Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper to be published in Vol. 25, No. 3, September 2026 issue

Original Research Paper

Deep Learning for AQI Prediction Using Multiple Feature Vectors: A Case Study of Colaba and Deonar Stations

Darakhshan Khan[1]†, Archana B. Patankar[1], Himani Deshpande[2] and Juhi Janjua[1]

[1]Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai, 400050, India

[2]Department of Artificial Intelligence and Data Science, Thadomal Shahani Engineering College, Mumbai, 400050, India

† Corresponding author: Darakhshan Khan; darakshan.khan@thadomal.org

https://orcid.org/0000-0002-9450-874X, https://orcid.org/0009-0001-2305-3776, https://orcid.org/0000-0002-9050-0732,

Key Words	Air pollutant, Meteorological parameters, Bidirectional LSTM, Deep learning for AQI prediction
DOI	https://doi.org/10.46488/NEPT.2026.v25i03.B4381 (DOI will be active only after the final publication of the paper)
Citation for the Paper	Khan, D., Patankar, A.B., Deshpande, H. and Janjua, J., 2026. Deep learning for AQI prediction using multiple feature vectors: A case study of Colaba and Deonar stations. *Nature Environment and Pollution Technology*, 25(3), B4381. https://doi.org/10.46488/NEPT.2026.v25i03.B4381

Abstract: Air quality monitoring and prediction are important for effective public health strategies as air pollution is one of the major contributors towards mortality. The goal is to design and evaluate different deep learning models - Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and a hybrid Conv1D-LSTM on different input sets of features. The dataset consists of pollutant and meteorological parameters spans from 2019 to 2024 with hourly frequency for two monitoring stations: Colaba and Deonar. The models were trained on three sets of features: pollutant-only, meteorological-only, and combined. Model accuracy was determined by root mean square error, coefficient of determination, mean absolute percentage error, and explained variance score. Results indicate that combined attributes significantly enhance prediction quality, with hybrid CNN-LSTM best performing at Colaba and LSTM on meteorological attributes performing best at Deonar. Bi-LSTM had consistent performance on feature sets. These results underscore the importance of using both pollutant as well as meteorological information and illustrate the efficiency of sophisticated deep learning structures for location-based air quality prediction.

1. INTRODUCTION

Globally, air pollution has emerged as a major cause of mortality and significant threat to human health. Monitoring, understanding, and predicting environmental factors, specifically, those related to air quality, remain critical challenges for governments and corporations worldwide. Given the increasing rate of pollution, there is a critical need to develop methods that could predict the Air Quality Index with accuracy and support in creating better public health strategies (Schürholz D et al. 2020). Human life is basically surviving on air, but its quality gets ruined due to vast pollution, that leads to acute multiple health hazards, such as, physiological and respiratory disorders. Scientific studies have proved that air pollution is the greatest environmental risk, driven by rapid industrialization and urban growth. This unplanned emission has degraded the air quality grossly, exposed people to hazardous substances and posing substantial risks to public health.

The need to mitigate these impacts underlines the importance of robust monitoring systems and predictive models in safeguarding air quality and reduce the adverse health effects of pollution on communities worldwide (Gupta S et al. 2023). Particulate matter (PM) has become a major pollutant of concern, hugely affecting air quality across borders. The particles in PM vary in size and chemical composition; therefore, the risks associated with them are also variable in space and time. Studies have identified a relationship between PM exposure and daily respiratory mortality in both construction workers and residents living nearby construction sites (Yan H et al. 2023) and it can exacerbate chronic obstructive pulmonary disease (COPD) symptoms, increasing the risk of mortality (Gou A et al. 2023) (DeVries R et al. 2017). Other common air pollutants are Ammonia, Sulphur dioxide, Carbon monoxide, Nitrogen oxides and Ozone, all of which have adverse effects on human health. Ammonia is an irritating gas that affects the skin, respiratory system, digestive system, and ophthalmic system. Ammonia combines with other air constituents to form $PM_{2.5}$, a key air pollutant with enormous health effects (Petrus M et al. 2022). Both SO_2 and ozone have been shown to significantly increase the risk of hospital admissions, cardiopulmonary diseases, myocardial infarction, chronic obstructive pulmonary diseases, respiratory diseases (Wu H et al. 2022) (Jian Z et al. 2024). Higher levels of these pollutants are directly correlated with the proportion of population having chronic respiratory diseases. Increase levels of CO in air can lead to reduction of oxygen level in human bodies which can damage cell, tissues and organs whereas high NOx levels can cause respiratory morbidity and lung diseases (Zhang C & Zhang L 2024).

Climate change is evident across globe and it is observed that variations in climatic conditions have strong influence on Air Quality Index (AQI) levels, which if not controlled can amplify risks to human health (Liu Y et al. 2022) (Yan M et al. 2024). Multiple studies have shown a strong correlation between AQI levels and various climatic factors such as, temperature, wind direction, wind speed, humidity, pressure, visibility, rainfall, sun exposure, diurnal temperature, precipitation, weather outlook conditions etc. Multiple studies also claims that the principal climatic factors that is affecting the AQI levels are temperature, pressure, humidity, dew point, and wind speed. Due to the local geomorphological and meteorological factors, AQI levels will be affected in an area within 100 kms of radius (Liu Y et al. 2022).

With the advancement of deep learning, various sequential deep architectures, such as, Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU) and Long Short-Term Memory cell (LSTM) have gained huge popularity in making accurate and robust AQI predictions for many cities across world. An RNN is created out of the feed-forward neural network, wherein some nodes have not only inter-node connections but also backward loops that allow the network to remember the present and the recent past (Ayus I et al. 2023). A variant of RNN, the GRU, utilizes two main gates: an update gate and a reset gate, to manage memory. The update gate controls addition of new information and the reset gate controls how much of the previous state to remember. It can be further extended as Bidirectional GRU, which learns sequences in both directions by combining two unidirectional GRUs for learning in forward and backward direction (Ayus I et al. 2023) (Mandal AK & Sen R

2024). Another variant of RNN is LSTM which provides solution to the problems of long-term dependency with forget, input, and output gates that control memorization and removal of an information (Ayus I et al. 2023) (Song Q et al. 2024). The Bidirectional LSTM further improves on this by processing input sequences in both forward and backward directions. CNN-GRU is a hybrid model where a Convolutional Neural Network is applied for feature extraction and a GRU for sequence prediction, hence creating a deep architecture capable of handling spatio-temporal data (Ayus I et al. 2023) (Shi T et al. 2023).

This research involves performing a statistical correlation analysis between AQI and various factors, such as meteorological parameters and atmospheric pollutants. It also aims to understand the effect of diurnal and seasonal variations on air quality data. The core goal is to understand how various feature vector simultaneously interact with different deep learning architectures. In addition, the study seeks to determine which deep leaning architectures provide the most robust prediction and have better generalization ability. Two stations with different environments are studied here: Colaba, which is located on the coast, and Deonar, an inland station influenced mainly by landfill and industrial pollution sources.

The rest of the paper is organized as follows, next section contains a review of the literature, which highlights global perception on AQI prediction, advancement in machine learning and deep learning techniques, and the role of climatic condition for AQI forecasting. The section ahead encompasses the understanding the relation of AQI with the factors influencing it, succeeded by temporal and spatial analysis. The proceeding section deals with the methodology adopted, that includes, detailed information on the data collection process, different model architectures used, feature engineering, and model validation. Final section shows the setup of the experiment, results, and observations of the research, followed by conclusion.

2. LITERATURE REVIEW

Air pollution has emerged as a global health crisis, and therefore, stable and steady forecasting models for air quality monitoring and prediction are necessary. Different researchers have utilized different methodologies to solve this problem, especially through the use of enhanced machine learning and deep learning techniques. This section discusses some of the most significant studies around the world to provide a complete context for the current research.

2.1. Cross-Continental Perspective on AQI prediction

China: Exhaustive research in China shows that the variations in AQI levels are largely influenced by atmospheric pollutants such as PM2.5, PM10, Ozone (O3), SO2, NOx, and CO and the concentration of these atmospheric pollutants are heavily influenced by climatic conditions. The studies (Zhang Y & Jiang W 2018) (Sun R et al. 2019) uses positive matrix factorization with chemical mass balance, regression and correlation analysis on data collected from different monitoring stations across multiple provinces in China, to gauge association between concentration of PM2.5 and various metrological factors, such as temperature, wind speed, relative humidity, and air pressure. It was observed that low speed of wind, coupled with high rising temperature and high humidity will lead to higher concentration of PM2.5 in air. Higher levels of PM2.5 were observed in winters and during peak hours in a day. This research (Tian J et al. 2021) have used wavelet transform and demonstrated that the ozone concentration shows strong temporal and spatial association with meteorological variables, like, wind speed, solar radiation, temperature, and humidity with noteworthy observation that the urban areas are prone to higher ozone levels .Whereas another study, in Linfen city, which is one of the most polluted areas in China, seasonal variation analysis was carried out to capture the climatic fluctuations on air pollution. Finding claims that during winter season, air quality was very poor which was heavily influenced by the speed of the wind and the precipitation (Cui H et al. 2018). Authors (Fang C et al. 2015) have performed and extensive research on data set collected from air quality monitoring station for 338 cities across China, using multiple tools such as ordinary least square, special lag models, geographically weighted regression and Spearman's rank correlation for understanding the effect of population size, industrialization and transportation network on air pollution.

India: Multiple researches conducted across different cities in India indicates that AQI levels are not only influenced by meteorological conditions but also location specific factors such as distances from dumping yard or how near or far monitoring stations are from the sea coast. The review article (Karthick K et al. 2024) integrates multiple pieces of research conducted at different cities namely, Ahmedabad, Delhi, Lucknow, Gurugram, Mumbai etc. It highlights an observation that the concentrations of PM2.5 and CO are the two most crucial factors influencing the AQI, thus, the reduction in the two aforementioned pollutants can contribute significantly to better air quality in Indian cities. Another study (Varaprasad V et al. 2024) have used five coastal Indian cities: Kolkata, Visakhapatnam, Chennai, Mumbai, and Thiruvananthapuram, as the research objects. The results seem to indicate that the most significant concentrations of PM2.5 occur during winters, December to February, whereas the lowest concentrations are shown during the monsoon, June to September. Among all these findings, an important outcome is that the eastern Indian coastline bears an intense influence from outflowing emissions from the Indo-Gangetic plain that gradually decreases from north to south-that is, Kolkata to Chennai. The analysis of the sea-land breeze cycles indicates a negative breeze magnitude versus PM2.5, which could suggest that weaker breezes might lead to worse air quality because of weaker ventilation. The researches (Rangaswamy PS et al. 2022) (Poyyamoli G & Boss CL 2014) have used data from Jawahar Nagar dump yard, Hyderabad and Kammiyampet dump yard, Cuddalore, to understand how improper garbage disposal practices around the landfills can increase the concentration of air pollutants in nearby areas. It is observed that due to the decomposition of waste or burning of garbage at a dumping site, there was much higher concentration of multiple air pollutants than the permissible limits, which can lead to various disorders related to respiratory system or cardiovascular problems. There was also formation of a chemical substance called as leachate, which pollutes land as well as ground water.

America and Europe: Studies from American and European continents inspect the interplay between air quality and various human activities such as urbanization, transportation and agriculture. Sterling M (2024) investigated how air quality varies with population density in any country, as well as the number of transportation-related emissions connected to it, in California. Notable findings are higher density is associated with increased emissions of vehicles; thus, the more a region contains human populations, the higher the emissions become, meaning more air pollutants from transports. In country like Romania, where 14% of civil workforce occupation is agriculture, a study (Petrus M et al. 2022) was conducted where proportion of NH3 was detected using laser photoacoustic spectroscopy. The key sources for higher NH3 levels are agricultural practices and emissions caused by fuel combustion. Other pertinent variables affecting the over-time dispersion and accumulation of NH3 include temperature, humidity, and wind speed. This study (Sajjad Abdollahpour et al. 2024) illustrates considerable links between urban spatial structure and air quality using a multi-decade analysis over the period of 481 U.S. cities spanning 1990–2015. Compact urban form, population density, circularity, and green spaces diminish pollutant concentrations, but expansion and industrial areas tend to enhance them. In total, a 10% improvement in key factors would potentially prevent over 10,000 deaths annually.

Asia and Africa: In Adıyaman (Kara Y et al. 2024), authors claim that natural and human activities drive air pollution. Major contributors include dust transport from southerly winds, crop burning, and heating, which are mainly found in the city centre. PM10 is positively correlated with SO2, while wind speed and temperature have negative effects on pollution levels. The study (Saadi D et al. 2021) used a Polar 810i monitor to test CO, Heart Rate Variability (HRV), and city size in Tel Aviv (metropolitan) and Afula (small city). Results have shown that pollution is high in Tel Aviv due to industrialization and vehicle emission, with CO levels showing a positive correlation and HRV showing a negative association, indicating greater cardiac effects. Weak positive correlations of PM2.5 and temperature ($r = 0.42$) and humidity ($r = 0.37$) have been reported by a study in Ratchaburi, Thailand. AirQ+ [33] states that PM2.5 during dry periods violated the limit set by WHO, the old age with more years lost due to disability. In Addis Ababa, deteriorating vehicle conditions and poor road status are drivers of air pollutants, especially in the heavily congested areas of Megenagna. SO2, NO2, PM2.5, and PM10 data from 43 sites show significant spatial heterogeneity, with SO2 often above safety thresholds. This study (Bizualem B et al. 2023) calls for further research on these traffic-related emissions to reduce this pollution. Global studies show that meteorological factors, urbanization, transportation, and waste management have a great impact on air quality, with most pollutants such as PM2.5, PM10, CO, and SO2 exceeding the safety limit through human activities and climatic conditions.

2.2. Innovation in AQI Prediction Techniques

Neural Network based Models: These models have garnered significant attention in predicting air quality indices (AQI) and pollutant concentrations. For example, one study utilizing data from four Chinese cities implemented machine learning algorithms such as extreme gradient boosting (XGBoost), light gradient boosting, and random forest. Among these, XGBoost demonstrated the highest forecasting accuracy, achieving a coefficient of determination score of 0.929 (Wang S et al. 2023). Another comparative analysis (Zhang C et al. 2020) explored various predictive models, including linear regression, CNN, LSTM, GRU, Bidirectional LSTM, and backpropagation neural networks. Using hourly AQI data from 1,615 locations across China, this study identified Bidirectional LSTM as the most effective model.

A broader review (M´endez M et al. 2023) examined global approaches to AQI prediction, with a specific focus on PM2.5. It highlighted the frequent use of weather data and pollutant concentrations as input features and identified models like LSTM, MLP, CNN, and GRU as the most commonly applied techniques. Evaluation metrics such as RMSE, R2 Score, and MAPE were consistently used to assess performance. Additionally, research involving data from Belfast city centre compared the predictive capabilities of LSTM, GRU, and ARIMA models for pollutants including NO2, O3, SO2, PM2.5, and PM10. The study found LSTM to outperform other models consistently, while ARIMA delivered subpar predictions (Naz F et al. 2023). Collectively, these findings underscore the increasing reliance on neural network architectures for precise air quality modelling.

2.3. Hybrid Models for Optimized Accuracy

These approaches to AQI prediction have demonstrated notable improvements by integrating simple and advanced methodologies. For instance, one study (Kleingchuay W et al. 2023) employed empirical mode decomposition to extract time-frequency information, which was combined with Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models for analyzing residuals and historical data. The output was subsequently processed through LSTM and SVM models, yielding enhanced MAE scores for LSTM and superior index of agreement values for SVM. Another study (Barve A et al. 2020), using meteorological data and AQI values from Beijing, introduced a hybrid framework with parallel dense neural networks and LSTM layers. This approach achieved significantly better MAE values compared to standalone LSTM models. Similarly, research (Thakur N et al. 2023) combined principal component analysis (PCA) with deep learning models like LSTM, GRU, and Bidirectional LSTM for multivariate AQI forecasting, with PCA and GRU delivering the best performance across various metrics.

For more advanced architectures, a study (Shi T et al. 2023) proposed a hybrid TCNbiGRU model that utilized Temporal Convolutional Networks (TCN) to capture long-term patterns and biGRU for short-term dependencies, outperforming standalone LSTM and GRU approaches. In another study (Lu Y & Li K 2023), researchers used a hybrid CNNBiLSTM model optimized through Bayesian methods to predict air pollutants in Tianjin, while this article (Nikpour P et al. 2024) introduces the ”Gelato” framework. This innovative approach combined particle swarm optimization, a transformer-inspired architecture, and XGBoost for multivariate AQI prediction. In research focusing on Indian metropolitan areas (Binbusayyis A et al. 2024), a stacked attention-based GRU model incorporating KL divergence was employed after imputing missing data with deep generative adversarial networks. This framework achieved high prediction accuracy. Lastly, this paper (Shankar L & Arasu K, 2023) explores hidden temporal and periodic patterns in AQI data, demonstrating that hybrid models like CNN-LSTM and GRU integrated with Empirical Mode Decomposition surpassed their standalone counterparts in predictive performance. Collectively, these studies underscore the effectiveness of hybrid architectures in enhancing the accuracy and robustness of AQI prediction, leveraging a blend of pre-processing techniques and sophisticated deep learning models.

A Transformer model was proposed for $PM_{2.5}$ prediction at 12 Beijing sites and performed better than the CNN-LSTM-Attention model, with EVS, MAE, MSE, and R^2 being improved by 12%, 9%, 6%, and 30%, respectively (B. Cui et al. 2023). It successfully captured short-term meteorological changes and long-term seasonal trends and has excellent ability in modeling long-range dependencies for air quality forecasting. This research (Wu. Q et al. 2024) presents a PLS-VAER model for indoor air quality prediction, where PLS extracts latent variables to enhance VAER input and improve accuracy. The model outperforms conventional approaches and provides a robust, eco-friendly solution for indoor air quality monitoring and parameter optimization. In contrast to current Bi-LSTM and hybrid models that utilize uniform sets of features across regions, this research focuses on station-specific feature engineering, adapting inputs for each site's environment and geography. This enables the capture of localized drivers of pollution, including coastal influence, and landfill distance, which improves predictive performance and interpretability over traditional approaches.

3. TEMPORAL AND CORRELATION ANALYSIS

This section discusses the temporal correlation of AQI values, followed by seasonal and diurnal oscillations between AQI and factors driving AQI levels, such as air contaminants and meteorological conditions. This research draws upon data from two sources: 1. CPCB site 2. OpenWeatherMap API. Atmospheric air pollutant data for the two stations namely, Colaba and Deonar, is taken from the CPCB portal (Central Pollution Control Board 2024) for the periods of July 2019 to March 2024 for Colaba station and November 2020 to March 2024 for the Deonar station. The meteorological data for the same stations and at same timestamp was captured via the OpenWeatherMap API (OpenWeather 2024).

3.1. Temporal Analysis

There is a strong autocorrelation between AQI values as shown in Fig. 1. Present AQI values are exhibiting association with past AQI values, that is, the present values are dependent of the past AQI values. In Fig. 1, for both stations, all points are centered around diagonal that indicates AQI time series have same descriptive properties such as, mean, variance, standard deviation over different time instances, that shows both time series are stationary in nature (Douglas C et al. 2015).

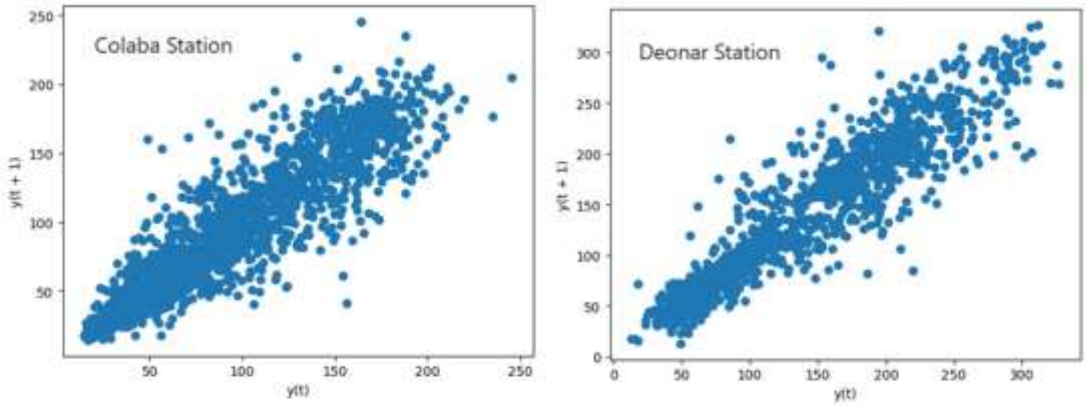


Fig. 1: Auto-lag plots of AQI values for Colaba and Deonar Stations.

3.2. Correlation Analysis

All six atmospheric pollutant have positive association with AQI levels in air, as shown in Fig. 2. Particulate matters, PM2.5 and PM10 have nearly a perfect positive correlation whereas NOx and SO2 also exhibit a strong positive association with AQI levels. For ozone, there is location specific dependency that is, it shows modest positive influence on AQI levels for Deonar station but not much for Colaba station.

Similarly, Pearson Coefficient was used to understand relationship between meteorological parameters and AQI levels. For both the stations from Fig. 3 and Fig. 4, it is apparent that temperature, dew point, minimum and maximum temperature in a day, humidity, wind speed and wind degree have negative association with AQI levels whereas diurnal temperature and pressure have positive impact on AQI levels. It is observed that out of all meteorological parameters, pressure, humidity and dew point have strong association with the target variable.

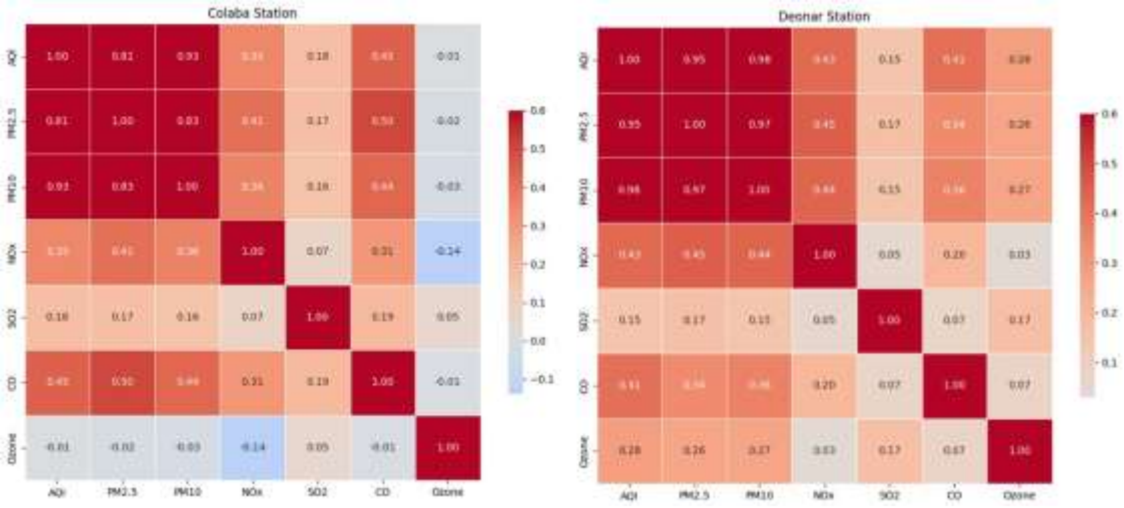


Fig. 2: Heatmap of Pearson correlation between atmospheric pollutants and AQI values for Colaba and Deonar Stations.

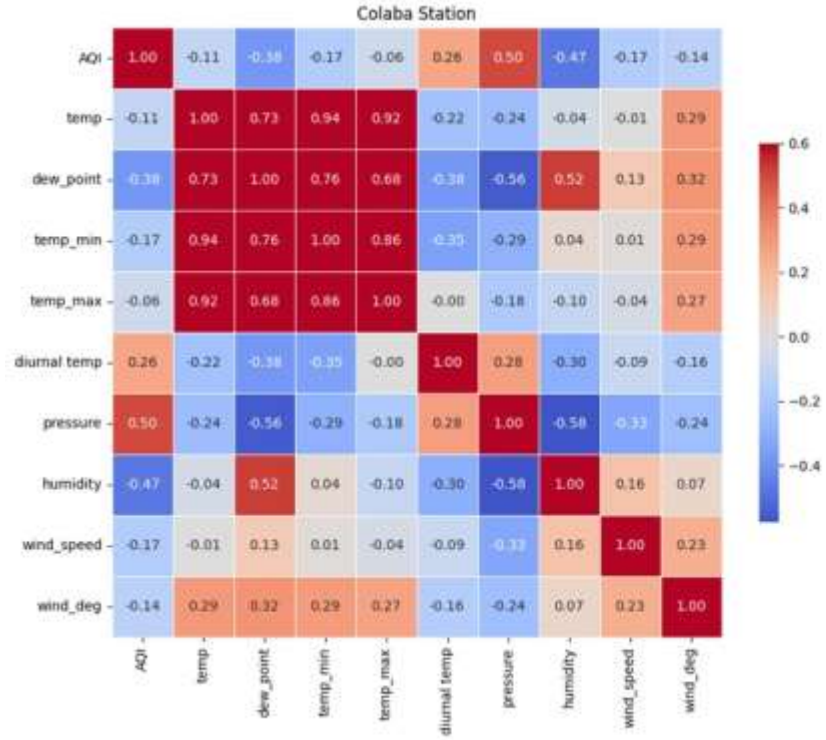
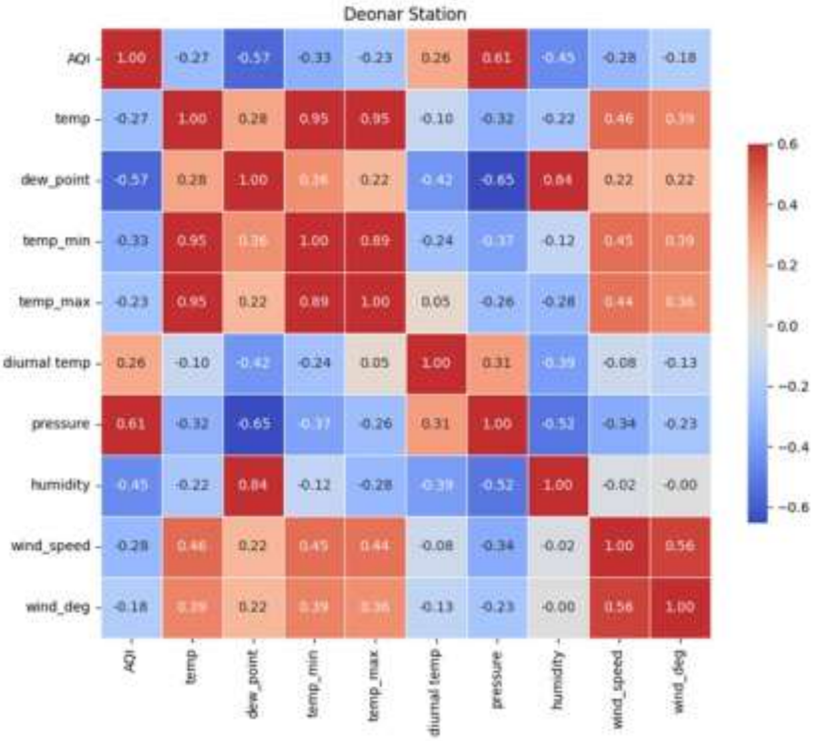


Fig. 3: Heatmap of Pearson correlation between meteorological factors and AQI values for Colaba Station.



	AQI	temp	dew_point	temp_min	temp_max	diurnal temp	pressure	humidity	wind_speed	wind_deg
AQI	1.00	-0.27	-0.57	-0.33	-0.23	0.26	0.61	-0.45	-0.28	-0.18
temp	-0.27	1.00	0.28	0.95	0.95	-0.10	-0.32	-0.22	0.46	0.39
dew_point	-0.57	0.28	1.00	0.36	0.22	-0.42	-0.65	0.84	0.22	0.22
temp_min	-0.33	0.95	0.36	1.00	0.89	-0.24	-0.37	-0.12	0.45	0.39
temp_max	-0.23	0.95	0.22	0.89	1.00	0.05	-0.26	-0.28	0.44	0.36
diurnal temp	0.26	-0.10	-0.42	-0.24	0.05	1.00	0.31	-0.39	-0.08	-0.13
pressure	0.61	-0.32	-0.65	-0.37	-0.26	0.31	1.00	-0.52	-0.34	-0.23
humidity	-0.45	-0.22	0.84	-0.12	-0.28	-0.39	-0.52	1.00	-0.02	-0.00
wind_speed	-0.28	0.46	0.22	0.45	0.44	-0.08	-0.34	-0.02	1.00	0.56
wind_deg	-0.18	0.39	0.22	0.39	0.36	-0.13	-0.23	-0.00	0.56	1.00

Fig. 4: Heatmap of Pearson correlation between meteorological factors and AQI values for Deonar Station.

3.3. Statistical Analysis

Table 1 gives the summary statistics of AQI values at Colaba and Deonar stations, showing dispersion in terms of variance and standard deviations, seasonal variation by calculating winter means and summer means, and also diurnal variation with day and night means for winter and summer seasons. Both time series are volatile in nature, with Deonar showing higher dispersion or spread.

Table 1: Statistical summary of AQI values for Colaba and Deonar Stations.

Stations	Dispersion		Seasonal Variation		Diurnal Variations			
	Variation	Standard Deviation	Winter Mean	Summer Mean	Winter Day Time Mean	Winter Night Time Mean	Summer Day Time Mean	Summer Night Time Mean
Colaba	4532.95	67.32	73.06	114.16	70.45	76.03	144.26	116.32
Deonar	7382.04	85.92	202.32	120.18	199.65	205.81	190.62	120.51

In all respects, Deonar always showed higher AQI values in both seasons and at all times of the day, indicating generally poorer air quality compared to Colaba. In winter, both stations have higher readings at night than during the day. During summer, both stations indicate higher values during the day as compared to night, with Colaba showing more pronounced day–night differences, manifesting time of day could have a notable impact on air quality. To understand variability in AQI levels, both seasonal and diurnal variations must be captured precisely.

4. MATERIALS AND METHODS

Fig. 5. shows the detailed diagram of the AQI prediction workflow using deep learning models. Basically, there are four stages: Data Collection, Data Pre-processing, Feature Engineering, and Model Development. In stage 1, data is collected from multiple sources which include concentration of various atmospheric pollutants, such as particulate with 15 matter and gaseous pollutants, along with meteorological parameters, namely, temperature, wind speed, wind direction, atmospheric pressure, dew point, humidity etc.

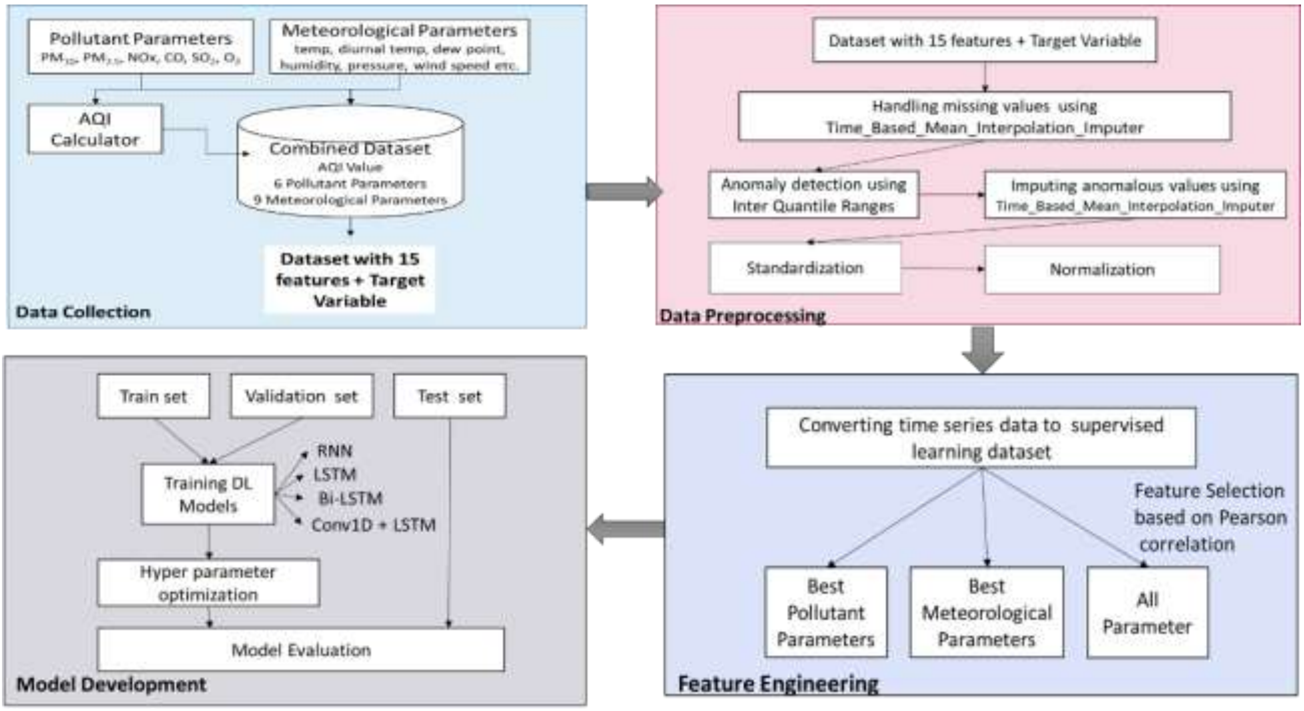


Fig. 5: A comprehensive Deep learning workflow for accurate AQI prediction.

A combined dataset features is created with pollutant levels and meteorological factors. The data is for hourly records, with Colaba having 41,617 data points from 01 July 2019 to 31 March 2024 and Deonar having 29,929 data points from 01 November 2020 to 31 March 2024, with the gap being due to non-availability of data on the CPCB portal for some dates. The AQI calculator (Central Control Room for Air Quality Management Delhi NCR 2024) is used to calculate AQI value, which is also appended in dataset. Data pre-processing is stage 2, which includes multiple steps, such as, filling missing value using time based mean interpolator, followed by detecting irregularities in data using inter quartile ranges and filling those irregularities with time based mean imputer. The TBMI algorithm imputes step by step, first computing the group means by time, month, day, and year, and then imputing missing values with calculated group means. The two-step anomaly treatment used in this study begins with the inter-quantile range (IQR) method at the 15th and 85th percentiles to mark values beyond this range. This method identifies medium-sized anomalies without a high degree of sensitivity towards extreme values. Before feeding for training, data is standardized and normalized, which is equivalent to adjusting values to a common scale, ensuring uniformity and comparability across the dataset.

Next stage is feature engineering where multiple feature sets were created based on results of Pearson's correlation matrix. This experimentation and combinations were created to understand how different deep learning topologies interact with different feature vectors. For this study, three different feature vectors, particularly, Best Pollutant parameters, Best Meteorological parameters and all 15 parameters were used, as shown in Table 2. Last stage of this research includes development of model, which involves splitting data into multiple set, namely, train, validation and test set with 70:20:10 ratio for both stations. Using optimized hyper parameter setting along with training set and validation set, multiple deep learning architectures, namely, RNN, LSTM, Bidirectional LSTM and Conv1D+LSTM as shown in Fig. 6, are trained using MSLE as loss function. For all these architectures, past one week data was used to make single step ahead AQI prediction. Shape of input was different for different architecture depending on feature set and location.

Table 2: Selected Features for AQI Prediction at Colaba and Deonar.

Stations	Colaba	Deonar
Best Pollutant Parameters	PM_{10}, $PM_{2.5}$, NOx, CO	PM_{10}, $PM_{2.5}$, NOx, CO, Ozone
Best Meteorological Parameters	Pressure, Humidity, Dew Point, Diurnal temperature	Pressure, Humidity, Dew Point, Temperature, Wind Speed
All Parameters	6 Pollutant Levels + 9 Meteorological Parameters + AQI (Target Variable)	6 Pollutant Levels + 9 Meteorological Parameters + AQI (Target Variable)

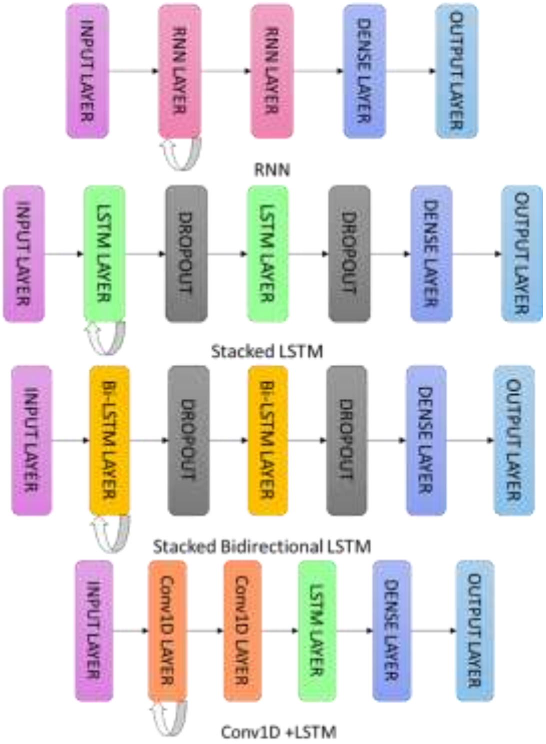


Fig. 6: Architectural Diversity in Deep Learning Models for AQI Predictions.

First architecture, that is, RNN architecture has an input layer followed by couple of RNN layers and eventually a dense layer before final output layer. The RNN layer processes sequential data and it captures the temporal dependencies using hidden state over multiple time steps. The dense layer is used to project the outputs of the RNN into the desired dimension of output. RNN suffers from problem of diminishing gradients (Khan et al. 2023). Second architecture, which is stacked LSTM, is enhancement over simple RNN, where RNN cells are replaced by LSTM cells. LSTM cells can reduce vanishing gradient issue with the help of gating mechanism, where different gates decide which information to pass and which information to remove (Le VD et al. 2020).

To avoid overfitting, dropout regularization is used and stacked configurations can help in learning more complicated patterns by increasing their depth. Third architecture studied is stacked Bidirectional LSTM, where input sequences is operated in both forward and backward direction. These models can perceive latent relationship between input and target variable on both past and future contexts (Mandal et al. 2024). Architecture is similar to LSTM architecture, with one change, LSTM cells in hidden layers are replaced with Bi-LSTM cells. Lastly, there is a hybrid architecture which used convolution layers for extracting local patterns and then fed those local extracted patterns to LSTM layers for capturing temporal relationships (Jiao L et al. 2020). This hybrid model can take advantages of both learning topologies.

4.1. Evaluation Metrics

To understand different aspects of model's performance, multiple performance metrics were used. These metrics gives a comprehensive evaluation of prediction models covering various facet, such as, magnitude of error in prediction, how much variation in the data was captured etc.

RMSE: RMSE stands for Root Mean Squared Error and is a measure of the average magnitude of the errors between the actual and predicted AQI values. It calculates the square root of the average of the squares of the differences between every actual value and its predicted value (Mandal et al. 2024) (Botchkarev 2019). Squaring gives more importance to larger errors than to smaller ones; hence, this metric is sensitive to outliers. Lower RMSE value is an indication that the average of the predicted value is closer to actual values.

R^2 Score: R^2 Score is the coefficient of determination (Plevris et al. 2022) (Botchkarev 2019), which indicates how well the model fits the variability of the target variable. It provides a number showing the proportion of variance in the dependent variable which might be predictable from the independent variables. The R^2 score can take any value from 0, which means that the model does not explain any variability in the target variable, to 1, which would mean that it explains all the variability. A higher R^2 score is indicative of how good the model fits the data and, therefore, how well it is at predicting the AQI values based on the input features.

MAPE: MAPE stands for Mean Absolute Percentage Error, which is an accuracy metric calculated by taking average absolute percentage difference between the predicted and actual values (Plevris et al. 2022) (Botchkarev 2019). This will be useful in a case where the accuracy of a model on different datasets or scales is compared. MAPE normalizes the errors. The lower the value of MAPE, the more accurate the predictions; the higher the value, the more the error.

EVS: Another explanatory metric is EVS stands for Explained Variance Score (Plevris et al. 2022), indicating the proportion of variance in a target variable explained by the predicted values from a model. This would lie between 0 and 1, where higher values mean that more of the variability is accounted by the model. An EVS of 1 would mean perfect explanation of the variance by the model, and one close to 0 would mean very little explanatory power. The EVS tells how well the model does in capturing underlying patterns in data.

5. RESULTS AND DISCUSSION

5.1. Experimental Setup

For this experimentation, every combination of deep learning networks along with different feature vectors were run for 50 epochs with a batch size of 256 samples. For recurrent hidden layers, activation function used was relu and number of hidden units were 64 followed by 32. For dense layer, linear activation is used with 16 and 1 neurons. For efficient training, loss function used was Mean Squared Logarithmic Error (MSLE), Adam optimizer with 0.001 as learning rate was used for stable convergence with dropout regularization of 0.2 is used.

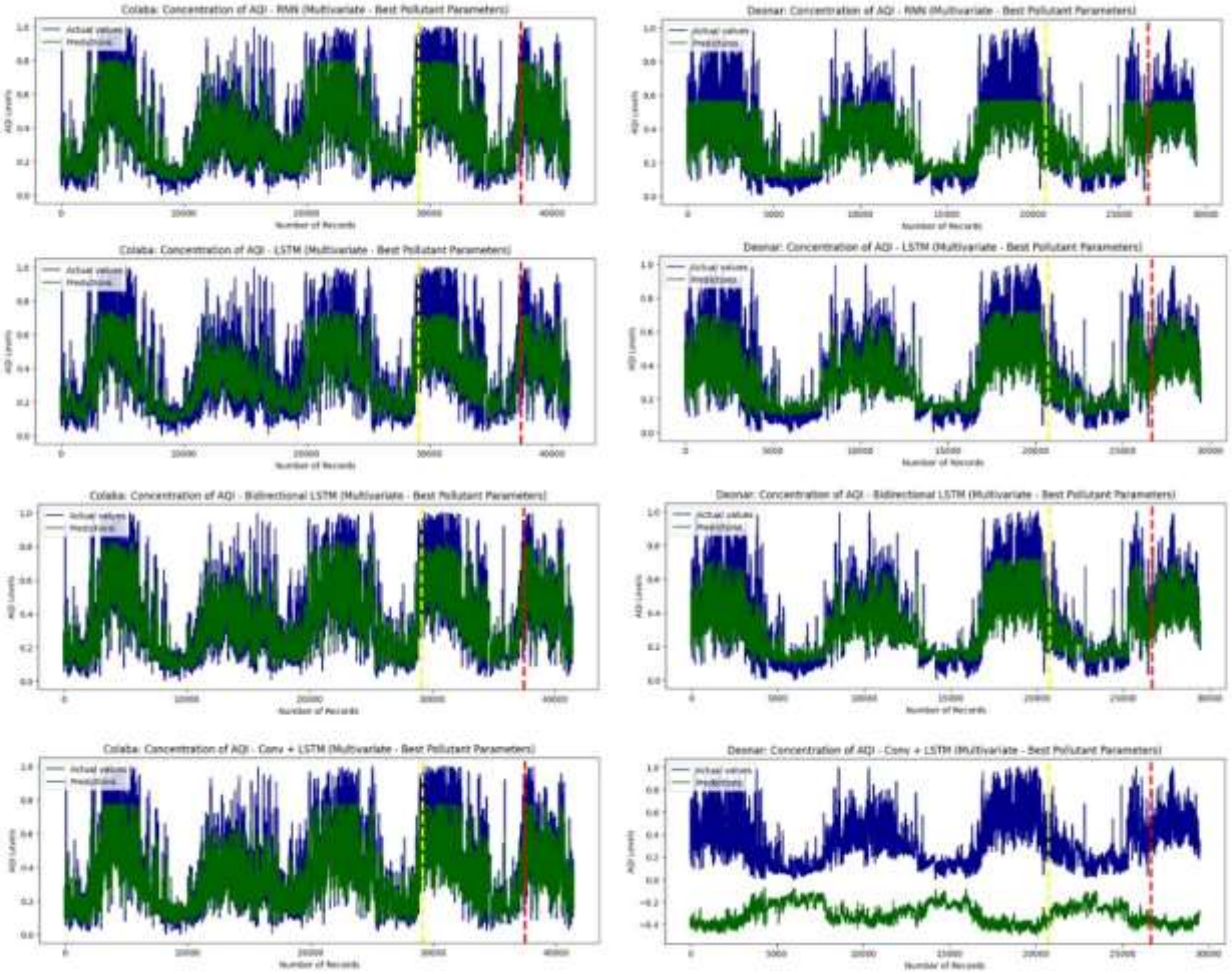


Fig. 7: Actual and predicted values of AQI using various deep learning models with Best Pollutant Feature Set.

5.2. Best Pollutant Feature Set

Fig. 7 represents AQI prediction for both stations using multiple deep learning topologies on best pollutant parameters. The actual values in blue and predicted values in green, 70% of dataset is used for training which is presented towards left of yellow dotted line, next 20% is used as validation set which is in between yellow and red dotted lines and remaining 10% is used as test set. It is observed that with pollutant parameters as feature set better predictions were made for Colaba station than Deonar station, this might be due to high spread of AQI values for Deonar station. For Colaba station, RNN and Bidirectional LSTM has made slightly better AQI prediction whereas for Deonar station, LSTM has given significantly better prediction and it is noticeable that Conv1D + LSTM with pollutant feature set has miserably fail.

5.3. Best Meteorological Feature Set

It is clear from Fig. 8, which demonstrates actual and predicted AQI values for all deep learning architectures for both stations using meteorological features, that every model have captured variation in AQI values more precisely with this set, indicating better predicting capability. Closely examining the plots and using climatic parameters, Conv1D + LSTM model has resulted in marginal improvement in making predictions for Colaba station whereas for Deonar station, LSTM has again out performed.

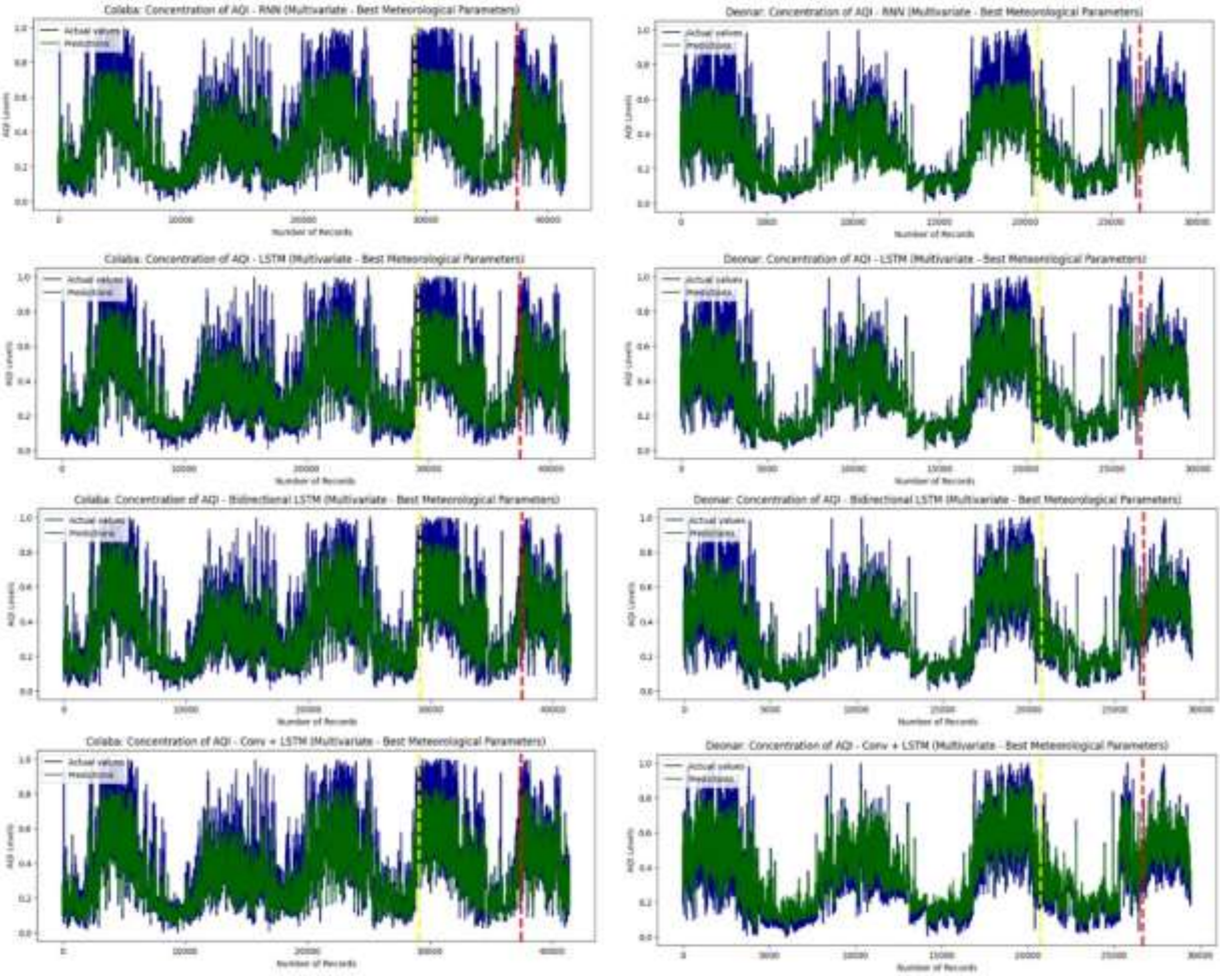


Fig. 8: Actual and predicted values of AQI using various deep learning models with Best Meteorological Feature Set.

5.4. Combined Feature Set with All Parameters

For combined feature set with total of 15 features and target variable, actual and predicted values for all deep learning architectures used for both stations in shown in Fig. 9. For Colaba station similar to meteorological feature set, Conv1D + LSTM model has made best predictions with little improvement over other models. Due to high dispersion in data for Deonar station, there is significant fluctuations in outcome of various models, it is clearly visible that RNN model has completely failed and other models has resulted in similar results with Bi-LSTM being slightly better.

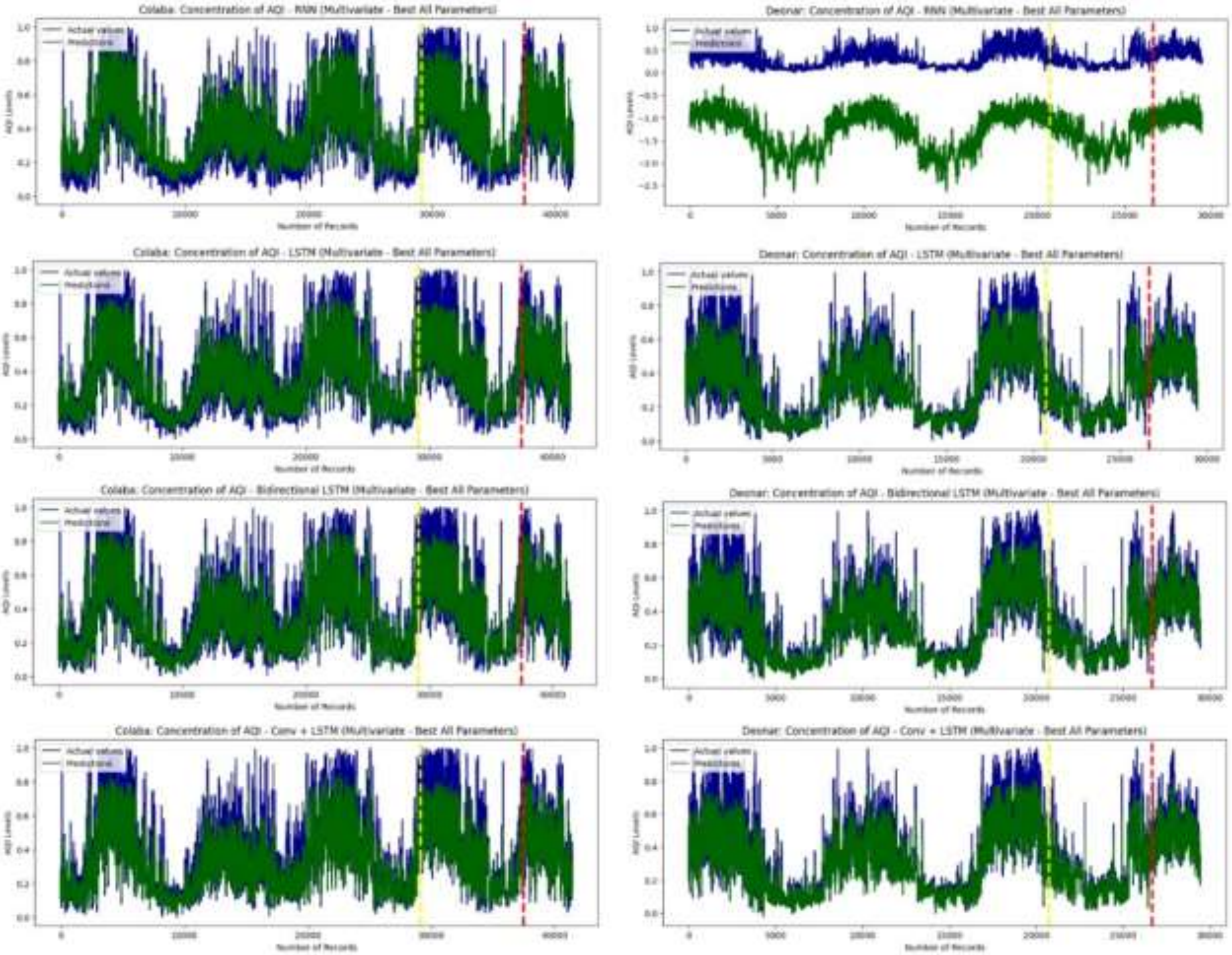


Fig. 9: Actual and predicted values of AQI using various deep learning models with Combined Feature Set with All Parameters.

5.6. Juxtaposed Analysis

Performance of various deep learning architectures for different input feature set is tabulated in Table 3 and Table 4 for Colaba and Deonar stations respectively. These tables include varying performance measures namely, RMSE, R^2 Score, MAPE, and EVS.

Across Models: Bi-LSTM and LSTM models have shown consistent performance across different stations and feature sets. For Colaba, it gave the lowest RMSE and MAPE with several feature sets and the highest R^2 Score and EVS, thus showing better prediction accuracy. At Deonar, the Bi-LSTM model performance was good, especially with meteorological parameters, producing competitive results. While Conv1D + LSTM have some bright spots, considering the overall model reliability and efficiency toward minimizing errors and explaining enhanced variance, Bi-LSTM and LSTM would be the best models, irrespective of the station or feature set.

Across Feature Sets: Clearly, the meteorological set of features outperforms others in general across both Colaba and Deonar stations. Indeed, models with meteorological parameters recorded the lowest RMSE and MAPE and higher values for R^2 Score and EVS, thus showing better variance explanation. The discovery that meteorological factors affect AQI more significantly than pollutant concentrations is important, since these factors directly regulate pollutant spreading and settling. Increased temperatures and humidity have the effect of reducing AQI by increasing atmospheric mixing and facilitating pollutant removal. Conversely, extended high-pressure conditions can elevate AQI by generating stable atmospheric flows that retain pollutants in the lower layers of the atmosphere. Increased wind speeds lower AQI by spreading out pollutants, while wind direction has a weaker, less predictable influence on air quality.

Table 3: Performance Metrics of Various Models across Different Feature Set for AQI Prediction for Colaba Station.

Feature Set	Models	RMSE	R^2 Score	MAPE	EVS
Pollutant Parameters	RNN	21.3230	**0.7619**	0.1268	0.8062
	LSTM	23.6293	0.6164	0.1502	0.7569
	Bi-LSTM	21.0952	0.7551	0.1260	0.8093
	Conv1D + LSTM	22.1642	0.6927	0.1334	0.8029
Meteorological Parameters	RNN	22.1090	0.6553	0.1312	0.7984
	LSTM	21.6329	0.6852	0.1322	0.7962
	Bi-LSTM	21.5187	0.7112	0.1327	0.7984
	Conv1D + LSTM	21.0152	0.7320	**0.1238**	0.8089
All Parameters	RNN	23.5945	0.6796	0.1545	0.8175
	LSTM	22.0414	0.7021	0.1400	0.8069
	Bi-LSTM	21.4796	0.7407	0.1329	0.8139
	Conv1D + LSTM	**20.8821**	**0.7502**	0.1246	**0.8189**

Table 4: Performance Metrics of Various Models across Different Feature Set for AQI Prediction for Deonar Station.

Feature Set	Models	RMSE	R^2 Score	MAPE	EVS
Pollutant Parameters	RNN	30.6543	0.1596	0.1057	0.6140
	LSTM	33.7968	0.2082	0.1407	0.6980
	Bi-LSTM	34.9731	0.1983	0.1548	0.6725
	Conv1D + LSTM	362.7671	-689.501	2.3102	-0.347
Meteorological Parameters	RNN	30.1466	0.2836	0.1156	0.7034
	LSTM	**26.2931**	0.5135	**0.0955**	0.7116
	Bi-LSTM	27.8235	0.4953	0.1015	0.6892

	Conv1D + LSTM	34.4509	0.2719	0.1288	**0.7180**
All Parameters	RNN	588.7852	-96.7797	1.5506	-1.163
	LSTM	26.3419	0.5340	0.0956	0.7088
	Bi-LSTM	26.6565	**0.5510**	0.0977	0.7124
	Conv1D + LSTM	27.4466	0.5264	0.1028	0.7171

Across Locations: Models fare poorly for Deonar primarily due to the high variability in its AQI values, which makes prediction difficult. Such variability is presumably due to numerous factors specific to the area, including its location near mixed sources of pollution like landfills, industrial areas, and busy traffic routes. These sources create fluctuations in pollutant concentrations that are non-linear and more difficult to incorporate using conventional modeling methods. Also contributing to decreased model reliability are inconsistencies and gaps within data as it currently stands. Aggregated, these factors produce more intricate, less predictable air quality patterns in Deonar compared to Colaba, resulting in increased error measures and lower explained variance across models.

6. CONCLUSIONS

Air pollution is one of the prime factors of mortality across the globe, and needs to be monitored and predicted with accuracy to reduce its impacts. Atmospheric pollutants, such as, $PM_{2.5}$, PM_{10}, ammonia, SO_2, CO, NOx, and ozone greatly affect human health. The climatic factors including temperature, humidity, wind speed, and pressure influence the level of the AQI. Research object of this study was dataset from two stations, namely, Colaba and Deonar, containing atmospheric pollutants and meteorological factors. Correlation study shows AQI levels exhibit strong influences from its past values and the have strong positive correlation with atmospheric pollutants and various meteorological conditions, such as, temperature, pressure, wind speed. Other climatic factors, humidity and dew point are negatively associated with AQI levels.

For this study based on results of Pearson's correlation, multiple input feature set: best pollutant set, best meteorological set and combined feature sets are created. For these sets, multiple deep learning topologies, such as, RNN, LSTM, Bi-LSTM and Conv1D + LSTM models were implemented with optimized hyperparameters for both locations. The assessment metrics used were RMSE, R^2 Score, MAPE and EVS. For Colaba station, RNN and Bi-LSTM has outperformed for pollutant feature set whereas Conv1D + LSTM has improved results for other two feature sets. For Deonar station, LSTM has produced best results in all scenarios and RNN with combined feature set have miserably failed. Meteorological factors have enhanced the accuracy, as deep learning model were capable of learning more complex patterns that models to nonlinear associations and interactions.

As different results were obtained for different location, this ensures there in need of more tailored location specific models for more accurate prediction. This research advances by comparing multiple feature set and multiple architecture with location specific information. This information can be helpful for policy makers and environment regulatory authorities to forecast episodes of bad air quality and release necessary early alerts and warnings, which is needed for public safety, support urban planning and eventually reduce the burden on the healthcare infrastructure of the city.

Every region has some sources of pollution, geographical features, climatic conditions, and human activities that affect air quality differently. Generic models, though helpful, may lack such local variations, hence leading to less accurate predictions. In future, studies can include other factors such as proximity to industrial areas, density of traffic, vegetation cover, distance from landfill, distance from sea shore etc can all have impact on the dispersion and concentration of pollutants in air. Attention based systems, variational autoencoders, probabilistic graph networks with a greater number of training samples and more inclusive feature set can further enhance prediction accuracy.

Author Contributions: All authors contributed to the study conception and design. Research flow, architecture designs was performed by [Darakhshan Khan]. [Dr. Arhcana Patankar] have performed interpretation of results. Data Collection and preprocessing were performed by [Himani Deshpande] and [Juhi Janjua]. The first and final draft of the manuscript was written by [Darakhshan Khan] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments: The authors recognize the National Environmental Engineering Research Institute (NEERI) in Nagpur, the Central Pollution Control Board, State Pollution Control Boards, and Pollution Control Committees for making data freely available for research purpose.

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

1. Ayus, I., Natarajan, N., Gupta, D. (2023). Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from china. Asian Journal of Atmospheric Environment, 17(4), , https://doi.org/10.1007/s44273-023-00005-w.
2. Barve, Atharva & Singh, Vishwa & Shrirao, Shivam & Bedekar, Mangesh. (2020). Air Quality Index forecasting using parallel Dense Neural Network and LSTM cell. 1-4. 10.1109/INCET49848.2020.9154069.
3. Binbusayyis, A., Khan, M., Ahmed, A., M.M. (2024). A deep learning approach for prediction of air quality index in smart city. Discover Sustainability, 5, 89, https://doi.org/10.1007/s43621-024-00272-9.
4. Bizualem, B., Angassa, K., Shibeshi, N. T., Feyisa, G. L. (2023). Assessing the status of air pollution related to traffic using dispersion modeling: the case of Megenagna, Addis Ababa, Ethiopia. Bulletin of Atmospheric Science and Technology, 4(1), 9. https://doi.org/10.1007/s42865-023-00062-7.
5. Board, C.P.C. (2024). Cpcb home page. https://app.cpcbccr.com/AQI India/. (Accessed 24 June 2024).
6. Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45–76. https://doi.org/10.28945/4184.

7. Cui, H., Ma, R., Gao, F. (2018). Relationship between meteorological factors and diffusion of atmospheric pollutants. Chemical Engineering Transactions, 71, 1417–1422. https://doi.org/10.3303/CET1871237.
8. DeVries, R., Kriebel, D., & Sama, S. (2017). Outdoor Air Pollution and COPD-Related Emergency Department Visits, Hospital Admissions, and Mortality: A Meta-Analysis. COPD, 14(1), 113–121. https://doi.org/10.1080/15412555.2016.1216956.
9. Douglas, C., Cheryl, L., Murat, K. (2015). Introduction to time series analysis and forecasting (2nd ed.). Wiley.
10. Fang, C., Lu, H., Li, G., Sun, D., & Miao, Z. (2015). Estimating the Impact of Urbanization on Air Quality in China Using Spatial Regression Models. Sustainability, 7(11), 15570–15592. https://doi.org/10.3390/su71115570.
11. For Air Quality Management Delhi NCR, C.C.R. (2024). Air quality index information - aqi calculator. https://app.cpcbccr.com/ccr docs/AQI%20-Calculator.xls. (Accessed 29 June 2024).
12. Gou, A., Tan, G., Ding, X. (2023). Urban-rural difference in the lagged effects of pm2.5 and pm10 on copd mortality in chongqing, china. BMC Public Health, 23, 1270, https://doi.org/10.1186/s12889-023-16113-9.
13. Gupta, S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., Ganeshan, A. (2023). Prediction of air quality index using machine learning techniques: A comparative analysis. Journal of Environmental and Public Health, 1-26, https://doi.org/10.1155/2023/4916267.
14. Jian, Z., Cai, J., Chen, R. (2024). A bibliometric analysis of research on the health impacts of ozone air pollution. Environmental Science and Pollution Research, 31, 16177–16187, https://doi.org/10.1007/s11356-024-32233-0.
15. Jiao, L., & Zheng, W. (2020). Multi-step traffic flow prediction method based on Conv1D + LSTM. 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 113–118. doi:10.1109/BigComp49089.2020..
16. K., K., S.K., A., Ravindiran, D.R.G. (2024). Integrating machine learning techniques for air quality index forecasting and insights from pollutant-meteorological dynamics in sustainable urban environments. Earth Science Informatics, , https://doi.org/10.1007/ s12145-024-01382-8.
17. Kara, Y., S¸ evik, S.Y., Toros, H. (2024). Comprehensive analysis of air pollution and the influence of meteorological factors: a case study of adiyaman province. Environmental Monitoring and Assessment, 196, , https://doi.org/10.1007/s10661-024-12649-4.
18. Khan, D., Patankar, A., Motwani, R., Ailsinghani, A. (2023). Investigating efficacy of rnn and its variants for aqi prediction. 2023 international conference on advances in computation, communication and information technology (icaiccit) (p. 1303-1307).
19. Kleingchuay,W., Srimanus,W., Srimanus, R., et al. (2022). The association of meteorological parameters and airq+ health risk assessment of pm2.5 in ratchaburi province, thailand. Scientific Reports, 12, 12971, https://doi.org/10.1038/s41598-022-17087-1.
20. Le, V.-D., Bui, T.-C., & Cha, S.-K. (2020). Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 55–62. doi:10.1109/BigComp48618.2020.00-99.
21. Liu, Y., Wang, P., Li, Y. (2022). Air quality prediction models based on meteorological factors and real-time data of industrial waste gas. Scientific Reports, 12, 9253, https:// doi.org/10.1038/s41598-022-13579-2.
22. Lu, Y., & Li, K. (2023). Multistation collaborative prediction of air pollutants based on the cnn-bilstm model. Environmental Science and Pollution Research, 30, 92417–92435, https://doi.org/10.1007/s11356-023-28877-z.
23. Mandal, A., & Sen, R. (2024). Medium-term aqi prediction in selected areas of Bangladesh based on bidirectional gru network model. SN Computer Science, 5, 627, https:// doi.org/10.1007/s42979-024-02971-6.

24. Mandal, N., & Sarode, T. (2024a). A framework for cloud cover prediction using machine learning with data imputation. International Journal of Electrical and Computer Engineering (IJECE), 14(1), 600-607, https://doi.org/10.11591/ijece.v14i1.9538.
25. Mandal, N., & Sarode, T. (2024). A framework for cloud cover prediction using machine learning with data imputation. International Journal of Electrical and Computer Engineering (IJECE), 14(1), 600–607. https://doi.org/10.11591/ijece.v14i1.pp600-607.
26. M´endez, M., Merayo, M.G., N´u˜nez, M. (2023). Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 56, 10031–10066, https://doi.org/ 10.1007/s10462-023-10424-4.
27. Naz, F., McCann, C., Fahim, M., Cao, T.-V., Hunter, R., Viet, N.T., . . . Duong, T.Q. (2023). Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas. IEEE Access, 11, 64015-64025, https://doi.org/10.1109/ACCESS.2023.3289153.
28. Nikpour, P., Shafiei, M., Khatibi, V. (2024). Gelato: a new hybrid deep learning-based informer model for multivariate air pollution prediction. Environmental Science and Pollution Research, 31, 29870–29885, https://doi.org/10.1007/s11356-024-33190-4.
29. OpenWeather (2024). Openweather homepage. https://openweathermap.org. (Accessed 21 April 2024).
30. Petrus, M., Popa, C., Bratu, A.-M. (2022). Ammonia concentration in ambient air in a periurban area using a laser photoacoustic spectroscopy detector. Materials, 15(9), 3182, https://doi.org/10.3390/ma15093182.
31. Plevris, V., Solorzano, G., Bakas, N., & Ben Seghier, M. E. A. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. Proceedings of the 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022), 155. https://doi.org/10.23967/eccomas.2022.155.
32. Poyyamoli, G., & Boss, C. (2014). Assessment of air pollution and its impacts near municipal solid waste dumping site Kammiyampet, Cuddalore, India. International Journal of Innovative Research in Science, Engineering and Technology, 3(5), 12588–12593.
33. Rangaswamy, P. S., Vasareddy, A., Saiteja, B., Sreeja, B., & Chandana, G. (2022). Impact assessment on air quality around integrated municipal solid waste management plant in Hyderabad. International Journal of Advanced Research in Science, Communication and Technology, 2(1), 666–677. https://doi.org/10.481.
34. Saadi, D., Tirosh, E., Schnell, I. (2021). The relationship between city size and carbon monoxide (co) concentration and their effect on heart rate variability (hrv). International Journal of Environmental Research and Public Health, 18(2), 788, https://doi.org/10.3390/ijerph18020788.
35. Sajjad Abdollahpour, S., Qi, M., Le, H.T.K., Hankey, S. (2024). Urban spatial structure and air quality in the united states: Evidence from a longitudinal approach. Environment International, 190, 108871, https://doi.org/10.1016/j.envint.2024.108871 Retrieved from https://doi.org/10.1016/j.envint.2024.108871
36. Sch¨urholz, D., Kubler, S., Zaslavsky, A. (2020). Artificial intelligence-enabled context-aware air quality prediction for smart cities. Journal of Cleaner Production, 271(121941),https://doi.org/10.1016/j.jclepro.2020.121941.
37. Shankar, L., & Arasu, K. (2023). Deep learning techniques for air quality prediction: A focus on pm2.5 and periodicity. Migration Letters, 20(S13), 468–484, https://doi.org/10.33182/ml.v20iS13.4763.
38. Shi, T., Li, P., Yang, W. (2023). Application of tcn-bigru neural network in pm2.5 concentration prediction. Environmental Science and Pollution Research, 30, 119506–119517, https://doi.org/10.1007/s11356-023-30354-6.

39. Song, Q., Zou, J., Xu, M. (2023). Air quality prediction for chengdu based on long shortterm memory neural network with improved jellyfish search optimizer. Environmental Science and Pollution Research, 30, 64416–64442, https://doi.org/10.1007/s11356-023-26782-z.
40. Sterling, M. (2017). Air quality: Population density and transportation emissions. The Equilibrium, 3(1). https://doi.org/10.5070/Q23141221.
41. Sun, R., Zhou, Y., Wu, J., Gong, Z. (2019). Influencing factors of pm2.5 pollution: Disaster points of meteorological factors. International Journal of Environmental Research and Public Health, 16(20), 3891, https://doi.org/10.3390/ijerph16203891.
42. Thakur, N., Karmakar, S., Shrivastava, R. (2023). Hybrid deep learning algorithms for predicting air quality index using dimension reduction technique in search of precise results. International Journal of Information Technology, 15, 3181–3187, https://doi.org/10.1007/s41870-023-01350-3.
43. Tian, J., Fang, C., Qiu, J., Wang, J. (2021). Analysis of ozone pollution characteristics and influencing factors in northeast economic cooperation region, china. Atmosphere, 12(7), 843, https://doi.org/10.3390/atmos12070843.
44. Varaprasad, V., Kanawade, V. P., & Narayana, A. C. (2024). Association between sea-land breeze and particulate matter in five coastal urban locations in India. Science of The Total Environment, 913, 169773. https://doi.org/10.1016/j.scitotenv.2023.169773.
45. Wang, P., Bi, X., Zhang, G. (2023). A new hybrid pm volatility prediction model based on emd and machine learning algorithms. Environmental Science and Pollution Research, 30, 82878–82894, https://doi.org/10.1007/s11356-023-26834-4.
46. Wang, S., Ren, Y., Xia, B. (2023). Estimation of urban aqi based on interpretable machine learning. Environmental Science and Pollution Research, 30, 1-13, https://doi.org/10.1007/s11356-023-29336-5.
47. Wu, H., Hong, S., Hu, M., Li, Y., Yun, W. (2022). Assessment of the factors influencing sulfur dioxide emissions in shandong, china. Atmosphere, 13(1), 142, https://doi.org/10.3390/atmos13010142 .
48. Yan, H., Li, Q., Feng, K. (2023). The characteristics of pm emissions from construction sites during the earthwork and foundation stages: an empirical study evidence. Environmental Science and Pollution Research, 30, 62716–62732, https://doi.org/10.1007/s11356-023-26494-4 .
49. Yan, M., & Li, T. (2024). A review of the interactive effects of climate and air pollution on human health in china. Current Environmental Health Reports, 11, 102–108, https://doi.org/10.1007/s40572-024-00432-z .
50. Zhan, C., Li, S., Li, J., Guo, Y., Wen, Q., & Wen, W. (2020). Prediction of air quality in major cities of China by deep learning. Proceedings of the 2020 16th International Conference on Computational Intelligence and Security (CIS), 68–72. https://doi.org/10.1109/CIS52066.2020.00023.
51. Zhang, C., & Zhang, L. (2024). The relationship between toxic air pollution, health expenditure, and economic growth in the european union: fresh evidence from the pmg-ardl model. Environmental Science and Pollution Research, 31, 21107–21123, https://doi.org/10.1007/s11356-024-32342-w .
52. Zhang, Y., & Jiang, W. (2018). Pollution characteristics and influencing factors of atmospheric particulate matter (pm2.5) in chang-zhu-tan area. IOP Conference Series: Earth and Environmental Science, 108, 042047, https://doi.org/10.1088/1755-1315/108/4/042047 .
53. B. Cui, M. Liu, S. Li, Z. Jin, Y. Zeng, and X. Lin, "Deep learning methods for atmospheric PM2.5 prediction: A comparative study of transformer and CNN-LSTM-attention," Atmospheric Pollution Research, vol. 14, no. 9, pp. 101833, 2023. doi: 10.1016/j.apr.2023.101833.

54. Wu, Q., Geng, Y., Wang, X. et al., "A novel deep learning framework with variational auto-encoder for indoor air quality prediction," Frontiers in Environmental Science & Engineering, vol. 18, no. 8, 2024, doi: 10.1007/s11783-024-1768-7.