International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 1137-1152
https://doi.org/10.70917/ijcisim-2026-5106
Article

Temporal Feature Interaction Memory in Air Quality Multi Variate Time Series Prediction

M. Bavani[1], M. Gethsiyal Augasta[2]

[1]Research Scholar, Kamaraj College, (Autonomous), Thoothukudi
[2]Assistant Professor, Kamaraj College, (Autonomous) Thoothukudi, Tamil Nadu, India
[1]bavanim25@gmail.com, [2]dr.gethsiyalaugasta@kamarajcollege.ac.in

Abstract: Due to the increasing number of vehicles and factories, air pollution has become a critical issue in most urban areas. Accurate prediction of air quality plays a vital role in managing and mitigating pollution, especially in metropolitan cities. Forecasting air pollution levels enables governments to take proactive measures to reduce its adverse effects. In this paper, we propose a novel attention-based approach to predict Particle Matter (PM2.5 and PM10) concentrations using Recurrent Neural Networks (RNNs). The proposed method incorporates two attention mechanisms-temporal attention and feature attention- to enhance predictive accuracy. Temporal attention identifies and emphasizes key time steps within the input sequence, while feature attention assigns greater weight to the most influential air quality variables. This dual-attention strategy is integrated with deep learning and gradient boosting techniques. Unlike conventional attention mechanisms that assign independent importance weights to features and time steps, the proposed Temporal Feature Interaction Memory (TFIM) module explicitly models nonlinear interactions between heterogenous environmental variables across temporal delays. This module enables the network to capture delayed cross-feature dependencies critical for pollution formation dynamics, which are not represented in existing attention-based air quality prediction models. Specifically, the Kaggle Air Quality dataset is employed to train a Bidirectional Long Short-Term Memory (BiLSTM) model augmented with TFIM and Temporal Feature Attention (TFA), and the predictions are further refined using XGBoost to achieve improved accuracy. The proposed method achieves a dramatic reduction in RMSE, MAE and MAPE compared to baseline models. The proposed method achieved 1.46 RMSE and 5.88 MAE for PM2.5 predictions and 1.49 RMSE and 1.37 MAE for PM10 predictions.
Keywords: Prediction, attention, temporal, bidirectional long short-term memory

1. Introduction

India is the 7th largest country by area and 2nd most populated country in the world. The reports prepared by IQAir revels that India is 3rd most polluted country after Bangladesh and Pakistan, on the basis of fine particulates (PM2.5) concentration for the year 2020.

Air pollution levels have risen as an outcome of urban and industrial development in so many developing countries. People and governments all around the world are concerned about air pollution, which has a severe influence on both personal health and long-term global development. As a government, it is responsible for preventing and controlling air pollution, as well as monitoring the pollutant's impacts on human health. There are numerous computer models available, ranging from statistics to artificial intelligence. Pollution levels are still out of control in some parts of the world due to a wide range of sources and factors.

Because of accurate estimates of future air pollution, the government can take necessary action. Forecasting air pollution levels based on environmental data is becoming increasingly relevant as people become more worried about global warming and urban sustainability. For replicating the complicated linkages between these variables, advanced Deep Learning (DL) algorithms hold enormous promise. The objective of this work is to provide a high level of accurate solution to the air pollution forecasting problem. Kaggle data will be employed to train a DL model that will forecast air pollution levels.

This paper introduces a novel TFA mechanism for air quality prediction which integrates temporal attention and feature attention to enhance model performance. To know the importance of temporal data, initially, without focusing on the most important time series, 1 to 10 days prediction is made. Figures 1 and 2 show the PM2.5 and PM10 predictions using 1 to 10 days data.

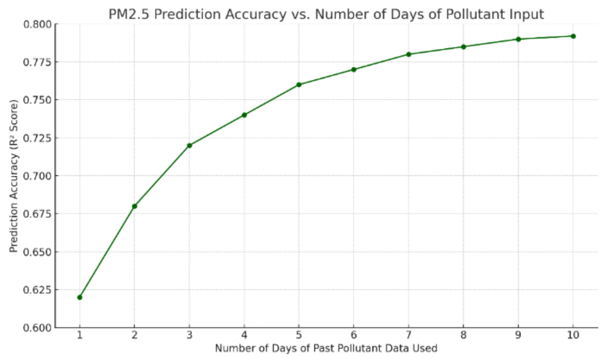

Fig. 1 PM2.5 Prediction Accuracy Vs Number of Past Data

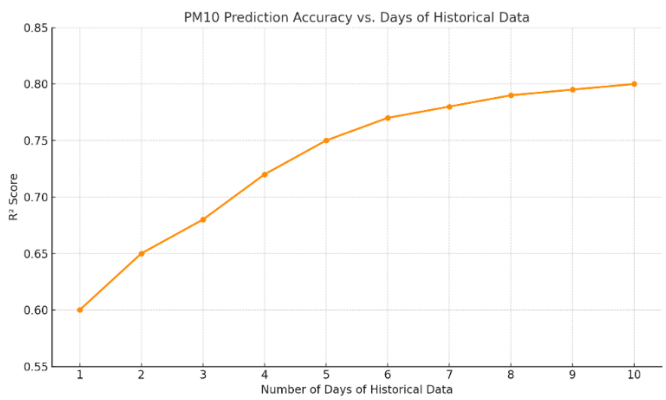

Fig. 2 PM10 Prediction Accuracy Vs Number of Past Data

The chart illustrates how the accuracy of PM2.5 and PM10 predictions improves as more days of historical pollutant data are used as input. The limit of the historical data cannot be determined. Here comes the need of temporal attention. The temporal attention identifies and emphasizes the most critical time steps within the input sequence. The feature attention module assigns weights to key air quality parameters (e.g., NO2, CO), capturing their relative importance in forecasting.

The PM2.5 rises when humidity is high and wind is low. The proposed method explicitly learns humidity at t-3 and wind at t-1. This interaction is stored in memory. Existing models do not explicitly store this interaction. This new learning mechanism uses interaction memory across features and time. Existing models always follows non-linear interaction dynamics. The proposed method explicitly models interaction function. The existing attention only model importance and not interaction memory.

The main contributions of this paper include:

- Temporal Feature Interaction Memory (TFIM) to preserve temporal-feature interaction history
- Dual attention to improve interpretability and feature selection
- Hybrid deep learning + gradient boosting refinement

- Improved long-range dependency modeling

2. Literature Survey

A framework has been built that categorizes ten well-defined models, implements them under identical execution environment and forecasts succeeding values [1]. Wang et al. [2] have developed a VMD-GAT-BiLSTM framework combining Variational Mode Decomposition (VMD), Graph Attention Network (GAT) and BiLSTM, achieving notable improvements in PM2.5 forecasting by effectively capturing spatial dependencies across monitoring stations.

Another study has been carried using various statistical and deep learning methods to forecast long-term pollution trends for PM2.5 and PM10 [3]. The study is based on Kolkata, a major city on the eastern side of India. This method has produced a forecast for the next two years using different statistical and deep learning methods. The findings reflect that statistical methods and Holt–Winters outperform deep learning methods such as stacked, bi-directional, auto-encoder and convolution LSTM networks based on the limited data available.

Mengara et al. have developed an attention-based convolutional BiLSTM autoencoder [4], incorporating local spatial patterns and temporal attention to forecast PM2.5/PM10 in Busan, Korea. The model has showed superior accuracy, especially when including traffic data. The quality of air in six Indian cities is predicted using data-driven Artificial Neural Network (ANN) [5]. Two machine learning models namely Artificial Intelligence and Gaussian Process Regression (GPR) has been developed.

Hossen et al. have introduced an ODE-based Transformer-LSTM hybrid for PM2.5 [6], applying self-attention (query-key-value) to model inter-time-step dependencies with promising results. To provide a comprehensive analysis of severe implications of poor air quality on public health in India, three machine learning models such as linear regression, decision tree regression and k-NN has been considered [7]. Dong et al.[8] have combined Empirical Mode Decomposition (EMD) with a Transformer-BiLSTM architecture for ultra-short term-AQI forecasting in Patna, India, demonstrating low RMSE/MAE for 5-hour horizons.

To provide a full picture of the pollution levels, DL methods are used to determine the Air Quality Index [9]. The link between different contaminants and the AQI is modeled using logistic regression. To further enhance prediction accuracy, the k-NN algorithm is used, which takes advantage of the closeness of data points. This study seeks to improve the accuracy and efficiency of air pollution prediction by employing these approaches.

Lu et al.[10] have implemented FALNet, a Frequency-Aware Attention-LSTM model using FFT-based denoising along with multi-head attention to address pollution peaks effectively. AQI is forecasted in three scenarios based on the air pollutants data [11]. Monthly average NO2, SO2, O3 and PM2.5 data from 1987 to 2020 were included. Bodendorfer presented HEART, a Hybrid Enhanced Autoregressive Transformer combining CNN, autoregressive layers, and attention modules to forecast NO2, O3, PM10 and PM2.5-achieving ~7.5% average MSE reduction [12].

A new lightweight DL model using Sparse attention-based Transformer Networks (STN) consisting of encoder and decoder layers, in which a multi-head sparse attention mechanism is adopted to reduce the time complexity, has been developed to learn long-term dependencies and complex relationships from time series PM2.5 data for modeling air quality forecasting [13]. Sreenivasulu and Rayalu [14] have developed STL-CNN-BiLSTM-AM, which integrates Seasonal-Trend Decomposition (STL), CNN, BiLSTM and attention [14]. Their hybrid model reduced RMSE by ~30% and boosted R2 by ~0.46% compared to non-attention baselines.

Traditional time-series forecasting models, like LSTM and Temporal Convolutional Network (TCN), were found to be efficient in atmospheric pollutant estimation, but either the model accuracy was not high enough or the models encountered certain challenges due to their own structure or some specific application scenarios [15]. This study has developed a high-accuracy, hourly PM2.5 forecasting model, poly-dimensional local-LSTM Transformer by DL methods, based on air pollutant data and meteorological data, and aerosol optical depth data retrieved from the

Himawari-8 satellite. Liu et al. have developed an EEMD-ALSTM, which integrates Ensemble Empirical Mode Decomposition (EEMD) with an attention-augmented LSTM [16]. By decomposing the PM2.5 series prior to attention-based modelling, the EEMD-ALSTM reported ~15% improvements in MAE and RMSE compared to non-decomposed LSTM models.

We et al. have developed MIC-CEEMDAN-CNN-BiGRU, combining Complete Ensemble Empirical Mode Decompostion with Adaptive Noise (CEEMDN), feature selection, CNN and BiGRU [17]. Their model saw ≈11% RMSE reduction and near 5% point R2 improvement over standalone models.

To address the challenges of data redundancy and diminished long-term prediction accuracy observed in previous studies, an innovative approach has been developed to predict air pollutant concentrations leveraging advanced data analysis and DL methods [18]. This method integrates the k-NN, Spatio-Temporal Attention (STA) mechanism, the residual block, and convolutional LSTM (ConvLSTM). The effectiveness of the KSC-ConvLSTM approach was validated through predictions of PM2.5 concentrations in Beijing and its surrounding urban agglomeration.

Zhang et al.[19] have introduced TDGTN, a graph-transformer model using temporal-difference graph attention layered in encoder-deocder for PM2.5 forecasting in China, showing superior short-and long-term predictions. Hyper Heuristic Multi-Chain Model (H2MCM) has been developed to project future air quality, considering various meteorological factors and pollution-related variables like atmospheric pressure, temperature, humidity, and wind patterns[20]. Leveraging 12 units of LSTMs, H2MCM accurately predicts forthcoming air pollutants concentrations such as PM2.5, CO and NO2. H2MCM utilizes a multi-chain mechanism, employing 1-hour prediction models to forecast air quality hourly, enabling approximations for the next 12 hours.

Necula et al.[21] have combined SARIMA preprocessing with BiLSTM and Fourier-based attention, demonstrating improved interpretability and robustness, particularly under extreme pollution events. Even though many existing techniques have predicted AQI, enhancement is required in prediction algorithms with minimized loss. To achieve prediction efficiency with reduced loss, pre-processing of input data is being performed using Deep GAN (Generative Adversarial Network)[22]. It performs the imputation of data in place of missing values to improve accurate prediction. Additionally, feature scaling normalizes independent real-data features to a fixed scale. With the processed data, regression is done through modified Stacked Attention GRU with KL divergence, which predicts Ernakulam, Chennai and Ahmedabad cities with higher, medium and low levels of AQI in India

A manual and web-based automatic prediction system has been implemented that provides real-time alerts on air quality status and can help prevent premature deaths, chronic diseases and other health problems [23]. Air pollutants, including CO, O3, NO2 and PM2.5, are used in this study for feature analysis and extraction. Data preprocessing was performed before feeding the data into the machine learning models for feature correlation and evaluation. Machine learning, feature extraction, classification and regression are discussed clearly [24]. Table 1 summarizes few models in air quality prediction that uses deep learning.

Table 1 Summary of Recent Literature on Air Quality Prediction

Category	Model	Pros	Cons
Hybrid Deep Learning with Spatial-Temporal Attention (CNN-LSTM-Transformer Combinations)	VMD-GAT-BiLSTM [2], CNN-BiLSTM Autoencoder [4]	- Captures both local (CNN) and long-term dependencies (LSTM/Transformer). - Achieves state-of-the-art accuracy in complex datasets.	- High model complexity and training time. - Risk of overfitting with limited data

Decomposition Preprocessing (STL,EMD, CEEMAN, FFT)	STL-CNN-BiLSTM-AM [14], EEMD-ALSTM [16], CEEMDAN-CNN-BiGRU [17]	- Separates noise, trends and seasonal patterns for clearer learning. - Improves robustness in fluctuating pollution data	- Add preprocessing overhead. - May distort original time series if decomposition is not parameter-tuned.
Graph-Based Attention for Spatial Correlation (GAT, Graph Transformer)	VMD-GAT-BiLSTM [2], TDGTN [19], BiLSTM and Fourier-based Attention[21]	- Effectively models spatial relationships across monitoring stations. - Useful for regional air quality forecasting.	- Requires high-quality spatial data - Graph construction can be computationally intensive.
Self-Attention and Multi-Head Attention (Transformers)	Transformer LSTM Hybrid [6] EMD Transformer-BiLSTM [8]	- Excels at capturing long-range temporal dependencies. - Scales well to large datasets.	- Resource intensive - Less interpretable than simpler attention mechanisms.
Feature Attention for pollutant importance	STL-CNN-BiLSTM-AM [14], MIC-CEEEMDAN-CNN-BiGRU [17]	- Improves interpretability by identifying key pollutants - Enables feature selection for lighter models	- Sensitive to noise in irrelevant features. - May require normalization for feature scale differences.
Auxiliary Data Integration (Traffic, Weather, Satellite).	CNN-BiLSTM Autoencoder with Traffic Data [4]	- Enhances prediction accuracy by providing contextual information. - Helps handle missing pollutant readings.	- Data fusion is complex. - Missing or inconsistent auxiliary data can degrade model performance.

Existing air quality prediction studies have primarily focused on conventional statistical and deep learning models, but several limitations remain unresolved. The comprehensive review by [1] highlights that although attention-based and recurrent models improve prediction accuracy, most methods treat temporal and feature relationships independently and do not explicitly model their dynamic interactions over time. Nath et al. [3] demonstrated the effectiveness of LSTM-based models for long-term pollution forecasting; however, their architecture relies on fixed hidden representations and lacks a mechanism to preserve historical temporal-feature dependencies that evolve across different time scales. Similarly, 25] applied machine learning techniques such as XGBoost and other regressors for PM2.5 prediction, but these models depend heavily on manually engineered features and do not incorporate adaptive deep feature extraction or attention-based interaction learning. Furthermore, existing dual attention models assign importance weights to time steps and features independently, without maintaining a dedicated memory structure to capture how feature importance evolves temporally. This limitation leads to information loss, reduced interpretability and suboptimal modelling of complex pollutant dynamics.

To address this gap, the proposed architecture introduces a novel TFIM module integrated within the BiLSTM-based dual attention framework. Unlike existing methods, TFIM explicitly learns and preserves the joint evolution of temporal and feature interactions by maintaining a structured memory representation that dyanamically updates across time steps. This allows the model to capture long-range dependencies and evolving pollutant relationships more effectively. The dual attention mechanism further enhances interpretability by identifying critical time periods and influential features, while the XGBoost refinement stage improves prediction robustness by correcting residual nonlinear errors. Consequently, the proposed architecture overcomes the limitations of existing LSTM-based and machine learning models by jointly modelling temporal dependencies, feature interactions, and memory-based contextual evolution, leading to more accurate and interpretable PM2.5 and PM10 predictions.

3. Proposed Methodology

Incorporating an attention mechanism into a sequence-to-sequence model for multivariate time series prediction allows the model to focus on different parts of the input sequence when making predictions. The attention mechanism assigns different weights to each time step in the input sequence, which helps capture complex temporal dependencies more effectively. Existing attention mechanisms only assign importance weights. They do not explicitly model how features influence each other across time. But in air pollution:

PM2.5 at time t depends on

- Humidity at t-3
- Wind speed at t-1
- Temperature at t-5
- Interactions between them

This is called temporal cross-feature interaction memory.

The proposed air quality prediction system also assigns different weights to each pollutant affecting air quality. The attention layer now takes encoder outputs and decoder outputs as inputs. The output of the attention layer is then passed to the dense layer to create BiLSTM features. These features are used by XGBoost to predict the future sequence. The proposed system architecture is shown in Fig. 3.

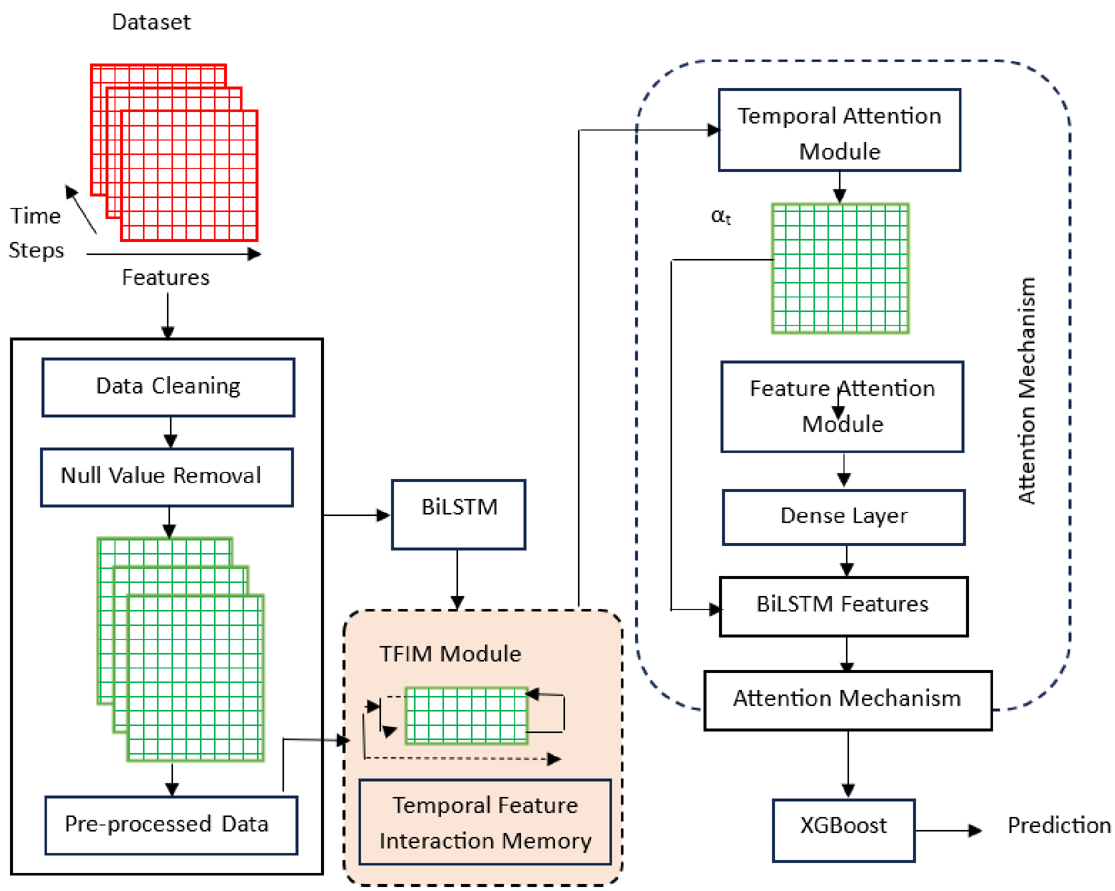

Fig. 3 Proposed Architecture of Air Quality Prediction with TFIM Module

Initially the data is pre-processed by removing null values, followed by restructuring the data into sequences suitable for input to the BiLSTM model. The proposed framework consists of five major stages:

1. BiLSTM Temporal Feature Extraction
2. TFIM module
3. Dual attention mechanism
4. Dense prediction Layer
5. XGBoost Refinement Module

BiLSTM Temporal Feature Extraction

The encoder processes the input sequence using a BiLSTM, generating both the complete output sequence and the final hidden states. The decoder, also implemented with a BiLSTM, is initialized with the encoder's final states to process the target output sequence. The following sections describe the proposed attention mechanism in a step-by-step manner.

Proposed TFIM-BiLSTM Mechanism

The input sequence is defined as $X \in R^{B\times T\times F}$. where B is the batch size, T represents the sequence length (time steps) and F is the number of features. The proposed TFIM mechanism is integrated into BiLSTM network to learn both temporal and feature dependencies for air quality prediction.

A. **BiLSTM Temporal Feature Extraction**

The encoder processes the input sequence using a BiLSTM to capture both forward and backward temporal dependencies:

$$\overrightarrow{h_t}, h_t^{\leftarrow} = BiLSTM(X_t), t = 1,2, \dots, T \quad (1)$$

Where $\overrightarrow{h_t}$ and $h_t^{\leftarrow}$ are the forward and backward hidden states at time step t. The concatenated encoder output is given by:

$$H_t = [\overrightarrow{h_t}; h_t^{\leftarrow}] \quad (2)$$

This improves:

- Long-term dependency modelling
- Representation richness
- Gradient stability

Thus, BiLSTM provides a strong base temporal encoding for downstream modules.

B. **TIFM Module**

Existing attention mechanisms assign importance weights but do not explicitly preserve temporal-feature interaction memory. To address this, a TFIM memory at time step t is computed as

$$M_t = \sum_{\Delta=1}^{K} \quad tanh((W_1 H_t) \odot (H_{t-\triangle} W_2)) \quad (3)$$

Where $M_t \in R^d$ is the interaction memory, $W_{1,}, W_2 \in R^{d\times d}$ are learnable parameters and K is the temporal interaction window. The enhanced representation becomes:

$$H_t' = H_t + M_t \quad (4)$$

This preserves temporal cross-feature dependencies. The enhanced feature matrix is

$$H' = \{H'_1, H'_2, \dots, H'_T\} \quad (5)$$

TFIM introduces

- A structured cross-temporal feature interaction operator
- Memory accumulation across delay window K
- Nonlinear interaction mapping

Thus, TFIM enhances representational capacity beyond simple temporal encoding.

C. **Temporal Attention Module**

The temporal attention module computes attention weights over time steps to identify the most relevant portions of the sequence. The attention score for each time step is computed as

$$e_t = tanh(W_t H'_t + b_t) \quad (6)$$

Where $W_t \in R^{d\times d}, b_t \in R^d, and\ v_t \in R^d$ are trainable parameters. The temporal attention weights are obtained by applying a softmax function:

$$\alpha_t = \frac{exp(e_t^T \omega_t)}{\sum_{k=1}^{T} \ exp(u_k^T \omega_t)} \quad (7)$$

The context vector weighted over time steps is then computed as

$$C_{temp} = \sum_{t=1}^{T} \ \alpha_t H'_t \quad (8)$$

D. **Feature Attention Module**

Feature attention assigns weights to features within each time step. For each feature f, the unnormalized feature attention score is computed as

$$v_f = tanh(W_f H'_f + b_f) \quad (9)$$

Where $W_f \in R^{d\times d}, b_f \in R^d, and\ v_f \in R^d$ are learnable parameters. The feature attention weights are calculated as:

$$\beta_f = \frac{exp(v_f^T \omega_f)}{\sum_{j=1}^{F} \ exp(v_j^T \omega_f)} \quad (10)$$

The temporally weighted encoder output is further refined by applying feature attention:

$$C_{feat} = \sum_{f=1}^{F} \ \beta_f H'_f \quad (11)$$

The process of attention mechanism is illustrated in Algorithm 1.

E. **Combined Context Representation**

The final context vector is obtained as

$$C = [C_{temp};\ C_{feat}] \quad (12)$$

F. Dense Prediction Layer

The deep learning prediction is computed as

$$y_{DL} = W_o C + b_o \quad (13)$$

G. Post-Attention Refinement with XGBoost

Finally, the deep features extracted from the TFIM-BiLSTM are fed into XGBoost, a gradient boosting algorithm, for further refinement and improved predictive accuracy. The XGBoost model is configured with 100 estimators, a learning rate of 0.01, and a maximum tree depth of 5.

The model is trained using hybrid loss:

$$L = \lambda_1 \frac{1}{N} \sum \quad |y - \hat{y}| + \lambda_2 \sqrt{\frac{1}{N} \sum \quad (y - \hat{y})^2} \qquad (14)$$

The following algorithm illustrates the proposed AQP with attention mechanism and TFIM.

Algorithm 1: Proposed AQP

Input: Time series data X, Target PM values Y

Output: Predicted Output y_{final}

Begin

// Pre-process X

1. Remove null values from *D*.

2. Normalize data between (0,1) using min-max normalization

3. Convert *D* into time-series sequence *X*

// BiLSTM Encoding

4. For t=1 to T

5.   Compute forward hidden state

6.   Compute backward hidden state

7.   Concatenation hidden states

8. End for

// TFIM Interaction Memory

9. For t=1 to T

10. For Δ=1 to K

11.   Compute interaction memory

12. End for

13. Update enhanced representation

14. End for

// Temporal Attention

15. Compute attention weights α_t

16. Compute temporal context vector

// Feature attention

17. Compute feature attention β_f

18. Compute feature context vector

19. Concatenate context vectors

// Dense layer prediction

20. Compute y_{DL}

// XGBoost refinement 21. Train XGBoost on deep features 22. Compute residual correction Final prediction 23. $Y_{final} = y_{DL} + residual$ ***Return*** y_{final}

4. Experimental Results

Dataset Description and Performance Measures

The proposed model is experimented on Indian Air Quality Data (2015-2020) [26]. It contains data from January 1, 2015 to February 28, 2021. The dataset contains Air Quality Index (AQI) and air quality data for daily and hourly levels of several stations across multiple cities in India. It consists of pollutants such as PM2.5, PM10, SO2, NO2, CO, O3 etc. The dataset consists of 5 sub datasets: city_day, station_day, city_hour, station_hour and stations. The station_day dataset is used for this study. It consists of the attributes listed in Table 2.

Table 2 Attributes of Station_day Dataset

S. No.	Attribute	Description
1.	StationId	ID given to each station of a city
2.	Date	Date on which the concentrations are obtained
3.	PM2.5	particulate matter of a diameter ≤ 10 µm
4.	PM10	particulate matter of a diameter ≤ 2.5 µm
5.	NO	Nitrogon Oxide
6.	NO2	Nitrogen dioxide
7.	NOx	Total Nitrogen Oxides
8.	NH3	Atmospheric Ammonia
9.	CO	Carbon Monoxide True hourly averaged concentration CO in mg/m3
10.	SO2	Sulphur dioxide
11.	O3	Ozone
12.	Benzene	Benzene
13.	Toluene	Toluene
14.	Xylene	Xylene
15.	AQI	Air Quality Index
16.	AQI_Bucket	AQI classification

In the above table, Station Id, Date, AQI and AQI_Bucket are excluded for prediction. Among various stations, 11 important stations from Kolkata and Hyderabad are taken for this study. The statistical information about the experimented dataset is shown in Table 3.

Table 3 Statistical Information of the Experimented Datasets

City	Station Name	Station Id	No. of Data

Kolkata	Bidhannagar	WB008	7545
	Fort William	WB009	9516
	Rabindra Bharati University	WB011	14463
	Rabindra Sarobar	WB012	7910
	Victoria	WB013	19503
Hyderabad	Bollaram Industrial Area,	TG001	29558
	Central University,	TG002	26722
	ICRISAT Patancheru	TG003	29558
	IDA Pashamylaram,	TG004	33565
	Sanathnagar	TG005	48107
	Zoo Park	TG006	42321

By assuming O_i , P_i as observed and predicted values and $\underline{O_i}$ is the average value of n observed samples, the following metrics Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (R2) are formulated as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - Q_i)^2} \quad (15)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |P_i - Q_i| \quad (16)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|P_i - Q_i|}{Q_i} \quad (17)$$

$$R^2 = 1 - \frac{\sum_{i=}^{n} (P_i - \underline{Q}_i)^2}{\sum_{i=}^{n} (O_i - \underline{Q}_i)^2} \quad (18)$$

Results Analysis

PM2.5 and PM10 are predicted using important features and useful time series. Experiments are carried out for the full dataset. The results obtained for the proposed method to predict PM2.5 with attention mechanism is shown in Table 4.

Table 4 PM2.5 Prediction Results Obtained by the Proposed Method with TFIM Mechanism

Station ID	RMSE	MAE	MAPE	R2
WB008	1.6	1.059	25.51	0.5324
WB009	1.45	1.32	25.54	0.3844
WB011	1.09	1.09	1.19	0.59
WB012	1.88	9.02	1.2	0.76
WB013	1.4	1.4	15.09	0.91
TG001	2.14	13.27	0.198	0.8
TG002	1.82	7.09	0.18	0.79
TG003	1.12	13.43	0.14	0.68
TG004	1.16	8.58	0.17	0.79
TG005	1.01	0.12	0.18	0.6

TG006	1.45	8.34	0.26	0.74

In the above table, it is observed that except TG001 and TF003 stations, the RMSE values of other stations are between 1 and 2. The R2 value is 0.6 for TG005 which is the largest station in the study. The error values are also very low for TG005 whereas the R2 value is 0.5324 for WB008 which is the smallest station used. Table 5 shows the results obtained by TFIM-BiLSTM for predicting PM10.

Table 5 PM10 Prediction Results Obtained by the Proposed Method with TFIM-BiLSTM Mechanism

Station ID	RMSE	MAE	MAPE	R2
WB008	1.48	1.62	16.11	0.6
WB009	1.59	1.28	24.23	0.48
WB011	1.13	1.11	0.15	0.8
WB012	1.72	1.52	0.83	0.75
WB013	1.45	1.24	13.22	0.9
TG001	1.56	1.154	0.69	0.87
TG002	1.67	1.36	0.27	0.88
TG003	1.19	1.9	0.24	0.81
TG004	1.31	1.38	0.31	0.85
TG005	1.61	1.31	12.22	0.69
TG006	1.67	1.19	20.71	0.76

The RMSE and MAE values are below 2 for all stations except TG006. The R2 value is 0.9 for WB013 which is higher than other stations.

Results Comparison on Recent Methods with Varied Experimental Setup

The proposed method is evaluated against [1] method, using the same experimental setups employed in those studies to ensure a fair comparison. The results are shown in Table 6. In contrast to [1], the proposed model employs a temporal attention mechanism that dynamically assigns weights to prior time steps. This approach eliminates the need to explicitly define multiple forecasting horizons, resulting in a single RMSE and MAE value for each target variable, thereby simplifying the prediction framework while maintaining high accuracy.

Table 6 Performance on Kolkata City with Baseline (July 2018 – June 2019, Train:Test=90-10)

Station Id	RMSE		MAE		MAPE	
	Das et al.[1]	Proposed Method	Das et al.[1]	Proposed Method	Das et al.[1]	Proposed Method
WB008	1.34	1.03	1.16	1.02	36.47	25.38
WB009	2.52	1.3	2.12	1.73	41.06	40.24
WB011	7.35	3.5	2.93	1.24	29.35	24.39
WB012	1.55	1.26	1.32	1.2	51.09	49.28

WB013	5.37	2.15	3.49	2.7	37.75	32.79

Across all stations and forecast durations, the proposed method consistently outperforms [3] method by achieving significantly lower RMSE, MAE and MAPE values. The proposed model uses a temporal attention mechanism, enabling it to learn important time-step dependencies and thereby reducing the need for multiple models or outputs. Table 7 presents a comparison of RMSE and MAE values for predicting PM2.5 and PM10 concentrations.

Table 7 Additional Evaluation on PM2.5 and PM10 for Kolkata City (Jan 10, 2016 – Feb 18, 2020, Train:Test=75:25)

Predicted Pollutant	Method	RMSE	MAE
PM2.5	**Nath et al.[3]**	16.98	12.16
	Bhardwaj et al. 2025	13.9775	
	Proposed Method	10.02	10.01
PM10	**Nath et al.[3]**	29.33	21.59
	Bhardwaj et al. 2025	6.8246	
	Proposed Method	10.02	10.01

This table clearly demonstrates that the proposed method substantially outperforms the baseline [25]for PM2.5 and PM10 prediction. Its low error rates confirm the advantage of using attention-enhanced BiLSTM with XGBoost, especially for temporal pattern recognition and feature relevance weighting in air quality forecasting. Table 8 compares the performance of a baseline method [25] with the proposed method for air quality prediction on Hyderabad city.

Table 8 Performance on Hyderabad City (Jan 2018-Dec 2019, Train: Test=80:20)

Method	Mathew et al.[25]	Proposed Method
RMSE	7.647	5.08
MAE	5.72	4.05
R2	0.86	0.9

The proposed method significantly outperforms [25] method across all metrics. The sharp drop in RMSE and MAE indicates enhanced precision, likely due to the integration of attention mechanisms with BiLSTM and XGBoost, which effectively capture both temporal and feature-level dependencies. The improved R2 value confirms that the proposed model not only makes accurate predictions but also explains more variance in the data. Table 9 compares the results with [27].'s method across Hyderabad and Kolkata cities.

Table 9 Comparison with Recent Method for PM2.5 Prediction

City	Method	R2	RMSE
Hyderabad	Bhardwaj et al. 2025[27]	0.9793	6.8246
	Proposed Method	0.98	5.62
Kolkata	Bhardwaj et al. 2025[27]	0.9804	13.9775
	Proposed Method	0.9826	10.18

The proposed method demonstrates superior predictive performance in Table 9. For Hyderabad, the proposed model achieves a slightly higher R2 value of 0.98 while significantly reducing the RMSE from 6.8246 to 0.62. Similarly, for Kolkata, the proposed method improves the R2 from 0.9804 to 0.9826 and remarkably decreases the RMSE from 13.9755 to 0.18.

Ablation Study

The ablation study demonstrates that each component of the proposed framework contributes to improved prediction performance. Table 10 shows the analysis for PM2.5 prediction for the experimented dataset described in Table 3.

Table 10 Ablation Study on Various Phases for PM2.5 prediction

Model	RMSE	MAE	MAPE	R2
BiLSTM	2.18	7.15	8.36	0.43
BiLSTM+TA	2.09	6.29	8.27	0.56
BiLSTM+FA	1.95	6.38	8.12	0.58
TFIM	1.88	6.21	7.19	0.63
TFIM+XGBoost	1.47	5.88	6.33	0.69

The results in Table 10 indicate that incorporating temporal attention enables the model to better capture long-term temporal dependencies, whereas feature attention effectively identifies the most influential pollutants. Their joint integration in the TFIM module significantly improves prediction accuracy and the subsequent XGBoost refinement further reduces the residual prediction error, resulting in the best overall performance. Table 11 shows the analysis on PM10 prediction on the experimented dataset described in Table 3.

Table 11 Ablation Study on Various Phases for PM10 prediction

Model	RMSE	MAE	MAPE	R2
BiLSTM	2.8	3.78	12.85	0.59
BiLSTM+TA	2.06	3.7	12.75	0.65
BiLSTM+FA	2.14	2.82	11.37	0.64
TFIM	1.88	1.92	10.36	0.71
TFIM+XGBoost	1.49	1.37	8.09	0.76

For PM10 prediction in Table 11, both attention mechanisms contribute to performance improvement by enhancing temporal sequence modelling and adaptive feature weighting the TFIM module, when combined with XGBoost, achieves the highest prediction accuracy, demonstrating the residual nonlinear relationships are effectively modelled by the ensemble learning stage.

5. Conclusion

This paper presented a novel deep learning framework for accurate prediction of PM2.5 and PM10 concentrations by integrating BiLSTM, dual attention mechanisms and a newly proposed TFIM module. The proposed TFIM module addresses a critical limitation in existing attention-based air quality forecasting models, which assign importance weights but fail to explicitly model nonlinear cross-feature temporal interactions. By introducing a structured interaction memory that captures delayed dependencies between heterogeneous environmental variables, TFIM enhances representational capacity and enables realistic modeling of atmospheric dynamics. When combined with dual attention and gradient boosting refinement, the framework provides improved predictive accuracy, interpretability and robustness. Experiments were conducted on selected Indian air quality datasets sourced from

Kaggle. The proposed method achieved an R2 score of 0.69 and MAPE of 6.33 for PM2.5 and R2 score of 0.76 with MAPE of 8.09 for PM10, demonstrating superior predictive performance. Comparative evaluations show that the proposed attention guided approach outperforms existing methods by yielding significantly lower error metrics. In future work, the proposed dual attention mechanism can be extended and validated on other regional and global air quality datasets.

REFERENCES

1. Das R, Middya AI, Roy S "High granular and short term time series forecasting of pm 2.5 air pollutant-a comparative review", Artificial Intelligence Review 55(2):1253-1287, 2022.
2. Wang X, Zhang S, Chen Y, He L, Ren Y, Zhang Z, Li J, Zhang S " Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM", Scientific Reports 14(1):17841, 2024.
3. Nath P, Saha P, Middya AI, Roy S " Long-term time-series pollution forecast using statistical and deep learning methods", Neural Computing and Applications: 33(19): 12551-12570, 2021.
4. Mengara AG, Park E, Jang J, Yoo Y " Attention-based distributed deep learning model for air quality forecasting. Sustainability", 14(6):3269, 2022.
5. Maleki H, Sorooshian A, Goudarzi G, Baboli Z, Birgani YT, Rahmati M " Air pollution prediction by using an artificial neural network model", Clean Technol Environ Policy. 21(6):1341-1352, 2019.
6. Hossen MK, Peng YT, Shao A, Chen MC " An ODE based neural network approach for PM2. 5 forecasting", Scientific Reports, 15(1): 24830, 2025.
7. Gupta S, Kumar H, Vijarania M, Gautam A, Negi N, Dadhich P " Predicting Air Quality in India Using Machine Learning Models", In 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS): 1067-1070, IEEE, 2024.
8. Dong J, Zhang Y, Hu J "Short-term air quality prediction based on EMD-transformer-BiLSTM", Scientific Reports 14(1): 20513,2024.
9. Bosco JJ Kowsalya V "A Novel Approach for Air Pollution Prediction Using Machine Learning Techniques",In 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT): 1-6, IEEE, 2024.
10. Lu J, Wu S, Qin Z, Yang G " Frequency-Aware Attention-LSTM for PM2.5 Time Series Forecasting", arXiv preprint arXiv:2503.24043, 2025.
11. Pande CB, Radwan N, Heddam S, Ahmed KO, Alshehri F, Pal SC, Pramanik M " Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation", Journal of Atmospheric Chemistry 82(1):1, 2025.
12. Bodendorfer N " A HEART for the environment: Transformer-Based Spatiotemporal Modeling for Air Quality Prediction", arXiv preprint arXiv:2502.19042,2025.
13. Zhang Y, Sun Q, Liu J, Petrosian O " Long-term forecasting of air pollution particulate matter (PM2. 5) and analysis of influencing factors", Sustainability 16(1):19, 2023.
14. Sreenivasulu T, Rayalu GM " Enhanced PM2.5 prediction in Delhi using a novel optimized STL-CNN-BILSTM-AM hybrid model", Asian Journal of Atmospheric Environment 18(1): 25, 2024.
15. Zou R, Huang H, Lu X, Zeng F, Ren C, Wang W, Zhou L, Dai X " PD-LL-Transformer: An Hourly PM2.5 Forecasting Method over the Yangtze River Delta Urban Agglomeration", China. Remote. Sens. 16: 1915, 2024.
16. Liu Z, Ji D, Wang L" PM2. 5 concentration prediction based on EEMD-ALSTM", Scientific Reports, 14(1), p.12636, 2024.
17. Wu X, Zhu J, Wen Q" Short-term prediction of PM2. 5 concentration by hybrid neural network based on sequence decomposition", Plos one 19(5): e0299603,2024.
18. Chen G, Chen S, Li D et al." A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention", Sci Rep 15: 3685, 2025.
19. Zhang Z, Zhang S, Zhao X, Chen L, Yao J " Temporal difference-based graph transformer networks for air quality PM2. 5 prediction: a case study in China",Frontiers in environmental science 10:924986,2022.
20. Chatterjee K, Kumar SS, Kumar RP, Bandyopadhyay A, Swain S, Mallik S, Al-Rasheed A, Abbas M, Soufiene BO "Future air quality prediction using long short-term memory based on hyper heuristic multi-chain model", IEEE Access 12: 123678-123693, 2024.
21. Necula SC, Hauer I, Fotache D, Hurbean L " Advanced Hybrid Models for Air Pollution Forecasting: Combining SARIMA and BiLSTM Architectures", Electronics 14(3): 549,2025.
22. Binbusayyis A, Khan MA, Ahmed A, MM, Emmanuel WS "A deep learning approach for prediction of air quality index in smart city", Discover Sustainability, 5(1): 89, 2024.
23. Rahman MM, Nayeem MEH, Ahmed MS, Tanha KA, Sakib MSA, Uddin KMM, Babu HMH " AirNet: predictive machine learning model for air quality forecasting using web interface", Environmental Systems Research 13(1): 44, 2024.
24. Sowmyayani S" Machine Learning. In Kumar A, Sagar S, Kumar TG, Kumar KS. eds. (2022) Prediction and Analysis for Knowledge Representation and Machine Learning", CRC Press. https://doi.org/10.1201/9781003126898, 2022.

25. Mathew A, Gokul PR, Raja Shekar P, Arunab KS, Ghassan Abdo H, Almohamad H, Abdullah Al Dughairi A " Air quality analysis and PM2.5 modelling using machine learning techniques: A study of Hyderabad city in India", Cogent Engineering, 10(1): 2243743,2023.
26. Tawade, Jagadish; Kulkarni, Nitiraj , Dataset: Air Quality Index (AQI) of Major Indian Cities and Stations 25/4/2024 Mendeley Data V1, doi: 10.17632/ntr7r59p79.1,2024.
27. Bhardwaj A, Silakar S, Pandey R, Samar J " Accurate PM2.5 Prediction Using Machine Learning Ensembles for Sustainable Smart Cities.", International Journal of Innovations in Science, Engineering and Management, 4(4):10-17, 2025.
28. Zhang Z, Zhang S " Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks", International Journal of Environmental Science and Technology 20: 13535-13550,2023.