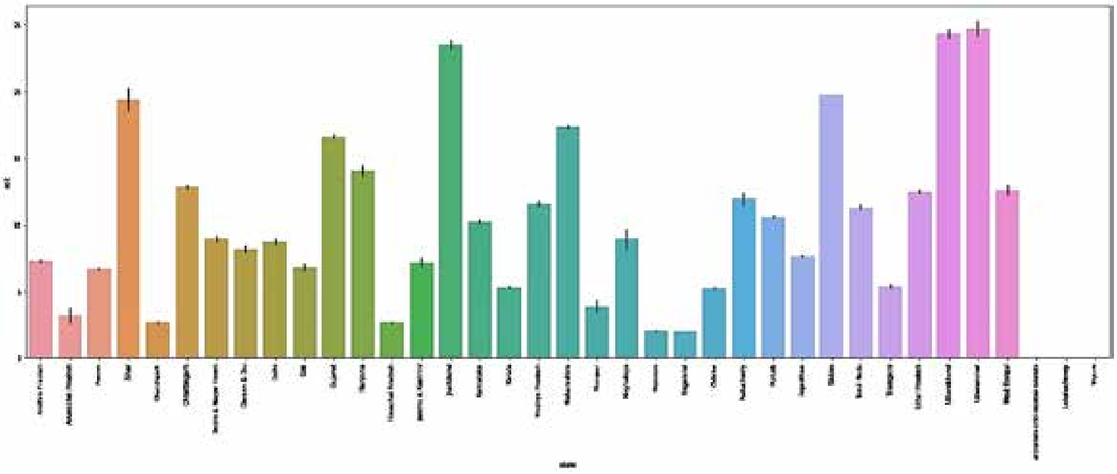

Figure 8. Result of SO_2 level in the air in selected states

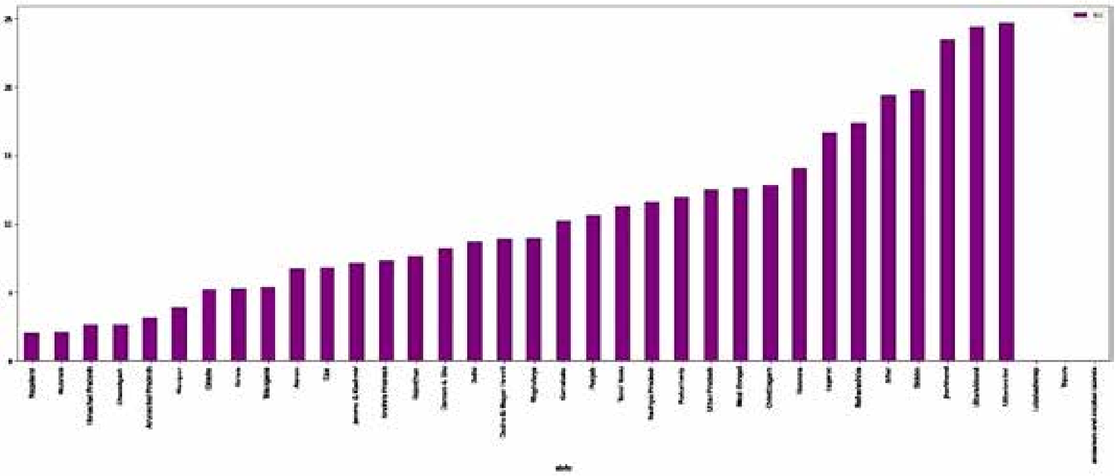

Figure 9. Increasing order of SO_2 level in the selected state.

Experimentation of the dataset

For testing purposes, the air quality frequency levels in 3 different areas such as residential & rural, sensitive and industrial were analyzed using LSTM-Convolutional Neural Network techniques. The areas with the highest rates, exceeding 175,000 MHz, were residential, rural, and other areas. The frequency in the sensitive areas fluctuated depending on the proximity to nearby residential and commercial districts. Figure 7 displays the air quality frequencies for 3 different areas based on the dataset.

The regression performance metric classification was assessed during the experimental process. The model's accuracy for both testing and training data was calculated using logistic regression, decision tree classifiers, random forest classifiers, and k-nearest neighbors. The Kappa score was then calculated to determine which outcomes were better. These techniques are used to predict a variable's value based on the value of another variable. By constructing a decision tree, these algorithms generate the classification model. Each node in the tree represents a test for a specific attribute, and each branch descending from that node represents a potential value for that attribute. Training data with a value of 1.0 was obtained for both the random forest and decision tree classifiers. With a score of 0.9998, the decision tree and random forest classifiers demonstrated excellent test data accuracy. In contrast, the logistic regression model showed relatively low-test data accuracy. Both the decision tree and random forest classifiers achieved a Kappa value of 0.9997, while logistic regression had a low score of 0.5843. Table 5 provides a detailed explanation of the results and the correctness of the training and testing data.

When mixed with water, this colorless, soluble gas emits a strong odor and forms sulfuric acid. The names of the states with the highest and lowest air SO_2 levels are displayed in Figure 8. Uttaranchal and Uttarakhand had the highest recorded atmospheric SO_2 concentrations, while Mizoram and Nagaland had the lowest. Figure 9 shows the states in increasing order based on their SO_2 levels. The NO₂ concentration patterns of the chosen states used in this study align with previous research examining the ambient air quality in those states.

This study found evidence of a connection between NO2 concentrations and a specific Indian state. In West Bengal, it was discovered that the morning rush hour led to high exposure to ambient pollutants, and that busy intersections and major thoroughfares were associated with increased NO2 concentrations due to high traffic volume. Compared to other states, Figures 10 and 11 show that West Bengal has higher levels of NO_2. Ten model configurations were tested in order to determine which produced better results for air purification. Nine distinct analyses were conducted.

This experiment used the following configurations: CNN-LSTM, CNN-GRU, CNNILSTM, SVR, RFR, MLP,

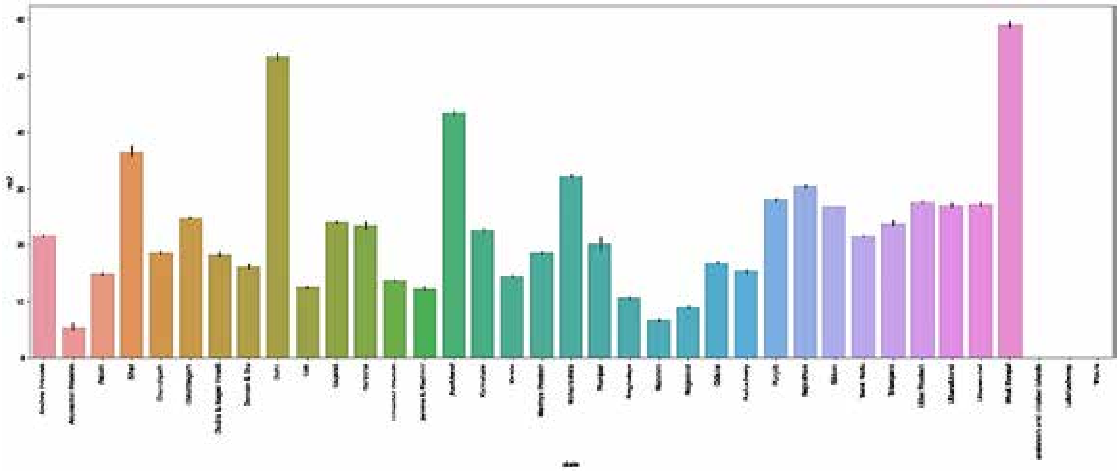

Figure 10. level of no2 in the air which is higher in west bengal

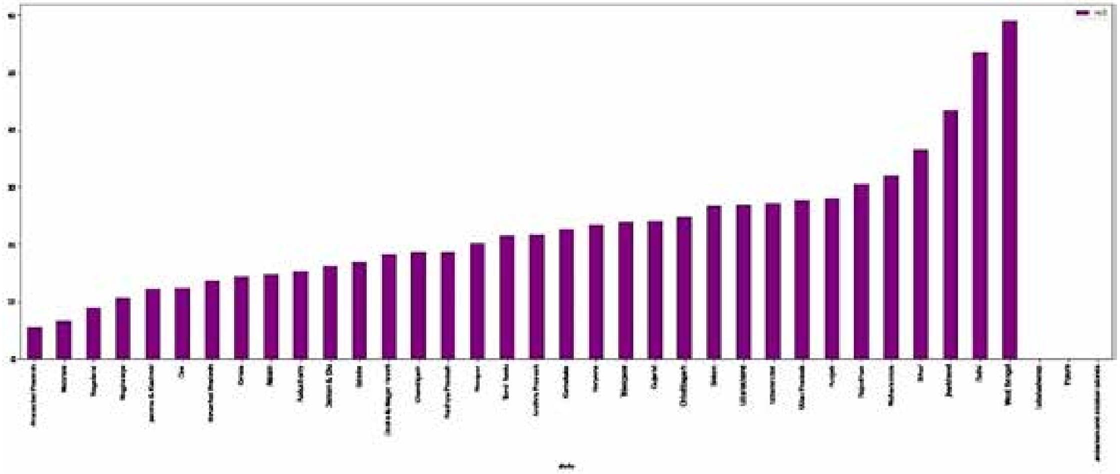

Figure 11. Results shows the increasing order based on their no2 levels.

LSTM, GRU, ILSTM, and CNN-LSTM. With a 8:1:1 ratio, Convolutional Neural Networks with BLSTM achieved a higher value of 0.9620. During a 7:2:1 configuration, CNN-ILSTM reached a maximum value of 0.9638. In SVR experiments, the lowest value was attained. CNN-ILSTM recorded a higher value of 0.9186 during a 6:2:2 configuration, while CNN-BLSTM recorded higher values of 0.9160 and 0.9361 during 6:1:3 and 4:3:3 configurations, respectively. In the evaluation, CNN-BLSTM outperformed the nine alternative model configurations, as clearly described in Table 6.

Conclusion

In conclusion, the application of Stacked Bi-LSTM and CNN for air quality prediction in India has proven to be highly effective. The following key findings were observed

Table 6. Model Performance Comparison with Different Configurations

Model Configuration	SVR	RFR	MLP	LSTM	GRU	ILSTM	CNN-LSTM	CNN-GRU	CNN-ILSTM	CNN-BLSTM
8:1:1	0.8852	0.9001	0.8932	0.9355	0.9492	0.9420	0.9410	0.9498	0.9510	0.9620
7:2:1	0.8762	0.8969	0.9061	0.9365	0.9507	0.9470	0.9487	0.9512	0.9638	0.9537
6:3:1	0.8633	0.8425	0.8821	0.9210	0.9315	0.9399	0.9280	0.9392	0.9330	0.9370
7:1:2	0.8232	0.8356	0.8692	0.8716	0.8796	0.8890	0.9001	0.9030	0.9068	0.9401
6:2:2	0.8269	0.8125	0.8793	0.8921	0.8880	0.8965	0.9034	0.9164	0.9186	0.9134
5:3:2	0.8132	0.8019	0.8611	0.8862	0.8760	0.8890	0.8962	0.8836	0.9020	0.9062
6:1:3	0.7795	0.7851	0.7932	0.8856	0.8569	0.8611	0.8960	0.8695	0.8741	0.9160
5:2:3	0.7487	0.7421	0.7752	0.8236	0.8525	0.8499	0.8321	0.8530	0.8499	0.9221
4:3:3	0.7516	0.7359	0.8003	0.8526	0.8210	0.8511	0.8561	0.8312	0.8501	0.9361

1. The model effectively captured temporal nuances, including diurnal patterns, seasonal trends, and long-term variations in air quality, showcasing its adaptability to the diverse environmental conditions across the country.
2. Spatially, the model exhibited its ability to predict air quality variations across different regions, accurately identifying regional differences and localized patterns. Quantitative evaluation using metrics like RMSE and R-squared (R^2) confirmed the model's reliability and highlighted its potential for precise air quality predictions.
3. While localized discrepancies were identified, these present opportunities for further refinement. Additional data collection and model fine-tuning can help address these discrepancies and enhance the model's accuracy.
4. The states of Uttrachand and Uttrachal recorded high SO_2 levels, while West Bengal exhibited elevated NO_2 concentrations in the air.

Overall, the successful implementation of advanced deep learning techniques has significant implications for environmental management in India. The model provides decision-makers with crucial insights, enabling the development of evidence-based strategies to combat air pollution and protect public health. This study contributes to the growing field of air quality prediction methodologies, laying the groundwork for more informed and proactive environmental stewardship in India's dynamic and diverse context.

References

Akinosho, T. D., Oyedele, L. O., Bilal, M., Barrera-Animas, A. Y., Gbadamosi, A. Q. & Olawale, O. A. (2022). A scalable deep learning system for monitoring and forecasting pollutant concentration levels on UK highways. *Ecological Informatics*, 69, 101609. DOI:10.1016/j.ecoinf.2022.101609

Al-Eidi, S., Amsaad, F., Darwish, O., Tashtoush, Y., Alqahtani, A. & Niveshitha, N. (2023). Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques. *IEEE Access*. DOI:10.1109/ACCESS.2023.3280129

Cao, Y., Zhang, D., Ding, S., Zhong, W. & Yan, C. (2023). A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition. *Tsinghua Science and Technology*, 29(1), 99-111. DOI:10.26599/TST.2023.2200016

Dobrzyniewski, D., Szulczyński, B., Rybarczyk, P. & Gębicki, J. (2023). Process control of air stream deodorization from vapors of VOCs using a gas sensor matrix conducted in the biotrickling filter (BTF). *Archives of Environmental Protection*, *49*(2). DOI:10.24425/aep.2023.144733

Drewil, G. I. & AlBahadili, R. J. (2022). Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors*, 24, 100546. DOI:10.1016/j.measen.2022.100546

Fang, Z., Yang, H., Li, C., Cheng, L., Zhao, M. & Xie, C. (2021). Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR. *Archives of Environmental Protection*, **47**(3). DOI:10.24425/aep.2021.138474

Fu, L., Li, J. & Chen, Y. (2023). An innovative decision-making method for air quality monitoring based on big data-assisted artificial intelligence technique. *Journal of Innovation & Knowledge*, 8(2), 100294. DOI:10.1016/j.jik.2023.100294

Godłowska, J., Kaszowski, K. & Kaszowski, W. (2022). Application of the FAPPS system based on the CALPUFF model in short-term air pollution forecasting in Krakow and Lesser PolandApplication of the FAPPS system based on the CALPUFF model in short-term air pollution forecasting in Krakow and Lesser Poland. *Archives of Environmental Protection*, *48*(3). DOI:10.24425/aep.2022.142698

Holnicki, P., Kałuszko, A. & Nahorski, Z. (2021). Analysis of emission abatement scenario to improve urban air quality. *Archives of Environmental Protection*, **47**(2). DOI:10.24425/aep.2021.137281

Iskandaryan, D., Ramos, F. & Trilles, S. (2023). A set of deep learning algorithms for air quality prediction applications. *Software Impacts*, 17, 100562. DOI:10.1016/j.simpa.2023.100562

Iskandaryan, D., Ramos, F. & Trilles, S. (2023). Graph Neural Network for Air Quality Prediction: A Case Study in Madrid. *IEEE Access*, 11, 2729-2742. DOI:10.1109/ACCESS.2023.3244295

Janarthanan, R., Partheeban, P., Somasundaram, K. & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, 67, 102720. DOI:10.1016/j.scs.2021.102720

Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J. & Wemmert, C. (2022). Deep learning methods evaluation to predict air quality based on Computational Fluid Dynamics. *Expert Systems with Applications*, 203, 117294. DOI:10.1016/j.eswa.2022.117294

Kanmani, P., Selvaraj, P. & Burugari, V. K. (2022). An energy efficient approach of deep learning based soft sensor for air quality management. *Measurement: Sensors*, 24, 100460. DOI:10.1016/j.measen.2022.100460

Liu, B., Yan, S., Li, J., Qu, G., Li, Y., Lang, J. & Gu, R. (2019). A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. *IEEE Access*, 7, 43331-43345. DOI:10.1109/ACCESS.2019.2903323

Liu, C., Pan, G., Song, D. & Wei, H. (2023). Air Quality Index Forecasting Via Genetic Algorithm-Based Improved Extreme Learning Machine. *IEEE Access*. DOI:10.1109/ACCESS.2023.3273346

Lu, T., Gu, C., Yuan, D., Zhang, K. & Shao, C. (2023). Deep learning model for displacement monitoring of super high arch dams based on measured temperature data. *Measurement*, 222, 113579. DOI:10.1016/j.measurement.2023.113579

Matthaios, V. N., Knibbs, L. D., Kramer, L. J., Crilley, L. R. & Bloss, W. J. (2023). Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data. *Atmospheric Environment*, 120233. DOI:10.1016/j.atmosenv.2023.120233

Prado-Rujas, I. I., García-Dopico, A., Serrano, E., Córdoba, M. L. & Pérez, M. S. (2024). A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on Deep Learning. *Engineering Applications of Artificial Intelligence*, 127, 107271. DOI:10.1016/j.engappai.2023.107271

Shao, Q., Chen, J. & Jiang, T. (2023). A novel coupled optimization prediction model for air quality. *IEEE Access*. DOI:10.1109/ACCESS.2023.3267475

Shin, S., Baek, K. & So, H. (2023). Rapid monitoring of indoor air quality for efficient HVAC systems using fully convolutional network deep learning model. *Building and Environment*, 234, 110191. DOI:10.1016/j.buildenv.2023.110191

Wang, X., Wang, M., Liu, X., Mao, Y., Chen, Y. & Dai, S. (2024). Surveillance-image-based outdoor air quality monitoring.

Environmental Science and Ecotechnology, 18, 100319. DOI:10.1016/j.ese.2024.100319

Wardana, I. N. K., Fahmy, S. A. & Gardner, J. W. (2023). TinyML Models for a Low-cost Air Quality Monitoring Device. *IEEE Sensors Letters*. DOI:10.1109/LSENS.2023.3247646

Wood, D. A. (2022). Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining. *Sustainability Analytics and Modeling*, 2, 100002. DOI:10.1016/j.susanm.2022.100002

Yadav, N., Sorek-Hamer, M., Von Pohle, M., Asanjan, A. A., Sahasrabhojanee, A., Suel, E., Arku, R., Lingenfelter, V., Brauer, M., Ezzati, M. & Oza, N. (2023). Using Deep Transfer Learning and Satellite Imagery to Estimate Urban Air Quality in Data-Poor Regions. *Environmental Pollution*, 122914. DOI:10.1016/j.envpol.2023.122914

Yang, Y., Mei, G. & Izzo, S. (2022). Revealing influence of meteorological conditions on air quality prediction using explainable deep learning. *IEEE Access*, 10, 50755-50773. DOI:10.1109/ACCESS.2022.3163935

Yu, W., Nakisa, B., Ali, E., Loke, S. W., Stevanovic, S. & Guo, Y. (2023). Sensor-based indoor air temperature prediction using deep ensemble machine learning: An Australian urban environment case study. *Urban Climate*, 51, 101599. DOI:10.1016/j.uclim.2023.101599

Zhang, B., Wang, Z., Lu, Y., Li, M. Z., Yang, R., Pan, J., & Kou, Z. (2023). Air pollutant diffusion trend prediction based on deep learning for targeted season—North China as an example. *Expert Systems with Applications*, 232, 120718. DOI:10.1016/j.eswa.2023.120718

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q. & Huang, L. (2019). A predictive data feature exploration-based air quality prediction approach. *IEEE Access*, 7, 30732-30743. DOI:10.1109/ACCESS.2019.2903346

Zwierzchowski, R. & Różycka-Wrońska, E. (2021). Operational determinants of gaseous air pollutants emissions from coal-fired district heating sources. *Archives of Environmental Protection*, **47**(3). DOI:10.24425/aep.2021.138473