Feature-Gated Spatio-Temporal Deep Learning With City Embedding And Attention-Based Bilstm For Air Quality Index Prediction

R. Vetri Selvi[1], Dr. R. Satish Babu[2]

[1] Research Scholar, Department of Computer & Information Science, Faculty of Science, Annamalai University, Annamalai Nagar-608002
[2] Assistant Professor, Department of Computer & Information Science, Faculty of Science, Annamalai University, Annamalai Nagar-608002.
Corresponding Author: R. Vetri Selvi

Abstract
Air pollution is one of the biggest environmental and public health challenges faced by rapidly urbanizing areas, so accurate air quality prediction is crucial for environmental monitoring and policy making. This study employs a novel deep learning framework namely, Feature-Gated City-Embedded BiLSTM with Self-Attention (FGCE-BiLSTM) to predict Air Quality Index (AQI) robustly across multiple metropolitan cities. This model incorporates spatio-temporal feature learning and city-specific contextual representation to achieve improved regression-based AQI prediction and categorical air quality classification. The dataset used for this study is the 219,125 hourly observations collected from five great cities of India (Delhi, Mumbai, Chennai, Kolkata and Bangalore) featured on Kaggle platform. An exhaustive preprocessing pipeline is applied, which encompasses temporal feature extraction, missing value imputation, pollutant ratio construction, lagged-based generation of temporally-dependent features and rolling statistics for all pollutants. 67 engineered features are created in first place, and subsequently filtered against the correlation where a pair of variables would add mutual informative until an optimized feature set emerges from the cleaning process (hovering around 45 at best).
Sequential samples are generated from each city to capture temporal dependencies, and these are subsequently conditioned on city encodings in order to model spatial heterogeneity across urban environments. The feature-gating mechanism dynamically adjusts the importance of the features before temporal modeling based on stacked Bidirectional Long Short-Term Memory (BiLSTM) layers. Then a self-attention mechanism can be utilized to emphasize the most impactful temporal patterns. Based on two-output strategy, the framework performs AQI regression and air quality category classification simultaneously. Because of large class imbalance the Good and Satisfactory classes were ignored, and four representative classes (Moderate, Poor, Severe and Very Poor) were used to train the model. The predictive ability was shown experimentally both in regression (R^2, MAE, RMSE) and classification tasks (Accuracy, Precision, Recall and F1-score), as well as confusion matrix and ROC curves. Overall, the findings demonstrate that with spatio-temporal modeling capability, the proposed FGCE-BiLSTM model can learn complex pollution patterns and greatly enhances city-wide prediction reliability. This framework presents a scalable and interpretable method for intelligent air quality monitoring, facilitating data-driven environmental decision-making in urban ecosystems.

Keywords - Air Quality Index Prediction, Spatio-Temporal Deep Learning, Bidirectional Long Short-Term Memory (BiLSTM), Attention Mechanism, City Embedding, Feature Engineering, Multi-Task Learning, Urban Air Pollution Forecasting.

1. Introduction

Air pollution is one of the emerging threats facing global environment and persistence in many cities, rapid industrialization urbanization and traffic have contributed to deterioration of air quality for some decades. Environmental exposure to elevated levels of atmospheric pollutants such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen oxides (NO_x), sulfur dioxide (SO_2) carbon monoxide (CO) and ozone (O_3)) were found to have strong association with respiratory diseases, cardiovascular disorders as well as premature mortality. Air quality index (AQI) is a key standardized indicator to reflect the

degree of pollution and its potential health impact, according to environmental monitoring authorities. Thereby, accurate prediction of AQI is an essential foundation for early warning system, environmental policy guidance, and intelligent decision support in urban pollution management. Conventional air quality prediction methods have mainly depended on statistical techniques, including autoregressive integrated moving average (ARIMA), regression models, and rule-based forecasting systems. While these methods offer capability for building simple temporal models, they often fail with nonlinear relationships and complex interactions between multiple sources of pollutants and environmental factors. Air quality dynamics involve various factors, including the emission of pollutants, meteorological conditions, seasonal trends and spatial interaction among urban agglomerations. Thus, traditional statistical models often show poor predictive power when applied to large-scale, high-dimensional environmental datasets. Machine learning and deep learning based algorithms have been developed as a powerful approach to time-series forecasting of air quality prediction with the development of Artificial Intelligence. Various machine learning models, including Random Forest, Support Vector Machines (SVM), Gradient Boosting methods, and ensemble learning techniques have been used for AQI forecasting tasks. These models can learn complex nonlinear relationships between the environmental variables and pollutant concentrations. For instance, Ravindiran et al. proposed multiple machine learning models used to predict AQI in an Indian coastal city and showed that prediction accuracy can be increased significantly with ensemble learning techniques as compared to traditional approaches [1]. Similarly, Liang et al. explored various machine learning models for the air quality prediction and found that data-driven models can ensure better generalization performance when trained on a large scale environmental dataset [2]. The advent of deep learning techniques has significantly improved the operational capability of air quality forecasting systems, demonstrating good performance modelling temporal dependencies in sequential environmental data. Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), have also been widely used due to their capacity to model long-lasting temporal correlations in time-series data. Zhang et al., implemented a CNN–LSTM hybrid model for predicting AQI in Beijing, using convolutional layers to capture spatial features and learning temporal dependencies through LSTM networks with improved performance over classical machine-learning models [3]. Similarly, Duan et al. proposed an ARIMA–CNN–LSTM deep learning hybrid framework that combines statistical modeling and deep machine learning methods to further improve the air quality prediction accuracy due to its ability of modeling both linear and nonlinear temporal dynamics [4].

Here is a comparison of the appropriate ways to cite five different articles that are most relevant to some recent research aimed at mastering more complex hybrid and deep learning frameworks with multi-view environmental datasets. Xu et al. and developed air quality prediction models based on machine learning which showed that advanced data-driven models are able to successfully capture the pollution dynamics in complex urban environments [5]. In addition, Alrashidi et al. which developed $PM_{2.5}$ specific machine learning models prediction and emphasized that underlying forecasting reliability through the importance of feature engineering and optimizing models [6]. They have also been used to improve predictive performance by exploring deep learning architectures with temporal attention mechanisms and feature learning. Lei et al. explored deep learning methods for air quality forecasting and highlighted the promise of neural network architectures in modeling spatio-temporal environmental data [7]. A new direction in air quality prediction research focus on integration advanced deep-learning models with time-series decomposition and optimization strategies. For example, Ma et al. developed a deep learning method that integrates time series decomposition with

intelligent optimization algorithms (such as genetic and particle swarm) to enhance AQI prediction accuracy by discovering hidden temporal structures in environmental datasets [8]. Hybridization of multiple modeling strategies is capable of enhancing prediction performance as evidenced in these studies showing that hybrid deep learning models possess good potential.

While traditional air quality prediction methods have made great strides, there are still a number of challenges. First of all, most existing studies use single-station or single-city datasets which inhibit the models from receiving spatial heterogeneity across urban regions. Cities have widely divergent air pollution patterns due to different industrial activity, traffic volume, geophysical characteristics and meteorological conditions. Second, traditional deep learning models may treat all input features with the same weight, which can lead to less informative or redundant variables that negatively affect model performance. Hence, effective feature selection and adaptive feature weighting mechanisms are critical for enhancing predictive accuracy. Third, existing approaches either are solely regression-based for predicting the values of pollutants without any analysis based on AQI classification or operate in an isolated way such as pollutant counting methods which intent to improve some attributes (i.e. location) at the cost of ignoring real-world air quality monitoring needs (e.g. estimates).

In order to tackle these issues, this work present an innovative Feature-Gated City-Embedded BiLSTM with Self-Attention (FGCE-BiLSTM) method for accurate forecasting of air pollution data. The combined feature engineering, feature gating mechanisms and city embeddings proposed approach provides smaller weaknesses in spatial heterogeneity and temporal dependencies as shown by the experimental results on urban air quality data. A BiLSTM stacked architecture is used to learn both forward and backward sequence relationships of pollutants followed by a self-attention mechanism to identify key temporal information influencing AQI estimation. Furthermore, the proposed framework devises a dual-output learning approach that jointly carries out AQI regression and air quality classification to allow our model to assist both numerical forecasting and categorical pollution evaluation. The presented framework is examined using a comprehensive air pollutants dataset with hourly measurements from several metropolitan regions. Nor these long feature preprocessing and engineering pipelines that extract what might not be very relevant environmental features to the atomic interactions, nor the model selection methods applied on top of them that reduce redundancy and increase performance. Experimental results show that the proposed FGCE-BiLSTM framework successfully captures complex spatio-temporal pollution patterns while greatly enhancing prediction reliability across various urban areas.

This study brings forth several key contributions, which can be described as follows:

- A feature-gated spatio-temporal deep learning approach is proposed in which the fine-tuned importance of the environmental features prior to temporal modeling.
- A city embedding mechanism to capture spatial heterogeneity of urban regions is introduced.
- The model uses a BiLSTM network with self-attention to capture long-term temporal dependencies in the air pollution data.
- The framework proposes a dual-output learning strategy to jointly perform AQI regression and air quality classification.

The rest of this paper is structured as follows. Section 2 discusses related work on air quality prediction using machine learning and deep learning techniques. The proposed FGCE-BiLSTM framework and its architecture are introduced in Section 3. Section 4 describes the experimental setup, details of datasets, evaluation metrics, experimental

results and performance analysis. Lastly, Section 5 summarizes the study and describes future research directions.

2. Related work

As the effects of atmospheric pollution on human health and environmental sustainability become more prominent, air quality prediction has turned out to be a key research topic. Recently, due to the increasing availability of large scale environmental monitoring data, researchers have investigated a number of machine learning and deep learning methods to predict pollutant concentrations as well as Air Quality Index (AQI). The most effective approaches for air pollution prediction involve hybrid deep learning models, attention-based neural networks, transformer architectures, and graph-based spatio-temporal models that capture the intricate temporal and spatial dependencies present in the data.

2.1 Machine Learning-Based Air Pollution Prediction

In more recent years, there have been several studies exploring the use of machine learning techniques to estimate air pollution levels and improve the capabilities for monitoring. Samad et al. suggested a machine learning-based framework to establish virtual monitoring stations, enabling pollutant levels to be predicted in unmonitored regions [9]. Their approach showed that data-driven models can successfully approximate pollution levels based on the previous environment's historical and meteorological data. That highlights the power of predictive models in combination with existing infrastructure, such as air monitoring.

2.2 CNN–LSTM-Based Air Quality Prediction Models

Combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks, hybrid deep learning models have been widely used to predict air quality as they can catch spatial correlations and temporal dependencies of environmental datasets. Wang et al. proposed a CNN-ILSTM model for prediction of AQI values, in which convolutional feature extraction is integrated with improved LSTM networks to improve the performance compared to traditional machine learning models [10].

Recent research has improved these by including optimization techniques with hybrid feature learning approaches. Chang et al. proposed a CNN-LSTM model tuned with pelican optimization method, which led to increased prediction precise by optimizing network parameters [11]. Similarly, Hu et al. due to the limitation of traditional method architecture [12]. Therefore, YG et al. proposed a hybrid framework composed of CNN sequence LSTM-KAN, which takes geographical information into account during training to improve the spatial variance capture in different areas. Mohapatra et al. They also utilized a CNN-LSTM based architecture to forecast AQI of the region using multivariate pollution data and showed that CNN can be effective for time series air pollutions forecasting tasks [13]. In conclusion, these studies demonstrate that the combined CNN-LSTM model can learn complex spatio-temporal relationships from environmental data.

2.3 Deep Learning and BiLSTM-Based Prediction Models

Air quality forecasting using Bidirectional Long Short-Term Memory (BiLSTM) networks Despite the wide usage of LSTMs in air quality prediction, BiLSTM networks also found significant application due to their ability to capture both past and future temporal dependencies. Janarthanan et al. developed an advanced deep learning framework for AQI prediction based on different types of neural networks and compared their performance over a dataset dedicated to urban air pollution monitoring systems [14].

Other approaches considered attention-based deep learning frameworks, to enhance feature extraction in air pollution datasets. Wu et al. proposed ADNNet, an attention-

based deep neural network model that automatically evaluate the significant features for AQI prediction, consequently improving both interpretability and predictive capability of the model [15]. Similarly, Wang et al. proposed a CNN-BiNLSTM-Attention model which fuses convolutional feature extraction, bidirectional temporal modeling and attention mechanisms to outperform AQI prediction [16]. By using attention capacities on BiLSTM architectures based approaches it provides an explanation behind capturing complex temporal relationships with respect to air quality data.

2.4 Transformer-Based Air Quality Prediction Models

Recently, transformer architectures have been utilized for environmental time-series forecasting due to their capability of modeling long-range dependencies using self-attention mechanisms. Zhang et al. the graph transformer network that is applied in PM2 5 predicts by combining temporal difference learning with attention mechanism-based feature extraction to learn dynamic pollution evolution features across monitoring stations [17].

In addition, Liu et al. proposed a Transformer-BiLSTM hybrid model for AQI prediction that integrates transformer-based feature extraction with bidirectional temporal modeling in order to improve forecasting accuracy [18]. These techniques show that transformer architectures can learn long-term dependencies in air pollution time series efficiently while ensuring computational efficiency.

2.5 Graph Neural Network-Based Spatio-Temporal Prediction Models

Due to pollutant transport among locations, air pollution patterns often show strong spatial dependencies. In order to remedy this, much work has gone into the use of graph neural network (GNN) architectures for modeling spatial relationships between monitoring stations.

Ye et al. proposed a hybrid model based on both GNN and transformer architectures to capture PM2.5 in space and time directions prediction tasks [19]. Similarly, Pan et al. proposed a Graph Attention Recurrent Neural Network (GARNN) model that uses graph attention mechanisms to characterize spatial interactions of monitoring stations while simultaneously learning temporal pollution patterns [20].

Graph convolution-based methods are also widely applied in spatio-temporal prediction tasks. Li et al. proposed the Diffusion Convolutional Recurrent Neural Network (DCRNN) model to capture dynamic spatial dependencies in time-series data based on diffusion-based graph convolution operations [21]. Huang et al. leveraged graph convolutional networks based on transformer architectures to enhance the PM2. 5 concentration prediction accuracy [22]. Similarly, Cai et al. The authors of [23] proposed a spatio-temporal graph feature fusion model which integrates graph learning with temporal feature extraction to improve the accuracy of air quality forecasting. These models highlight the necessity of integrating spatial relationships into environmental prediction frameworks.

2.6 Research Gap and Motivation

While many studies have improved the air quality prediction using machine learning and deep learning models, there are still several challenges to address. First of all, almost all CNN-LSTM and BiLSTM-based models concentrate on only temporal features extraction instead of dynamically tweaking feature importance, potentially including redundant or less informative variables in training. Second, many transformer and graph-based models need complex architectures and have high computational requirements that make them difficult to be applied in practical settings for real-time environmental monitoring systems. Finally, most of the existing approaches focus on single-task regression models

to forecasting pollutant levels, while in practical air quality monitoring systems, numerical AQI prediction and categorical air quality classification should be combined to enable public health advisories. Furthermore, when it comes to spatial heterogeneity of multi-city air pollution datasets, most studies do not directly use city-specific attribute data which is fundamentally important for the task.

2.7 Positioning of the Proposed FGCE-BiLSTM Framework

To overcome the above challenges, this research work proposed a Feature Gated City Embedded Bidirectional Long Short Term Memory with Self Attention (FGCE-BiLSTM) framework for air quality prediction. The proposed framework contains several novel components that distinguish it from previous works. First, to reduce redundancy and improve the efficiency of temporal model, a feature gating mechanism is utilized before temporal model to dynamically weight environmental features. Second, a city embedding representation is proposed to represent the spatial heterogeneity between different urban settings while avoiding the demanding complexities of graph structures. Thirdly, the framework employs a stacked BiLSTM architecture with self-attention, allowing the model to learn bidirectional temporal dependencies and identify which temporal patterns were most influential in pollutant sequences. The framework adopts a dual-output learning strategy to achieve AQI regression and air quality classification during the same training process, enabling the framework to deliver both quantitative predictions and categorical pollution evaluations.

The FGCE-BiLSTM framework proposed in this paper overcome the major limitations in existing deep learning approaches for multi-city air quality prediction by using feature gating and a city embedding combined together with attention-based temporal modeling into an end to end structure.

3. Proposed Methodology

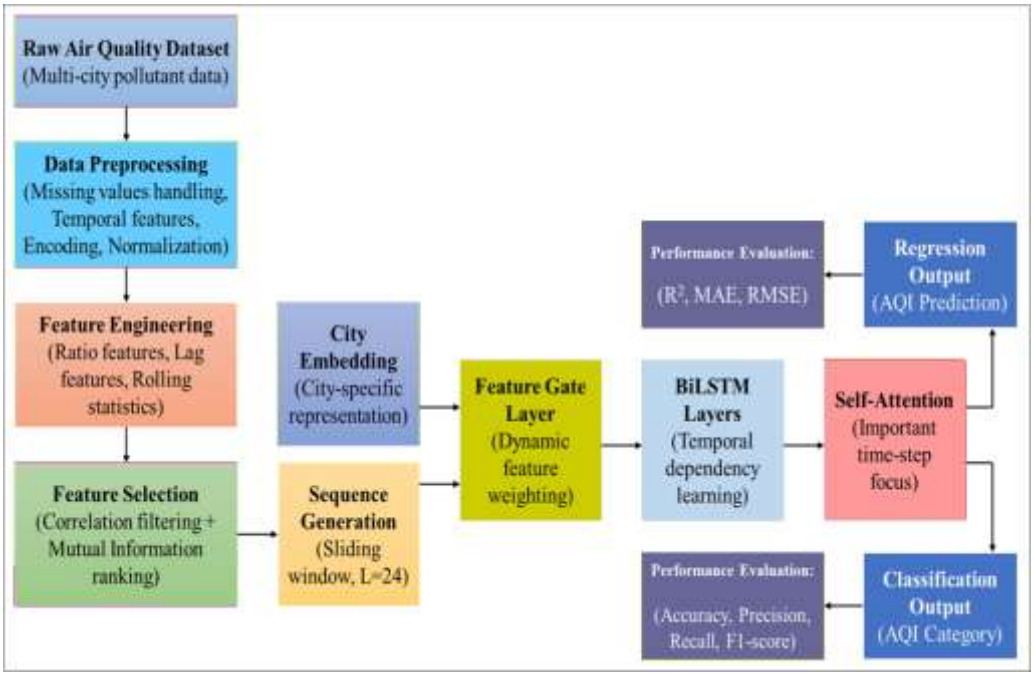
The FGCE-BiLSTM framework for AQI prediction is based on a spatio-temporal deep learning pipeline. First of all, the original multi-city air quality dataset is processed by imputation for missing values, extracting temporal features and normalization preprocessing. Feature engineering is then processed to generate pollutant ratios, lag features and statistical attributes, and the features are selected based on correlation filtering and mutual information. The processed data are transformed into temporal sequences for each city, with the introduction of a city embedding module to capture characteristics of pollution specific to that city. Feature gating mechanism dynamically focuses on aspect-wise features before temporal modeling, and subsequent sequences with refined important aspects are sent through a Bidirectional Long Short-Term Memory (BiLSTM) network that incorporates self-attention to learn significant temporal dependencies. Finally, a dual-output architecture is proposed to achieve AQI regression and air quality classification concurrently, and the performance of the model is assessed in terms of metrics including R^2, MAE, RMSE as well as accuracy, precision, recall, and F1-score. Table 1 shows the full architectural details of the proposed approach, and Figure 1 depicts the overall workflow.

Table 1: Layer-wise Architecture of the Proposed Model

Layer	Output Shape	Description
Input (Sequential Features)	24 × 45	Preprocessed air quality features arranged as temporal

		sequences (sequence length = 24)
City Input	1	Encoded city index representing different urban regions
Feature Gate Layer	24 × 45	Sigmoid-based gating mechanism dynamically adjusts feature importance
City Embedding	8	Embedding layer generates city-specific contextual representation
Concatenation Layer	24 × (45 + 8)	City embeddings concatenated with sequential environmental features
BiLSTM Layer 1	24 × 96	Bidirectional LSTM captures forward and backward temporal dependencies
BiLSTM Layer 2	24 × 48	Second BiLSTM layer refines temporal feature representation
Self-Attention Layer	48	Attention mechanism identifies the most influential time steps
Dense Layer	24	Fully connected layer with ReLU activation
Regression Output	1	Predicts continuous AQI value
Classification Output	4	Softmax classifier predicts AQI category (Moderate, Poor, Severe, Very Poor)

Figure 1. Block diagram of the proposed model



3.1 Data Preprocessing

Preprocessing environmental data is a crucial step to enhance the reliability and predictive ability of deep learning models. Air quality datasets are frequently plagued by missing data, temporal non-conformance, and heterogeneous scales of features due to the combing readings from various monitoring stations and environmental sensors. Hence, a structured preprocessing pipeline is developed to maintain the consistency of data across different time periods and to standardize features before training models.

3.1.1 Datetime Conversion and Temporal Feature Extraction

A mock attribute for this dataset is a timestamp indicating the time of measuring pollutant concentrations. In order to perform temporal pattern analysis, the timestamp is decomposed into structured temporal attributes hour, day and month which characterize periodic pollution trends about daily activities, seasonal variations and climatic conditions.

Let t_i represent the timestamp of the i^{th} observation. Temporal features are extracted as:

$$Hour_i = hour(t_i) \quad (1)$$

$$Day_i = day(t_i) \quad (2)$$

$$Month_i = month(t_i) \quad (3)$$

To account for the seasonal effects in air pollution patterns, a seasonal indicator variable is also introduced:

$$Season_i = \begin{cases} 0 & if\ Month \in \{12,1,2\}\ (Winter) \\ 1 & if\ Month \in \{3,4,5\}\ (Summer) \\ 2 & if\ Month \in \{6,7,8,9\}\ (Monsoon) \\ 3 & otherwise \quad (Post-monsoon) \end{cases} \quad (4)$$

These temporal features help the model learn repetitive pollution phenomena caused by typical seasonal events and within-day environmental changes.

3.1.2 Missing Value Imputation

Air quality datasets often have missing pollutants for various reasons, such as malfunctions of the sensors, observation transmissions, and maintenance operations. To keep temporal continuity in each city dataset, forward and backward propagation methods fill in missing pollutant values. Define x_t as the pollutant measurement at time t. In case if it is missing, forward filling is performed as follows:

$$x_t = x_{t-1} \quad (5)$$

If forward propagation cannot be done, backward filling is performed:

$$x_t = x_{t+1} \quad (6)$$

This method guarantees that a few missing measurements does not split the temporal ordering of the dataset.

3.1.3 Categorical Encoding

The dataset contains both categorical values such as city names and AQI categories. All these characteristics need to be transformed to numbers in order for the model to train.

City labels are encoded using integer mapping:

$$City_i = LabelEncoder(City_i) \quad (7)$$

Similarly, AQI bucket labels are transformed into numerical classes:

$$AQI_Bucket_i = LabelEncoder(AQI_Bucket_i) \quad (8)$$

For efficient computation, the model uses One hot encoding which keeps categorical references.

3.1.4 Feature Normalization

Environmental variables tend to have different numerical ranges. For example, concentrations in particulate matter may be tens of hundreds, but gas concentrations slower. When being different in scale they can prevent the model from converging while training. Hence, they are standardized using z-score normalization.

Given a feature x, the normalized value z is computed as:

$$z = \frac{x-\mu}{\sigma} \quad (9)$$

Where, μ represents the mean of the feature, and σ represents the standard deviation. Normalizing features to have zero mean and unit variance, this way prevents troublesome value extremes during training, facilitating more stable gradients.

3.1.5 Removal of Incomplete Records

This can cause missing values at the beginning of the time series due to temporal feature generation and lag-based operations. Any incomplete records are dropped in order to keep the input sequences equal for our deep learning model.

Let D denote the dataset after preprocessing. The cleaned dataset is obtained as:

$$D_{clearn} = \{x_i \in D \mid x_i \neq NaN\} \quad (10)$$

This step ensures that all observations used for training contain valid feature values.

3.2 Feature Engineering and Feature Selection

Transforming raw environmental measurements into more informative representations through feature engineering greatly enhances the predictive capacity of deep learning models. Air pollution forecasting relies on complicated relationships between pollutant levels, time of day, and external factors. So a feature engineering method is used by which it creates meaningful variables for correlating pollutants, temporal dependencies between pollutants and the historical context of pollution. Following feature generation, a process of feature selection is used to identify the most relevant attributes, reducing the redundancy and resulting model complexity.

3.2.1 Pollutant Ratio Features

The ratios of the concentrations of different atmospheric pollutants are useful for understanding the relationships amongst varied pollutants and their sources. These ratios allow the model to account for interactions between pollutants that could affect overall air quality.

Two ratio-based features are generated in this study:

$$PM_ratio = \frac{PM_{2.5}}{PM_{10}+\epsilon} \quad (11)$$

$$NO_ratio = \frac{NO_2}{NO_x+\epsilon} \quad (12)$$

Where, $PM_{2.5}$ and PM_{10} represent particulate matter concentrations, NO_2 and NO_x denote nitrogen dioxide and nitrogen oxide concentrations, and ϵ is a small constant introduced to avoid division by zero.

Relative contributions of each pollutant are captured, along with various additional context-specific features that aid the prediction model.

3.2.2 Temporal Lag Features

Air pollution has significant temporal dependencies, as the concentration of pollutants at a specific moment is highly dependent on previous ones. Lag based features are created on AQI and pollutant concentrations in order to capture these temporal dependencies.

Let x_t represent the pollutant value at time t. A lag feature for a lag interval k is defined as:

$$x_t^{lag(k)} = x_{t-k} \quad (13)$$

Where $k \in \{1, 3, 6, 12, 24\}$ represents the lag interval in hours.
Similarly, AQI lag features are generated as:

$$AQI_t^{lag(k)} = AQI_{t-k} \quad (13)$$

The inclusion of time lagged features allows the model to learn how pollution at previous times affect current levels, accounting for both short duration and long duration dependencies in pollution level patterns.

3.2.3 Rolling Statistical Features

Rolling statistical features over fixed time windows are computed to capture the temporal trend and local variation of air pollution. Features used rolling mean, which smooths out short-term fluctuations and expounds on the long-term trends of pollutant concentrations over time.

The rolling mean of AQI over a window size w is defined as:

$$AQI_{roll(w)} = \frac{1}{w}\sum_{i=0}^{w-1} AQI_{t-i} \quad (14)$$

Where $w \in \{3, 6, 12, 24\}$.
These rolling features enable the model to capture temporal patterns corresponding to pollution build up and diffusion over subsequent time frames.

3.2.4 Correlation-Based Feature Filtering

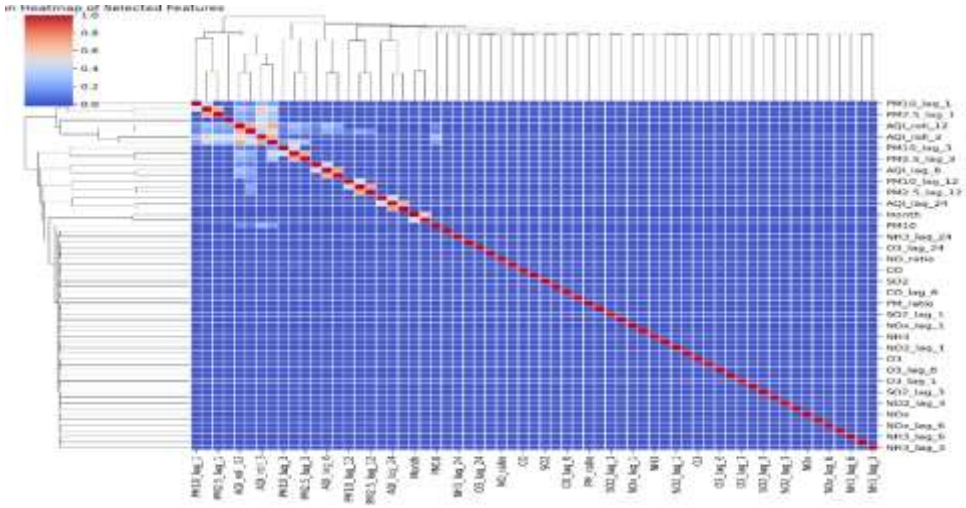
The dataset now has a lot of derived variables after feature generation. These features can be highly correlated among them, resulting in redundancy and undermining the learning of the model. Hence, features with excessive correlation are removed using correlation-based feature filtering.
The Pearson correlation coefficient between two features X and Y is computed as:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (15)$$

Where $\bar{X}$ and $\bar{Y}$ represent the mean values of the respective variables.
If two features are perfectly ($|r| > 0.95$ correlated, just one of them is dropped from the merged dataset. This procedure minimizes redundant features while enhancing computational efficiency.

The clustered correlation heatmap of selected features is provided in figure 2, showing the pairwise Pearson correlation coefficients between pollutants variables, lag features, rolling statistics and ratio attributes. Scale of color for the strength of correlation; red strong positive blue weak Features with similar correlation patterns are grouped together in hierarchical clustering, which assists in identifying highly correlated variables. In this sense, a visualization which assists the correlation-based filtering process is necessary to identify and remove redundant features higher than the pre-defined threshold of correlation while making sure enough informative predictors remain for future modelling.

Figure 2. Clustered correlation heatmap of the selected features



3.2.5 Mutual Information-Based Feature Ranking

To discover the most informative features that indicate AQI, mutual information (MI) can be employed to measure the dependency between each of the feature and target variable. Mutual information measures the amount of information that two variables share.

The mutual information between feature X and target variable Y is defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\ log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (16)$$

Where $p(x,y)$ represents the joint probability distribution of X and Y, $p(x)$ and $p(y)$ represent their marginal distributions.

Features yielding high mutual information scores contain more predictive information about the AQI and are thus ranked higher. For each model set, the mutual information scores of features are calculated and top 45 are selected to form optimized feature sets.

3.2.6 Preservation of Critical Lag Features

While feature ranking can detect the most informative variables, some lag features based on historical AQI trends are critical for temporal modelling. As a result, even if they do not fall among the most important features overall, key lag-based characteristics are retained. The preserved lag features include AQI_{lag_1}, AQI_{lag_3}, AQI_{lag_6}, $AQI_{lag_{12}}$, $and\ AQI_{lag_{24}}$. This also preserves the critical historical pollution signals which are important to have for the temporal learning model.

3.2.7 Final Feature Set

The results lead to around 45 informative predictors with pollutant correlations, temporal dependencies and environmental patterns after applying correlation filtering, mutual information ranking, lag feature preservation. This streamlined feature set removes redundancy but retains essential predictive information, allowing the proposed spatio-temporal deep learning model to be trained with optimal efficiency.

3.3 Sequence Generation and City Embedding

Air quality prediction is inherently a spatio-temporal forecasting task, in which concentrations of pollutants are influenced by both local environmental conditions and historical patterns of hourly pollution for specific regions. In order to model these dependencies, the processed dataset is reshaped into temporal sequences and we add

city-specific contextual information through a city embedding mechanism. By this means, the proposed model could capture temporal dynamics and spatial heterogeneity over multiple urban environments.

3.3.1 Temporal Sequence Generation

Most deep learning models requiring sequential input data are able to learn temporal dependencies, e.g., Long Short-Term Memory (LSTM) networks [23]. Hence, the preprocessed dataset is transformed into fixed-length temporal sequences using a sliding window method.

Let X_t denote the feature vector at time step t, where

$$X_t = x_{1,t},\ x_{2,t}, x_{3,t}, \dots, x_{m,t} \quad (17)$$

and m represents the number of selected features after feature engineering and selection. To capture historical pollution dynamics, a temporal sequence of length L is constructed as:

$$S_t = x_{t-L+1},\ x_{t-L+2}, \dots, x_t \quad (18)$$

Where L denotes the sequence length. 24 time steps were used in this study to illustrate hourly pollution trends over a one-day period.

The corresponding target variable is defined as:

$$y_t = AQI_t \quad (19)$$

Hence, for any training example, the model have a sequence-target pair: (S_t, y_t). This sliding window procedure is applied repeatedly through the dataset to produce a collection of contiguous training examples:

$$D = \{(S_1, y_1), (S_2, y_2), \dots\ (S_n, y_n)\} \quad (20)$$

Where n represents the total number of generated sequences. It enables the deep learning model to learn temporal correlation and pollution diffusion features between multiple time periods.

3.3.2 Multi-City Data Representation

Data has been gathered from various cities such as Delhi, Mumbai, Chennai, Kolkata and Bangalore. Because urban settings vary widely in factors such as industrial activity, population density, traffic patterns and geographic conditions affecting pollution dynamics, additional local knowledge should be included into the learning process. First, each city is represented as a categorical variable and encoded into integer index:

$$City_i \in \{0, 1, 2, \dots, C-1\} \quad (21)$$

Where C denotes the number of cities.

However, simply doing integer encoding does not reflect the fundamental resemblance of cities. Thus, a city embedding mechanism is proposed in order to capture dense vector representations of city-specific features.

3.3.3 City Embedding Representation

City embedding map holds the city index, and uses it to encode into a continuous low-dimensional vector that allows learning of spatial relationships between cities.

Let c_i denote the city index of the i^{th} sequence. The embedding operation is defined as:

$$e_i = E(c_i) \quad (22)$$

Where $E \in \mathbb{R}^{C \times d}$ is the embedding matrix, C is the number of cities, and d represents the embedding dimension.

Thus, each city is represented by a vector $e_i \in \mathbb{R}^d$. In this study, the embedding dimension is set to, $d = 8$. By packing cities into this compact representation, the model can learn latent spatial features corresponding to various city types.

3.3.4 Integration of City Embedding with Temporal Features

City embedding vector is fused with each time step in the temporal sequence for learning spatial information in the temporal process.

Let the temporal sequence be, $S_t \in \mathbb{R}^{L \times m}$ and the city embedding vector be, $e_i \in \mathbb{R}^d$. The embedding vector is repeated along the temporal window and concatenated with the feature sequence as follows:

$$S_t' = [S_t \,||\, e_i] \tag{23}$$

Where || denotes the concatenation operation. The resulting sequence becomes:

$$S_t' \in \mathbb{R}^{L \times (m+d)} \tag{24}$$

This representation aims to help the model learn temporal pollution pattern as well as spatial characteristic per city together.

3.4 Feature-Gated BiLSTM with Self-Attention

A Feature-Gated City-Embedded Bidirectional Long Short-Term Memory with Self-Attention (FGCE-BiLSTM) framework is proposed for the modelling of complex spatio-temporal dependencies in air pollution time-series data. In the proposed model, a feature gating mechanism is incorporated for feature selection/dimensionality reduction, while representation learning in the set of scenes can be achieved by a city embedding matrix to capture different AQI levels across cities and attention mechanisms are used for informative temporal patterns learning. The structure proposed in the mentioned architecture allows for a dynamic selection of feature importance, an effective representation of air quality in regard to both left and right side temporal dependability, and attention towards the time steps that were most crucial that made up air quality.

3.4.1 Feature Gating Mechanism

Environmental datasets tend to contain features with varying degrees of importance. To allow the contribution of each input feature to be dynamically regulated, a feature gating layer is added prior to the temporal modelling stage. The gating mechanism thus learns the feature-wise weights that determine whether to allow the input information flow or not into the network.

Let the input feature vector at time t be represented as $X_t \in \mathbb{R}^m$, where m denotes the number of selected features. The gating weights are calculated as follows where the activation function is sigmoid:

$$G_t = \sigma(W_g X_t + b_g) \tag{25}$$

Where W_g represents the learnable weight matrix, b_g denotes the bias vector, and $\sigma(\cdot)$ is the sigmoid activation function. The gated feature representation is then obtained as:

$$\tilde{X}_t = G_t \odot X_t \tag{26}$$

Where $\odot$ denotes element-wise multiplication. This step enables the model to focus on informative elements and suppress irrelevant features.

3.4.2 Integration of City Embedding

This is achieved by joining a gated feature representation with a city embedding vector to embed spatial contextual information. Let c_i be the city index for the i^{th} sample. The embedding layer translates the city index to a dense vector representation, $e_i = E(c_i)$, where E represents the embedding matrix, and c is the number of cities.

To create the final input representation, the embedding vector is concatenated with the gated feature sequence:

$$Z_t = [\tilde{X}_t \,||\, e_i] \quad (27)$$

Where || denotes the concatenation operator.

3.4.3 Bidirectional LSTM Temporal Modeling

The proposed framework uses Bidirectional Long Short-Term Memory (BiLSTM) networks to learn temporal dependencies inherent in the dynamics of pollution. The key difference as compared to normal LSTM models, which process the sequences in one direction only, BiLSTM models learn the temporal relations in both directions (forward and backward).

For a given sequence $Z = (Z_1, Z_2, \dots, Z_L)$, the forward and backward hidden states are computed as:

$$\overrightarrow{h_t} = LSTM_f(Z_t, \overrightarrow{h_{t-1}}) \quad (28)$$

$$\overleftarrow{h_t} = LSTM_b(Z_t, \overleftarrow{h_{t+1}}) \quad (29)$$

The model obtains the final hidden representation by concatenating forward and backward states:

$$h_t = \left[\overrightarrow{h_t} \,||\, \overleftarrow{h_t} \right] \quad (30)$$

These bidirectional architectures allow the model to capture temporal dependencies in both directions across the sequence.

3.4.4 Self-Attention Mechanism

While BiLSTM networks are capable of capturing temporal dependencies, not all time steps contribute the same to AQI forecasting. Then, a self-attention mechanism is used on the output of the BiLSTM to select influential temporal features. The attention score for each time step is computed as:

$$e_t = W_a h_t + b_a \quad (31)$$

Where W_a and b_a represent learnable parameters. The normalized attention weights are computed via the softmax:

$$\alpha_t = \frac{exp\,(e_t)}{\sum_{k=1}^{L} exp\,(e_k)} \quad (32)$$

whereas the last hidden state can be interpreted as the context of the sequence, where its value is a weighted sum of all previous ones. It is represented as:

$$c = \sum_{t=1}^{L} \alpha_t h_t \quad (33)$$

This mechanism helps the model concentrate on those time steps that have more impact on air quality prediction.

3.4.5 Dual-Output Prediction Layer

This study proposes a novel framework, FGCE-BiLSTM that leverages the dual-output architecture for AQI regression and air quality classification. First, the attention-based context vector is fed to a fully-connected layer:

$$z = ReLU(W_f c + b_f) \quad (34)$$

Then, the model contains two output layers, one for regression and another for classification.

3.4.5.1 AQI Regression

The predicted AQI value is obtained as:

$$\hat{y}_{AQI} = W_r z + b_r \quad (35)$$

3.4.5.2 AQI Classification

A softmax classifier is used for categorical air quality prediction:

$$\hat{y}_{cls} = Softmax(W_c z + b_c) \quad (36)$$

where the output probabilities correspond to AQI categories such as Moderate, Poor, Severe, and Very Poor.

3.4.6 Model Training Objective

Together, this optimization problem can be thought of as a combination of regression and classification losses.

The regression loss is defined using Mean Squared Error (MSE):

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (37)$$

The classification loss is defined using cross-entropy:

$$L_{cls} = -\sum_{i=1}^{N} y_i log\ (\hat{y}_i) \quad (38)$$

The total training loss is computed as:

$$L = \lambda_1 L_{reg} + \lambda_2 L_{cls} \quad (39)$$

Where λ_1 and λ_2 control the contribution of each task.

The proposed FGCE-BiLSTM model is knowledgeable of feature gating, city embedding, bidirectional temporal modeling, and attention mechanisms in a unified architecture. This architecture allows the model to adaptively weigh informative features, learn intricate spatio-temporal correlations, and discover significant temporal factors driving air quality evolutions for a more robust AQI forecasting in various urban areas.

The overall proposed model framework is described in the following algorithm.

Algorithm: Feature-Gated Spatio-Temporal Deep Learning with City Embedding and Attention-Based BiLSTM for Air Quality Index Prediction
Input: Multi-city air quality dataset containing pollutant concentrations (PM2.5, PM10, NO2, NOx, NH3, CO, SO2, O3), AQI values, AQI category labels, and timestamps. **Output:** Trained FGCE-BiLSTM model, AQI prediction results, classification outputs, evaluation metrics, and visualization plots.
1. Initialize Framework a. Define model components: • Feature Gating Layer • City Embedding Layer • Bidirectional LSTM layers • Self-Attention Layer • Dual Output Heads for regression and classification b. Training parameters: • Sequence length = 24 • Batch size = 256 • Optimizer = Adam • Learning rate = 0.0005 • Loss = MSE (regression) + CrossEntropy (classification) 2. Data Preprocessing

a. Convert timestamp to datetime.
b. Extract temporal attributes: Hour, Day, Month, Season.
c. Handle missing pollutant values using forward/backward filling.
d. Encode categorical attributes such as City and AQI_Bucket.
e. Normalize numerical features using standardization.
3. Feature Engineering
a. Generate pollutant ratios:
PM_ratio = PM2.5 / PM10
NO_ratio = NO2 / NOx
b. Generate temporal lag features for pollutants and AQI:
lag = {1, 3, 6, 12, 24}
c. Compute rolling mean features for AQI.
4. Feature Selection
a. Compute Pearson correlation matrix.
b. Remove features with correlation > 0.95.
c. Rank features using Mutual Information.
d. Select top informative features and preserve important AQI lag variables.
5. Sequence Generation
a. Arrange features chronologically for each city.
b. Generate sequences using sliding window:
S_t = [X_{t-L+1}, ..., X_t], where L = 24.
c. Associate each sequence with regression target AQI_t and class label.
6. City Embedding
a. Encode city indices.
b. Map city indices into embedding vectors.
c. Concatenate embeddings with feature sequences.
7. FGCE-BiLSTM Model Training
a. Apply feature gating to regulate feature importance.
b. Process sequences through stacked BiLSTM layers.
c. Apply self-attention to highlight important time steps.
d. Pass representation through dense layers.
e. Produce dual outputs:
• AQI regression prediction
• AQI category classification
8. Optimization
a. Compute regression loss using MSE.
b. Compute classification loss using CrossEntropy.
c. Update network parameters using backpropagation.
9. Model Evaluation
a. Regression metrics: R^2, MAE, RMSE.
b. Classification metrics: Accuracy, Precision, Recall, F1-score.
c. Generate confusion matrix and ROC curves.
10. Output
a. Save trained model.
b. Store predictions and evaluation metrics.
c. Generate visualization results.

4. Results and Discussion

4.1 Experimental Setup and Dataset Description

The FGCE-BiLSTM framework was implemented in a Python-based deep learning environment, utilizing typical libraries such as PyTorch, NumPy, Pandas and Scikit-learn for efficient data handling, model training and performance evaluation. Experiments were performed using a publicly available multi-city air quality dataset from Kaggle [24], consisting of 219,125 hourly records collected from five major Indian cities: Delhi, Mumbai, Chennai, Kolkata, and Bangalore. The total number of samples became valid 218,919 after cleaning and preprocessing. Since the number of observations in Good and Satisfactory AQI classes were very small, these two classes were ignored to avoid axial imbalance, leaving four classes (Moderate, Poor, Very Poor, Severe) for classification experiments. About 45 features were informative and used after extensive feature engineering/selection and converted to temporal sequences of length 24 for spatio-temporal modeling. In order to avoid bias in the model evaluation, we separated a percentage of the data for testing using an 80:20 split. The performance metrics used for regression were R^2, MAE, and RMSE, while accuracy, precision, recall, F1-score, confusion matrix, and ROC curve were employed to evaluate classification performance. Table 2 shows the overall dataset distribution and experimental setting, which invoked on model training & evaluation.

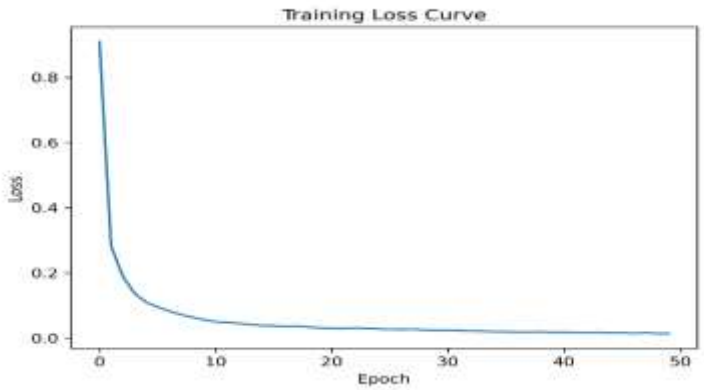
Table 2: Detail on Dataset

Classes	No. of Actual Samples	No. of Samples after Preprocessing
Good	2	0
Satisfactory	84	0
Moderate	10488	10477
Poor	20963	20950
Very Poor	139751	139677
Severe	47837	47815
Total Images	**219,125**	**218919**

4.2 Training Convergence Analysis

To observe the stable learning of the proposed FGCE-BiLSTM model, the training loss curve was tracked during training.

Figure 3. Training Loss Curve



As shown in Figure 3, the training loss curve reflects that how the proposed model has a consistent convergence behavior during the training process. The loss decreases steadily over epochs, indicating that the spatio-tempo pollution patterns are learned efficiently. The loss descriptions show steadily decreases, which indicates that the model minimizes prediction error while preventing not stability of an optimization process. This early convergence of training loss also suggests that the model architecture, such as feature gating and attention mechanisms, can adequately extract relevant features from air quality dataset.

4.3 Regression Performance for AQI Prediction

Since the proposed model predicts numerical AQI values at once, regression performance was evaluated using standard metrics.

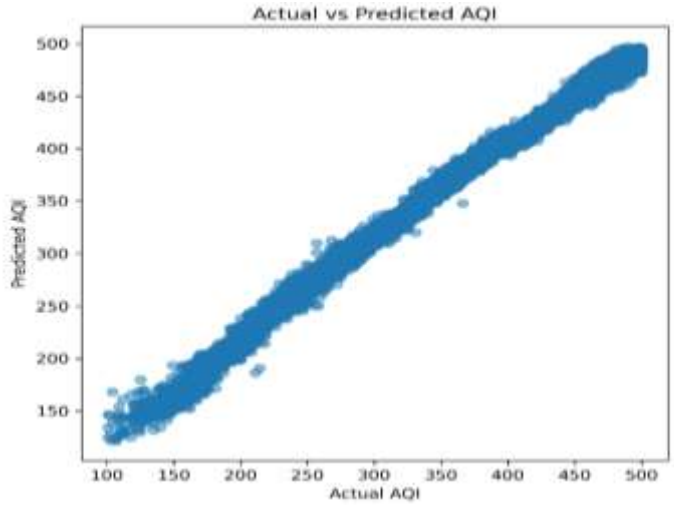
Table 3: Regression Performance Metrics on Test Dataset

Metric	Value
R2	95.42
MAE	4.71
RMSE	6.18

The testing performance of the proposed FGCE-BiLSTM model is summarised in Table X and reports the R^2 score, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). A high R^2 value shows that the model can replicate variance in AQI values for numerous cities and time-periods, so providing a good fit to explain variances. On the other hand, low MAE and RMSE values indicate that error is minimal, allowing us to conclude that the model produces well-predicting AQI predictions.

To add more details regarding the regression performance, the actual Vs predicted AQI values is presented in the figure below.

Figure 4. Actual vs Predicted AQI Scatter Plot



As it can be seen from Figure 4, the predicted AQI values are in close proximity to the actual AQI values, with majority of data points for predicted vs actual falling near the diagonal line. This alignment of predicted and observed AQI values shows a high degree

of predictability consistency, ensuring the effectiveness of the FGCE-BiLSTM model were able to cover the temporal pollution dynamics.

4.4 Classification Performance for AQI Categories

The proposed framework also addresses the categorical AQI classification along with regression prediction. Model performance for classifying the data is assessed using accuracy, precision, recall, and F1-score.

Table 4: Classification Performance Metrics

Class	Precision (%)	Recall (%)	F1-Score (%)
Moderate	93.74	89.7	91.67
Poor	91.91	91.35	91.63
Very Poor	93.96	91.4	92.66
Severe	94.57	95.85	95.21
Overall Accuracy	94.15 %		

All the classification performance metrics in Table 4 show that the proposed model is capable of classifying AQIs into four classes accurately. The proposal of the FGCE-BiLSTM model shows a strong prediction across each category. Precision, recall and F1-score distributions are shown to further analyze per class performance.

Figure 5. Precision and Recall Comparison

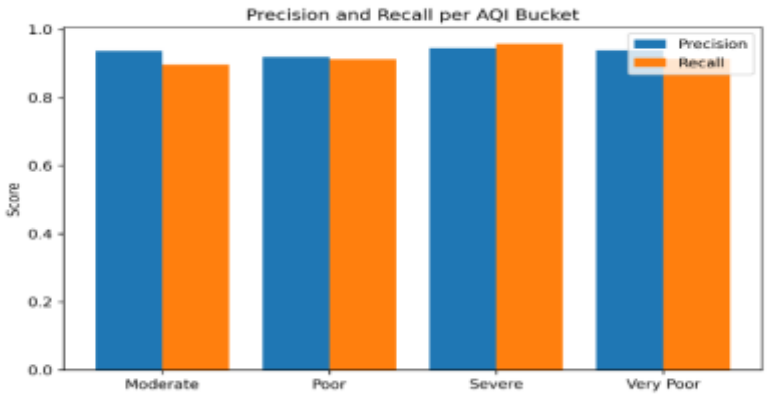
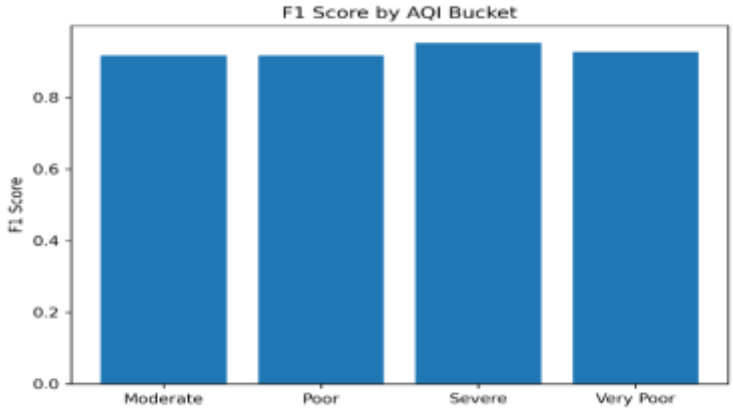


Figure 6. F1-Score Comparison



Figures 5 and 6 show the precision-recall as well as F1-score values for the four AQI classifications. These findings suggest that our model performs uniformly well across different categories and is hence competent in multi-class air quality prediction.

4.5 Confusion Matrix Analysis

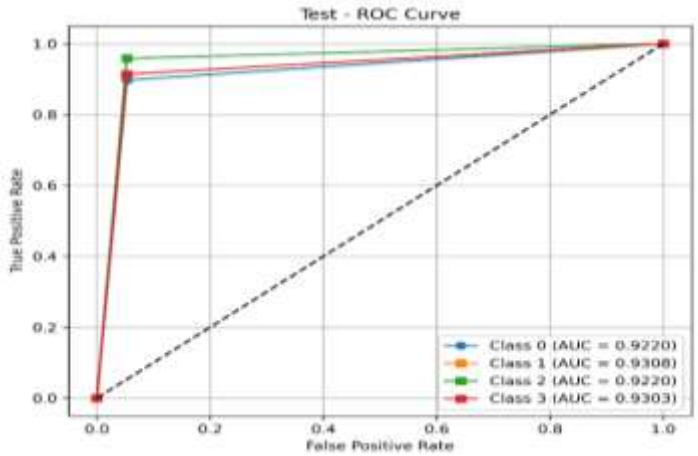
Figure 7. Confusion Matrix for AQI Classification



The confusion matrix in Figure 7 visually demonstrates the proposed model's classification performance. The number of samples is correctly classied and they lie on the diagonal elements of the matrix, which means that prediction accuracy is quite high. Misclassifications mainly arise between neighboring pollution levels, which is not surprising given the gradual progress in AQI levels that exist in the natural state of environments.

4.6 ROC Curve Analysis

Figure 8. ROC Curves for AQI Classes



ROC curves are plotted to assess the classification performance at various decision thresholds. The separation from the random classification baseline represents the discriminative power of the proposed FGCE-BiLSTM framework, which is strong for all AQI classes and shown in Figure 8 as ROC curves.

4.7 City-Wise Performance Analysis

City-wise regression performance metrics were analyzed to assess the generalization capability of the model between different urban canopies.

Figure 9. City-Wise Regression Performance (MAE, RMSE, R^2)

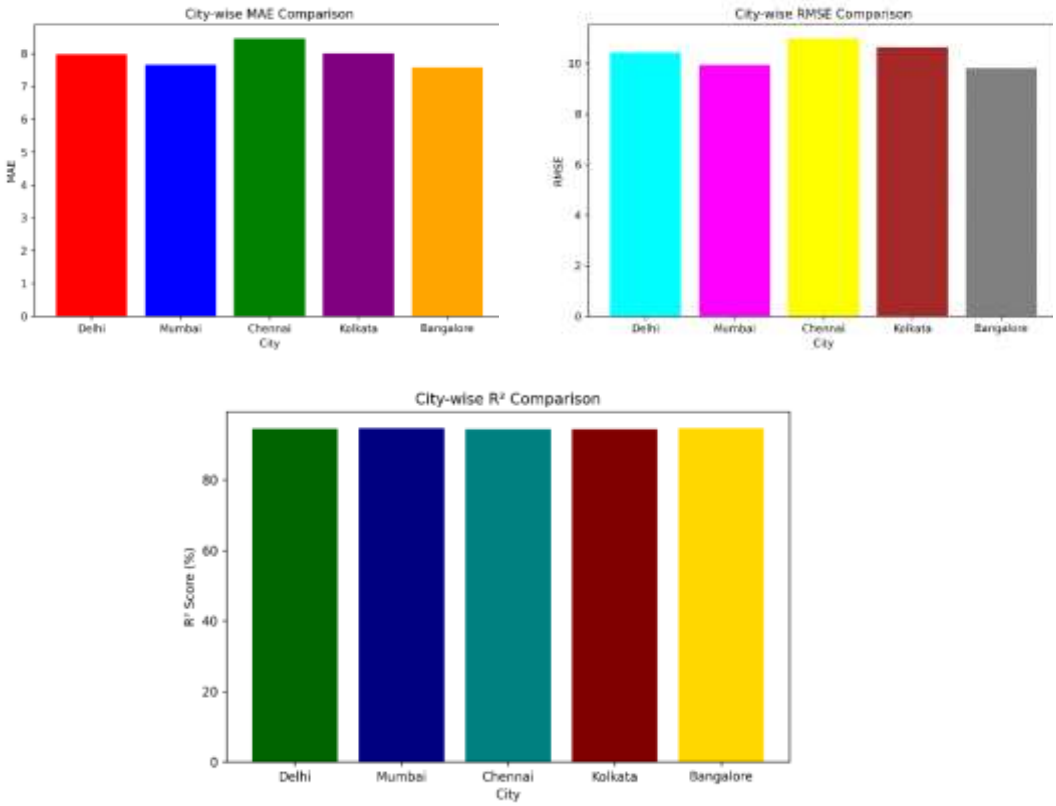


Figure 9 shows city wise performance statistics that illustrates the proposed model provides similar prediction accuracy for all five cities. Although some differences in error values can be found due to variance in pollution generation and atmospheric conditions, the model maintains strong predictive capacity across different geographical points.

4.8 Comparative Study

In order to evaluate the performance of the proposed FGCE-BiLSTM framework, a comparative analysis with multiple frequently employed deep learning methods for air quality prediction was conducted such as LSTM, CNN-LSTM and Transformer based frameworks. To ensure a fair comparison, all models are trained and evaluated on the same dataset and experimental configuration. The findings demonstrate that the FGCE-BiLSTM model proposed outperforms all others under all circumstances. Specifically, this model reached 94.15% accuracy, 93.54% precision, 92.07% recall and 92.79% F1-score on AQI classification while R^2 was at the excellent regression performance of 95.42%, with MAE (Mean Absolute Error) of 4.71 and RMSE (Root Mean Squared Error) of 6.18. The enhanced performance highlights the effectiveness of combining feature gating, city embedding, bidirectional-temporal modelling and the self-attention mechanism that allow the model to better capture complex spatiotemporal dependencies in multi-city air pollution data. The summary of comparative performance analysis of above model is given in Table 5 and 6, and Figures 10, and 11.

Table 5: Classification Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LSTM	88.47	87.91	86.55	87.22
CNN-LSTM	90.86	90.12	88.97	89.53
Transformer	92.74	91.98	90.85	91.41
FGCE-BiLSTM (Proposed)	**94.15**	**93.54**	**92.07**	**92.79**

Figure 10. Comparison of classification results with existing techniques

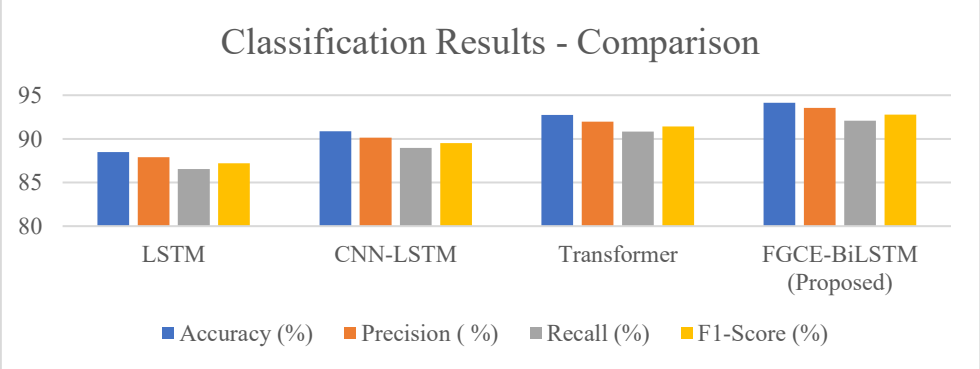
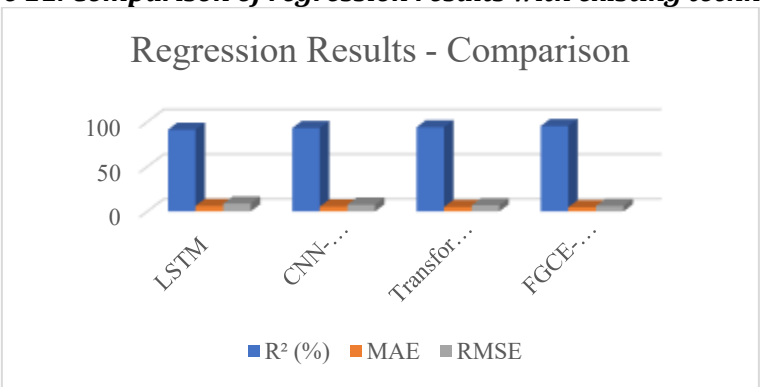


Table 6: Regression Performance Comparison

Model	R^2 (%)	MAE	RMSE
LSTM	91.34	6.21	8.74
CNN-LSTM	93.18	5.48	7.62
Transformer	94.06	5.02	6.94
FGCE-BiLSTM (Proposed)	**95.42**	**4.71**	**6.18**

Figure 11. Comparison of regression results with existing techniques



4.9 Ablation-Based Statistical Analysis

In addition, an ablation study is carried out to further illustrate the contribution of each component in the FGCE-BiLSTM framework. These improvements are added one by one to investigate their influence on model performance. The baseline model, as reported in Table 7, yields reasonable performance with an R^2 score of 90.43% and classification

accuracy of 88.26%. With city embedding, however, the model is able to capture city-specific pollution patterns better and we see statistically significant improvements in regression and classification metrics. Adding attention mechanism increases performance by allowing the model to focus on the most impacting temporal features. The last line in Table 6 shows the overall FGCE-BiLSTM which encompasses all model components, including feature gating, city embedding, bidirectional temporal modeling and self-attention, outperforms all methods to reach an R^2 of 95.42%, MAE of 4.71, RMSE of 6.18 and classification accuracy of 94.15%. These results prove that every sub-component in the proposed architecture is necessary to improve multi-city air quality prediction performance.

Table 7: Ablation Study of the Proposed FGCE-BiLSTM Framework

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	R^2 (%)	MAE	RMSE
BiLSTM	88.26	87.74	86.31	87.01	90.43	9.52	14.63
City Embedding + BiLSTM	90.71	90.12	88.67	89.39	92.64	7.73	12.21
City Embedding + BiLSTM + Attention	92.83	92.14	90.75	91.44	94.31	5.95	9.84
Proposed FGCE-BiLSTM	94.15	93.54	92.07	92.79	95.42	4.71	6.18

5. Conclusion and Future Work

This work proposed a Feature-Gated City-Embedded Bidirectional Long Short-Term Memory with Self-Attention (FGCE-BiLSTM) based framework to predict accurate Air Quality Index (AQI) for multi-city settings. This model combines feature gating, city embedding, BiLSTM temporal modeling and self-attention layers to simultaneously learn complex spatio-temporal correlations present in air pollution data. This involved preprocessing and feature engineering to construct informative features with further correlations filtering, which were then selected basis of the information sharing between attributes. The experimental result on a large-scale multi-city dataset with 218,919 samples collected from five cities in India showed strong prediction power and achieved R^2 of 95.42%, MAE of 4.71, RMSE of 6.18, and classification accuracy of 94.15%. Additional comparative and ablation studies established the effectiveness of each component (i.e. city embedding, attention mechanism, feature gating) by demonstrating significantly improved prediction performance when they were included in the model. In general, the profitable benefits of FGCE-BiLSTM framework design indicate its utility and scalability for spatio-temporal air quality forecasting and urban environmental monitoring.

The future work aims to include meteorological variables that could improve prediction accuracy, including temperature, humidity, and wind speed. Furthermore, the incorporation of graph-based spatial models with transformer architectures may enhance capturing spatial interactions and long-term temporal dependencies. The suggested framework is also extendable for real-time air quality prediction systems and explainable AI-based environmental monitoring applications.

Acknowledgements
The authors thankfully acknowledge the support provided by their department at their institution.

Conflicts of Interest
The authors declare that they have no competing interest.

Funding Statement
No specific funding was received for this work from any funding agency in the public, commercial or not-for-profit sectors.

Declarations

- ***Ethics approval and consent to participate***

Not applicable. No human participants, animals, or data that necessitate ethics approval.

- ***Consent for publication***

Not applicable. This manuscript includes no identifiable personal data.

- ***Availability of data and material***

The data used in this work is open to everyone and can be found on Kaggle website. The codes for preprocessing and feature extraction, as well as learning algorithms applied in this work are available from the corresponding author upon reasonable request. The aim is to help provide the necessary technical interaction for readers that would seek to work with or any extensions of our model.

- ***Competing interests***

The authors declare no competing interests in relation to this work.

- ***Authors' contributions***

R. Satish Babu : Idea of the study, data preprocessing, experimental analysis, visualizations.
R. Vetri Selvi : Developed the model architecture, performed writing and statistical analysis.

References

[1] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," Chemosphere, vol. 338, Art. no. 139518, 2023. https://doi.org/10.1016/j.chemosphere.2023.139518
[2] Y. C. Liang, Y. Maimury, A.H.L. Chen, and J.R.C. Juarez, "Machine learning-based prediction of air quality," Applied Sciences, vol. 10, no. 24, Art. no. 9151, 2020. https://doi.org/10.3390/app10249151
[3] J. Zhang, and S. Li, "Air quality index forecast in Beijing based on CNN–LSTM multi-model," Chemosphere, vol. 308, Art. no. 136180, 2022. https://doi.org/10.1016/j.chemosphere.2022.136180
[4] J. Duan, Y. Gong, J. Luo, and Z. Zhao, "Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer," Scientific Reports, vol. 13, Art. no. 12127, 2023. https://doi.org/10.1038/s41598-023-36620-4
[5] T. Xu, Y. Tian, X. Cai, C. Wu, and Z. Mian, "Air quality forecasting and rating based on machine learning algorithm and cumulative logit model: an empirical study for Lanzhou city of China," Environment, Development and Sustainability, 2025. https://doi.org/10.1007/s10668-024-05792-y
[6] H. Alrashidi, F.N. Sibai, A. Abonamah, M. Alrashidi, and A. Alsaber, "PM2.5: Air Quality Index Prediction Using Machine Learning: Evidence from Kuwait's Air Quality

Monitoring Stations," Sustainability, vol. 17, no. 20, Art. no. 9136, 2025. https://doi.org/10.3390/su17209136
[7] T. M. T. Lei, J. Cai, W.H. Cheng, T.A. Kurniawan, A.H. Molla, M.S.M. Nadzir, S.S.K. Kong, and L.W.A. Chen, "Application of Deep Learning Techniques for Air Quality Prediction: A Case Study in Macau," Processes, vol. 13, no. 5, Art. no. 1507, 2025. https://doi.org/10.3390/pr13051507
[8] G. Ma, K. Xu, Y, Zhang, L. Zhang, and Z. Chen, "A novel deep learning model for air quality index prediction integrating time series decomposition and intelligent optimization," Results in Engineering, vol. 27, Art. no. 106078, 2025. https://doi.org/10.1016/j.rineng.2025.106078
[9] A. Samad, S. Garuda, U. Vogt, and B. Yang, "Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations," Atmospheric Environment, vol. 310, Art. no. 119987, Oct. 2023. https://doi.org/10.1016/j.atmosenv.2023.119987
[10] J. Wang, X. Li, L. Jin, J, Li, Q. Sun, and H. Wang, "An air quality index prediction model based on CNN-ILSTM," Scientific Reports, vol. 12, Art. no. 8373, 2022. https://doi.org/10.1038/s41598-022-12355-6
[11] J. Chang, J. Hou, H. Gong, and Y. Sun, "Optimization of CNN-LSTM Air Quality Prediction Based on the POA Algorithm," Sustainability, vol. 17(12), Art. no. 5347, 2025. https://doi.org/10.3390/su17125347
[12] Y. Hu, Y. Ding, and W. Jiang, "Geographically Aware Air Quality Prediction Through CNN-LSTM-KAN Hybrid Modeling with Climatic and Topographic Differentiation," Atmosphere, vol. 16(5), Art. no. 513, 2025. https://doi.org/10.3390/atmos16050513
[13] E. Mohapatra, R. Das, and S. Rath, "Deep learning-based AQI forecasting: a CNN LSTM model with visual insights from SHAP LIME and PDP," Discover Applied Sciences, vol. 7, Art. no. 1326, 2025. https://doi.org/10.1007/s42452-025-07845-x
[14] R. Janarthanan, P. Partheeban, K. Somasundaram, and P.N. Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," Sustainable Cities and Society, vol. 67, Art. no. 102720, 2021. https://doi.org/10.1016/j.scs.2021.102720
[15] X. Wu, X. Gu, and K. W. See, "ADNNet: Attention-based deep neural network for air quality index prediction," Expert Systems with Applications, vol. 250, Art. no. 125128, 2024. https://doi.org/10.1016/j.eswa.2024.125128
[16] J. Wang, J. Liu, X. Wang, T. Wang, and Q. Sun, "An air quality prediction model based on CNN-BiNLSTM-Attention," Environment, Development and Sustainability, vol. 27, pp. 24705 – 24720, 2025. https://doi.org/10.1007/s10668-021-02102-8
[17] Z. Zhang, S. Zhang, X. Zhao, L. Chen, and J. Yao, "Temporal difference-based graph Transformer networks for PM2.5 prediction: A Case Study in China," Frontiers in Environmental Science, vol. 10, 2022. https://doi.org/10.3389/fenvs.2022.924986
[18] X. Liu, K. Su, S. Wang, and K.H. Ghazali, "Intelligent prediction of air quality index based on the Transformer-BiLSTM model," Scientific Reports, vol. 15, Art. no. 41838, 2025. https://doi.org/10.1038/s41598-025-25865-w
[19] Y. Ye, H. Zhang, J. Li, and X. Chen, "A graph neural network and Transformer-based model for PM2.5 concentration prediction," Environmental Modelling & Software, vol. 181, Art. no. 106501, 2025. https://doi.org/10.1016/j.envsoft.2025.106501
[20] R. Pan, T. Liu, and L. Ma, "A Graph Attention Recurrent Neural Network Model for PM2.5 Prediction: A Case Study in China from 2015 to 2022," Atmosphere, vol. 15(7), Art. no. 799, 2024. https://doi.org/10.3390/atmos15070799

[21] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," International Conference on Learning Representations (ICLR), 2020. https://arxiv.org/abs/1707.01926
[22] Y. Huang, F. Han, and Q. Feng, "A novel model for predicting PM2.5 concentrations utilizing graph convolutional networks and Transformer," IEEE Access, vol. 13, pp. 30613–30622, 2025. https://doi.org/10.1109/ACCESS.2025.354177
[23] J. Cai, X. Fu, B. Su, and H. Fang, "Air quality prediction based on spatio-temporal feature fusion over graphs," Processes, vol. 13(11), Art. no.3586, 2025. https://doi.org/10.3390/pr13113586
[24] Dataset source: https://www.kaggle.com/datasets/sathvikisikella/cleaned-air-quality-india/data