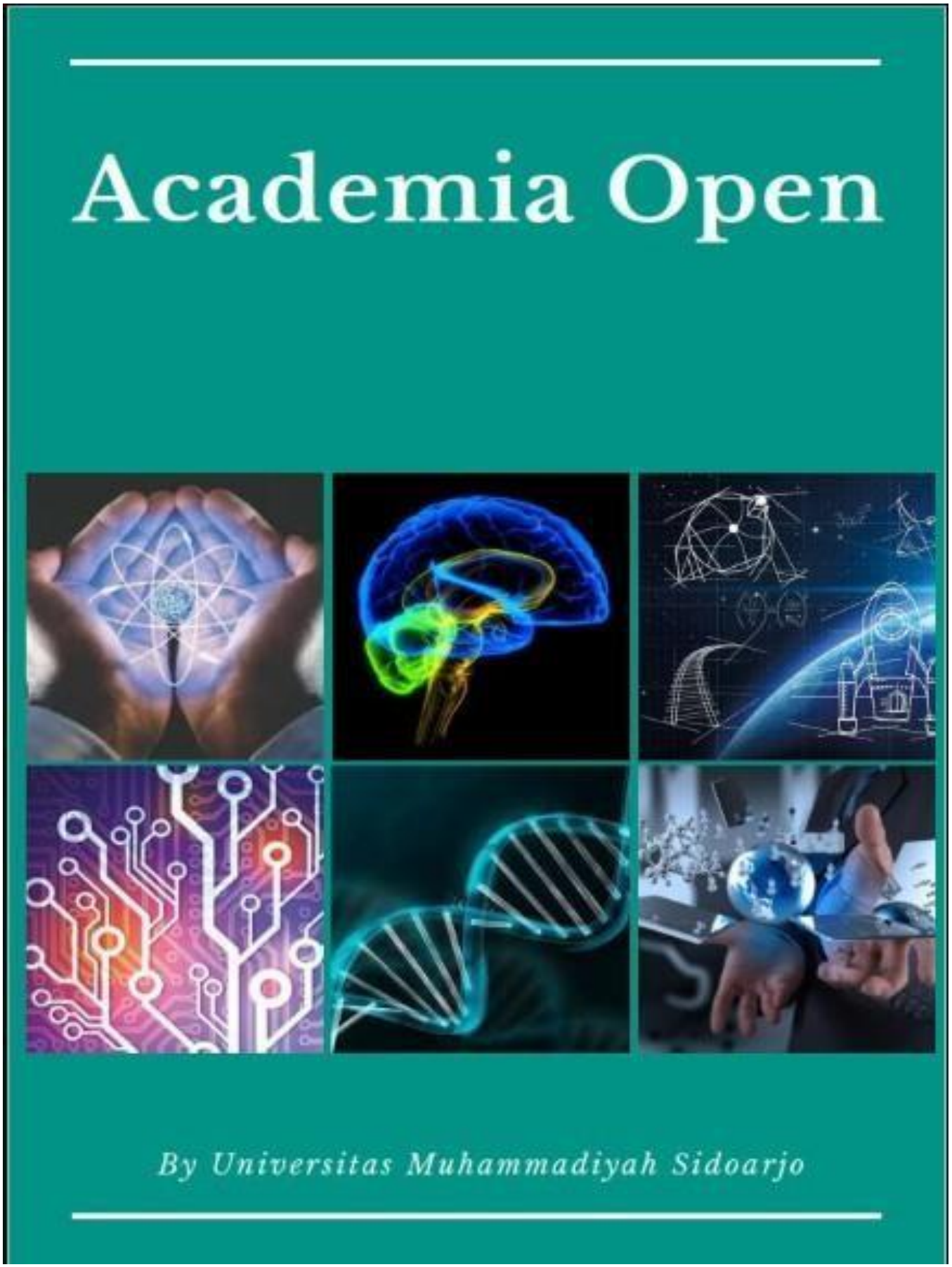
Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837



ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

Table Of Contents

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

EDITORIAL TEAM

Academia Open

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

A Hybrid Transformer–BiLSTM–Attention Framework for High Accuracy Multivariate Air Quality Prediction

Nebras Jalel Ibrahim, nebras.jalel@uodiyala.edu.iq (*)
Diyala University, Computer Center, Diyala, Iraq

(*) Corresponding author

Abstract

General Background: Air pollution has become a critical global issue affecting environmental sustainability and public health, creating a strong demand for accurate air quality prediction systems. **Specific Background:** Traditional statistical models and conventional machine learning techniques often struggle to capture the nonlinear and multivariate characteristics of environmental data, particularly when dealing with complex temporal dependencies. **Knowledge Gap:** Many existing forecasting approaches focus primarily on either short-term sequential learning or long-range temporal modeling, which limits their ability to represent both bidirectional temporal patterns and long-term dependencies in multivariate air quality datasets. **Aims:** This study proposes a hybrid deep learning framework integrating Transformer, Bidirectional Long Short-Term Memory (BiLSTM), and an Attention mechanism for accurate multivariate air quality prediction. **Results:** Experiments conducted on the UCI Air Quality dataset demonstrate that the proposed model achieves superior predictive performance with RMSE of 0.0799, MAE of 0.0589, and R^2 of 0.9621, outperforming baseline models such as standalone Transformer and BiLSTM architectures. **Novelty:** The proposed framework combines global temporal dependency modeling from Transformer encoders with bidirectional sequence learning from BiLSTM and adaptive temporal weighting through the attention mechanism. **Implications:** The framework provides a reliable computational approach for environmental monitoring systems, supporting intelligent air quality forecasting, early warning mechanisms, and data-driven environmental decision-making.

Highlights

- Hybrid architecture captures both long-range temporal dependencies and bidirectional sequence relationships in environmental data.
- Multivariate forecasting shows strong predictive consistency across several pollutants and meteorological variables.
- Experimental evaluation reports very low prediction errors and strong statistical correlation with observed measurements.

Keywords: Air Quality Prediction, Multivariate Time Series, Hybrid Deep Learning, Transformer, BiLSTM Model, Environmental Monitoring

Published date: 2026-03-03

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

1. Introduction

Cities expanding and economies developing drive air pollution. For cities today, air pollution is thought to be a serious issue since it negatively affects people's health and well-being in so many ways. People in cities don't care as much about lowering air pollution or making things green. The distribution of vegetation, air quality indices (PM2.5, PM10, CO2, and AQI), and the health risks connected with air pollution for city people across regions. For this reason, contemporary environmental management systems include rather accurate air quality predictions. It supports wise policy decisions, early warning systems, and intelligent city design.

Air quality forecast methods now being used are based primarily on statistical regression or deterministically computer modeled , based upon oversimplified assumptions about how pollutants behave and how as the atmosphere impacts all pollutants; even though these forecasting methods have been very widely applied, their ability to successfully predict variations and changes to air quality are typically limited, especially with regard to the complexity of real-world air pollution (highly nonlinear, nonstationary , and multidimensional) data. To get beyond these restrictions, modern studies have used more and more machine learning and deep learning techniques, which provide more freedom in simulating complex relationships between different points in time and between different variables.

Two kinds of RNNs (recurrent neural networks) that have been utilized extensively to forecast air quality are Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM). Compared to conventional techniques, their capacity to record temporal relationships has produced significant improvements [1, 2]. Despite the excellent performance of long-term neural network (LSTM) models, they typically focus on short-term patterns and may struggle to accurately depict long-term correlations. This is particularly evident when attempting to predict how different pollutants will interact over longer time periods.

Recent studies in time series prediction have focused on attention mechanisms and transformer designs to address long-term dependency models. Transformers, originally designed to model self-attention-based sequences, enable the direct modeling of global time relationships without the need for iteration. Recent research shows that transformer-based models, especially with complex multivariable inputs [3, 4], outperform conventional recurrent neural networks in predicting environmental and air quality measurements.

A new field of study, leveraging the best aspects of various deep learning techniques, seeks to address these problems. Combining transformers with recurrent networks allows models to utilize sequential inductive bias and comprehensive temporal awareness. For environmental data, where certain times or conditions significantly influence pollutant levels, adding attentional techniques enables a focus on the most important temporal steps and characteristics. According to recent research, hybrid models with enhanced attentional capabilities, compared to single-structure models, perform better in terms of the accuracy and reliability of air quality estimates [5, 6].

Building on these developments, this study presents a hybrid model combining Transformer, BiLSTM, and Attention for accurate multivariate air quality prediction. The proposed design leverages the properties of contrasting modeling approaches; the Transformer encoder captures interactions between features and long-term temporal dependencies. By processing data in both forward and bidirectional, the BiLSTM component enhances sequential learning. The Attention mechanism dynamically identifies the most relevant temporal patterns for prediction. This approach is well-suited for complex environmental datasets with multiple components that vary over time .The Air Quality UCI dataset, which consists of real hourly data gathered from a network of chemical sensors set up in an urban area together with meteorological variables, is used to assess the suggested model. Advanced forecasting models can be tested well with the dataset since it reveals real-world monitoring scenarios like missing values and multivariate coupling. Below are the main points of this study's contributions:

- Design a new hybrid model that contains three key components: 1) Transformer; 2) Bi-LSTM; and 3) an attention-based mechanism. This hybrid model will be able to model both long-range dependencies and bidirectional temporal dynamics for the purpose of predicting multivariate air quality in urban environments in Italy.·
- The proposed model demonstrates the relationships between pollutants by demonstrating how pollutants interact with each other and are dependent upon each other. This will lead to more accurate assessments of future air quality, both currently and in the long term.·
- The proposed framework encourages the creation of advanced environmental monitoring and early warning systems that will have a real effect on how air quality is managed and decisions are made.·
- This research contributes to the transition towards more sustainable cities, a low-emission economy, and a healthier future for generations to come.

2. Literature Review

Air Quality is a global concern and a major public health/environmental issue. Airborne particulate matter (like PM2.5, PM10), as well as gases (like NH3, NO2, O3, and SO2), can cause death through excess mortality by respiratory or cardiovascular diseases. Airborne contaminants also cause much economic loss. Therefore, being able to predict the concentrations of these pollutants provides important information for the purpose of early warning, exposure mitigation, and policy support by providing accurate short- and medium-term forecasts. Therefore, from a modeling perspective, predicting air quality contaminant concentrations is a challenging multivariate time series forecasting task since local emissions and meteorological conditions will have an impact on pollutant concentrations and generate time series dependencies between

pollutant concentrations, as well as spatial correlations between pollutant concentrations from different monitoring stations [7][8][9][10].

Early adoption of Transformers in air-quality forecasting focused on univariate targets (often PM2.5) with multivariate inputs.Cao et al.'s TD-CS-Transformer decomposes PM2.5/PM10 series into trend, seasonal and irregular components, then uses a convolutional sparse self-attention Transformer to model long sequences efficiently [11]. Time-series decomposition simplifies patterns and reduces model complexity, while sparse convolutional attention improves long-range dependency capture; TD-CS-Transformer outperforms conventional deep baselines on long-sequence PM2.5/PM10 forecasting [11]. Several works embed Transformers within broader systems. Xu & Jiang construct a CNN–Transformer daily AQ forecasting system with multisource data fusion; their CNN–Transformer model achieves lower RMSE and MAE than standalone LSTM and plain Transformer models, while being integrated into an online system for multiscale, realtime forecasts [12]. Zhang et al. propose an LSTM–Transformer with adaptive temporal attention, where an LSTM first encodes historical airquality and weather data and a Transformer with an adaptive attention mechanism focuses on informative timesteps; this hybrid outperforms LSTM and a CNN–BiLSTM–Attention baseline on Jiaozuo data [13].He et al. compare six deep models (RNN, ANN, CNN, BiLSTM, Transformer, and a CNN–BiLSTM–Transformer hybrid) for daily PM2.5 forecasting in Qingdao . Their hybrid extracts local patterns via CNN, captures bidirectional temporal dependencies via BiLSTM, and enhances global temporal patterns and salient information via a Transformer block. It achieves the lowest RMSE and MAE and the highest correlation coefficient RR, outperforming all individual components [14].

Zou et al.'s PDLLTransformer tackles hourly PM2.5 across the Yangtze River Delta with a polydimensional embedding layer, a local LSTM block, and a Transformer over the enriched embeddings . The polydimensional embedding fuses pollutant, meteorological and satellite AOD features; the local LSTM captures shortterm dynamics, while the Transformer models global temporal interactions. PDLLTransformer surpasses LSTM and TCN baselines in accuracy [15]

Wang et al. proposed MSTTNet, which couples multiscale Temporal Convolutional Networks (TCNs) with a Transformer for PM2.5 forecasting in multiple Chinese cities [16]. Multiscale TCNs capture local correlations at different temporal resolutions, and the Transformer handles global temporal dependencies. MSTTNet outperforms LSTM and CNNbased models, demonstrating that TCN + Transformer hybrids are effective for multi city AQ prediction .

Chen et al.'s integrated dualLSTM framework trains perpollutant seq2seq LSTM models (singlefactor) and a multifactor LSTM with attention using neighborstation and weather inputs, then fuses them via XGBoost; this ensemble improves both error and model expressiveness relative to single models [17].Mo et al.'s TSTM framework uses two CNN–BiLSTM–Attention encoders one for "pollution source" variables (time, space, type) and one for "meteorology"—and fuses them via ConvLSTM to produce multistep pollutant concentrations, AQ levels, chief pollutant types, and heavy pollution events; the multioutput, multistream design yields accuracy gains across these tasks [18]. Nguyen et al. design a pipeline where ARIMA removes linear components, an Attention CNN (ACNN) encoder with multihead attention and multiscale convolutions feeds into a BiLSTM decoder with masked attention, and XGBoost refines AQI predictions for Seoul; they report up to ∼31% MSE reduction and ∼19% MAE reduction vs conventional models [19].Other works refine temporal modeling and linear trends. Wang & In Zhu's DLARN, a combined use of Convolutional Neural Networks (CNN) and Bi-directional Long Short-Term Memory networks (BiLSTM), with temporal attention and an explicit Autoregressive (AR) module, is used to model linear trends in air quality series; thus improving prediction performance by 7.04% to 10.81% versus the state-of-the-art baseline results [20]. Liu et al. employ a two-layer LSTM for temporal encoding and a Transformer with multi-head self-attention with residual connections on top, which yields better accuracy (RMSE) and stronger correlations (R^2) than the two previous methods on an urban multivariate dataset [21].

In light of this context, this study intends to critically analyse existing models that here are both Transformer and BiLSTM/GRU-based architectures; as well as models that use attention mechanisms and hybridisation of either Transformers or BiLSTMs to forecast multivariate air quality, focusing on patterns of design in these architectures that can assist with providing a Hybrid Model Framework for accurate multi-pollutant and multi-station forecasting of air quality. The review first surveys Transformer-centric approaches, then BiLSTM/GRU- attention hybrids, and finally graph/ConvLSTM spatio-temporal models and explicit Transformer-(Bi)LSTM hybrids. By organizing and comparing these strands, we identify common architectural principles, strengths and limitations, and open gaps that motivate the proposed hybrid framework. Table 1 summarizes the relevant literature reviewed in this work, presenting a comparison of previous works in terms of models used, datasets, and research outcomes.

Table1: summarizes of the Literature Review

Work	Model Type	Target & Data	RMSE (Hybrid / Proposed)	R^2 (Hybrid)
Liu et al. [21]	LT-Hybrid (2-layer LSTM + Transformer + attention)	Urban AQ dataset, 9 features incl. PM2.5	0.1021	$R^2 = 0.9382$
He et al. [14]	CNN–BiLSTM–Transformer	Daily PM2.5 in Qingdao, met + pollutant inputs	5.4236	$R^2 \approx 0.95$
Zhang et al. [13]	LSTM–Transformer + adaptive temporal attention	AQ + weather data, Jiaozuo (2015–2022)	N/R (lower than LSTM)	N/R (higher R^2 than LSTM)
Xu & Jiang [12]	CNN–Transformer vs. LSTM and plain Transformer	Daily AQ forecasting with multi-source data	N/R (best among compared)	N/R
Cao et al. [11]	TD-CS-Transformer vs. LSTM/GRU/TCN, etc.	Long sequence PM2.5/PM10	N/R (lowest among models)	N/R

Nguyen et al. [19]	ACNN + BiLSTM + attention + XGBoost vs. LSTM, GRU, etc.	AQI, 25 Seoul stations, 6 pollutants	N/R	N/R
Wang & Zhu [20]	DLARN (CNN + BiLSTM + Att + AR) vs. BiLSTM, LSTM	Two AQ datasets	N/R	N/R
Zou et al. [15]	PD-LL-Transformer vs. LSTM/TCN	Hourly PM2.5 over Yangtze River Delta	N/R	N/R
Wang et al. [16]	MSTTNet (multi-scale TCN + Transformer) vs. LSTM/CNN	PM2.5 in multiple Chinese cities	N/R	N/R

3. Methodology

This section discusses the methodological framework utilized in this work, which includes the dataset, preprocessing technique, construction of the suggested Hybrid Transformer-BiLSTM-Attention model, and training and evaluation processes.

3.1 Dataset Description

The dataset includes 9358 hourly averaged answers from a network of 5 metal oxide chemical sensors implanted in an Air Quality Chemical Multisensor Device. The device was found on a field in a heavily polluted location, at street level, in an Italian city. Data were collected from March 2004 to February 2005 (one year), making them the longest openly available recordings of on-field deployed air quality chemical sensor device responses. A co-located reference certified analyzer supplied ground truth hourly averaged quantities of CO, non-metanic hydrocarbons, benzene, total nitrogen oxides (NOx), and nitrogen dioxide (NO2)[22]. Evidence of cross-sensitivities, as well as concept and sensor drifts, are apparent, as stated in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required), ultimately compromising sensor concentration estimate capabilities. Missing values are marked with a -200 value [23].

3.2 Data Processing

Before modeling, a structured preparation approach is utilized to make sure that the air-quality dataset is clean, consistent, and good for deep learning. This step is very important since time-series data on air quality often has missing values, features that change size, and variables that are noisy or not needed. The main steps in the preprocessing stage are as follows:

3.2.1 Missing Value Imputation

Missing Values imputation (MVI) has been researched for decades as a primary approach to addressing problems with incomplete datasets. particularly when a dataset contains one or more missing attribute values. The missing data is completed using a completion technique such as forward/backward interpolation or statistical interpolation, to maintain the consistency of the time series and avoid losing key records. [24].

3.2.2 Normalization

Normalization is a preprocessing stage for any type of problem. It plays a crucial role in fields like elastic computing , cloud computing, and others, enabling data processing such as miniaturizing or maximizing data before use in later stages. Several normalization techniques exist, but we use the minimum and maximum normalization in the proposed model. This makes training more stable and helps the model reach its final state faster [25].

3.2.3 Time-Series Windowing

One way to use time series data in deep learning is Time-Series Windowing[26].This method relies on sequential sampling windows for model prediction, meaning it divides the time series into smaller segments consisting of two parts:

- Input: A data sequence containing a specific number of data points (previous time steps)
- Output: The corresponding next data points in the sequence (future time steps) .

3.2.4 Feature Selection

Feature selection, as an approach to data preprocessing, has proven effective and efficient in preparing data (especially high-dimensional data) for various data mining and deep learning problems. The objectives of feature selection include building simpler and clearer models, improving data mining performance, and producing clean and understandable data. This choice depends on the relevance of the data to the pollutants we want to eliminate. This allows the model to focus on the most useful inputs, such as sensor readings, critical pollutant data, and climate parameters[27].

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

3.3 Proposed Hybrid Model Architecture

The proposed approach uses a hybrid deep learning architecture that combines the best features of the Transformer encoder, a bidirectional long-term memory network (BiLSTM), and an attention mechanism to improve multivariable air quality prediction. The proposed approach uses a hybrid deep learning architecture that combines the best features of the Transformer encoder, a bidirectional long-term memory network (BiLSTM), and an attention mechanism to improve multivariable air quality prediction. Every element has a distinct purpose in gathering various facets of time-series data. The architecture of the suggested hybrid model is shown in Figure 1. To acquire high-level representations of the input sequences, the Transformer encoder initially employs a self-attention technique [28]. Long-range dependencies and general correlations between various air quality indicators are well captured by this component of the model. This model aids in comprehending how intricate relationships evolve over time.The BiLSTM layer receives the gathered characteristics and processes the sequence in both directions. This improves the model's capacity to simulate how pollutant concentrations fluctuate over time by allowing it to simultaneously learn past and future patterns [29]. The model can selectively focus on significant temporal aspects thanks to the attention mechanism, which permits a dynamic distribution of weights on hidden states. In multivariable air quality forecasting tasks, this improves the model's prediction performance by strengthening its capacity to learn long-term dependencies [30].

Lastly, the output layer makes precise predictions regarding the concentrations of air pollutants using the derived representations. The proposed hybrid model effectively embodies global and local temporal interactions by integrating these complementary components, thus improving prediction and generalization performance.

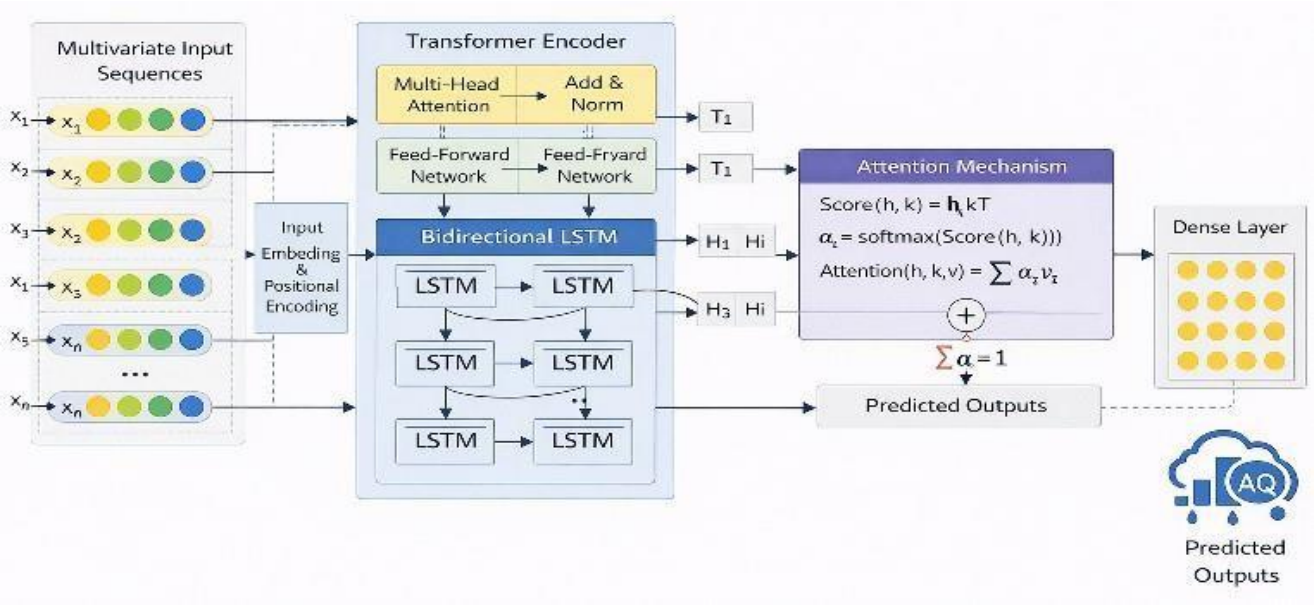


Figure 1. Proposed Hybrid Model Architecture

3.4 Model Training

The proposed hybrid model learns using a pre-selected dataset that includes multivariate air quality measurements. To ensure the consistency of the learning process and the fairness of the evaluation, the data is divided into training, validation, and testing subsets. The training subset aims to identify patterns in the data, while the validation subset helps to fine-tune the hyperparameters and monitor the model's performance throughout the training period. Table 2 shows the proposed training settings for the long-term, bidirectional, attention-based hybrid model, including an overview of the basic hyperparameters and the optimization settings needed to produce the results of the experiment.

Table 2: The training configuration of the proposed hybrid model

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Mean Squared Error (MSE)
Batch Size	32
Number of Epochs	50
Regularization	Dropout (0.3)
Early Stopping	Based on validation loss
Evaluation Metrics	MAE, RMSE, R^2

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

3.5 Evaluation Metrics

The Evaluation metric plays a crucial role in identifying the optimal classifier during classification training. Therefore, selecting the appropriate assessment tool is key to differentiating between classifiers and achieving the optimal classifier. The performance of the suggested hybrid model is clearly and consistently measured using a number of widely used regression indicators. These metrics demonstrate the degree to which the actual and expected levels of air pollution coincide. The root mean squared error (RMSE) is the square root of the average of the squared errors. It is a useful measure for numerical predictions and is primarily used to compare the prediction errors of different models or configurations of the same variable, due to its dependence on the scale. RMSE measures how well a regression line fits the data[31]. It has proven remarkably efficient in identifying models that lead to large predictive errors in air quality forecasting, as illustrated in Equation (1).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \quad (1)$$

The mean absolute error (MAE) is a metric used to measure the average size of errors between predicted and actual values in time series prediction. It calculates the absolute difference between each predicted value and its corresponding observed value, and then averages these differences. Unlike metrics that square errors (such as mean squared error), the mean absolute error treats all errors equally, making it easy to understand and resistant to outliers [32].The mean absolute error is a straightforward and practical way to measure how far your predictions deviate from actual values. It is easy to calculate and explain, and useful in a wide range of applications., as shown in Equation (2).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \quad (2)$$

The coefficient of determination (R^2) is a statistical measure used in regression analysis. In regression, we typically deal with dependent and independent variables. Any change in the independent variable is likely to lead to a change in the dependent variable[33]. The model can provide better predictions and the expected and actual measurements are likely to match as shown in equation (3).

$$R2 = 1 - \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{\sum_{t=1}^{n}(A_t - \bar{A})^2} \quad (3)$$

These evaluation metrics work together to provide a complete and fair picture of the model's accuracy, stability, and generalizability. Figure 2 illustrates the research methodology and the components of the proposed hybrid model.

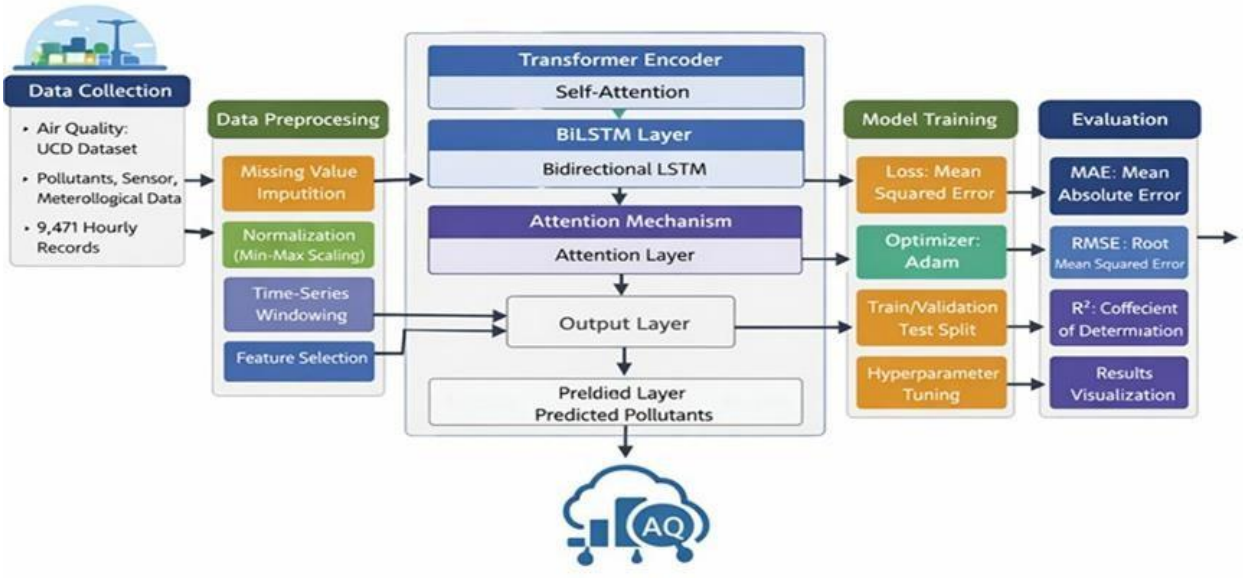


Figure 2. A hybrid deep learning framework for predicting air quality

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

4. Experimental Setup

To obtain the best results, all experiments were conducted using the same software environment and computer resources. Table 3 illustrates the experimental setup for data preparation for this work.

Table 3: Experimental Setup

Item	Description
Dataset	Air Quality UCI Dataset
Data Type	Multivariate time-series
Training Set	70%
Validation Set	10%
Test Set	20%
Data Normalization	Min–Max scaling
Framework	Tensor Flow / Keras
Hardware	GPU-enabled workstation

4.1 . Baseline Models

To objectively assess the efficacy of the proposed methodology, it is juxtaposed with two well-established deep learning benchmarks:

4.1.1 BiLSTM

Bidirectional Long Short-Term Memory (BiLSTM) is an expansion of the regular LSTM network. Unlike standard Long Short-Term Memory (LSTM) systems, which process sequences in only one direction, BiLSTMs allow information to flow both forward and backward, allowing them to collect more contextual data. This makes BiLSTMs especially useful for tasks that need knowing both the past and the future context. A Bidirectional LSTM (BiLSTM) is composed of two distinct LSTM layers. Forward LSTM processes a series from beginning to end, while backward LSTM processes the sequence from end to begin. The outputs of the two LSTMs are then merged to create the final result.

4.1.2 Transformer

The transformer is an artificial neural network architecture based on the multi-head attention mechanism, in which text is translated to numerical representations known as tokens, and each token is converted into a vector via lookup from a word embedding table. At each layer, each token is contextualized with other (unmasked) tokens within the context window using a concurrent multi-head attention method, allowing the signal for key tokens to be increased while less significant tokens are diminished.

Transformers have the advantage of having no recurrent units, therefore they require less training time than previous recurrent neural architectures (RNNs) like long short-term memory (LSTM).All baseline models are trained under identical experimental conditions, utilizing uniform input characteristics, window widths, and training methodologies. This makes sure that differences in performance may be explained by the design of the architecture instead of bias in the experiments.

5. Results

This section talks about and looks at the experimental results from the Air Quality UCI dataset that were gotten using the suggested Hybrid Transformer-BiLSTM-Attention framework. It evaluates the model's predictive accuracy, temporal simulation capability, multivariate prediction accuracy, training process stability, and comprehensibility. Using a complete collection of graphs improves the analysis and gives a better understanding of the model's operation. The efficacy and dependability of the suggested approach for simulating intricate air quality dynamics are precisely assessed by this quantitative and visual study.

5.1 Prediction Performance

The results demonstrate that the tested devices had very different levels of performance. The BiLSTM and Transformer can see general trends in air quality in an Italian city. Data were recorded from March 2004 to February 2005 , but they aren't very good at making predictions. Figure 3 shows that the proposed Hybrid Transformer-BiLSTM-Attention model always does better than other models when it comes to MAE, RMSE, and R^2. Table 4 presents an evaluation of the performance of the proposed method compared to the baseline models. The Comparative result with previous studies is illustrated in Table 5.

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

Table 4: Model Performance Comparison on the Air Quality Dataset

Model	MAE	RMSE	R^2
BiLSTM	0.1116	0.1426	0.2884
Transformer	0.1041	0.1347	0.1270
Our Proposed model	0.0589	0.0799	0.9621

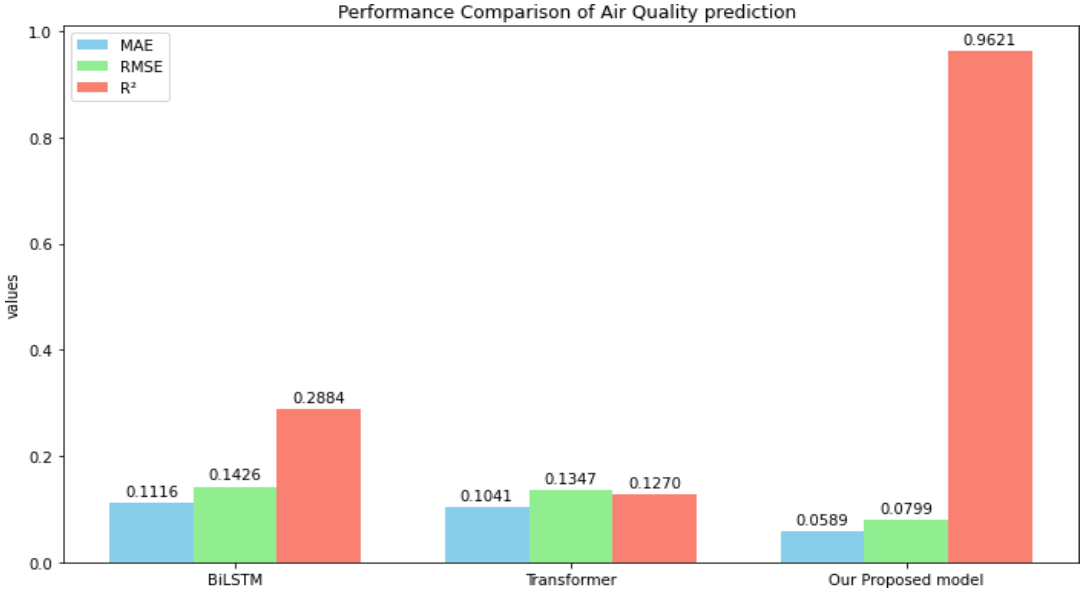


Figure 3. Model Performance Comparison on the Air Quality prediction

Table 5: Comparative result with previous studies

Reference	Model	Dataset	City / Region	MAE	RMSE	R^2
[21]	LT-Hybrid (LSTM–Transformer)	Urban air quality dataset (5000 records)	Not specified	—	0.1021	0.9382
[14]	CNN–BiLSTM–Transformer	Integrated meteorological & pollutant dataset	Qingdao City, China	4.0220	5.4236	0.95
Our proposed model	Transformer–BiLSTM–Attention (Proposed)	UCI Air Quality Dataset, 9,358 hourly records, 15 features	Italian city	0.0589	0.0799	0.9621

5.2 Temporal Prediction Accuracy

The time-series data in Figure 4 reveal significant variations in the model's behavior. Large patterns can be seen by BiLSTM and Transformer models, but they struggle with rapid shifts and often smooth out sudden peaks in pollution. In contrast, the suggested hybrid model closely resembles actual CO(GT) changes. It responds quickly to abrupt changes and remains steady throughout periods of minimal change. Predictions regarding air quality are more accurate and dependable when comprehensive attention and bidirectional temporal learning are combined.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

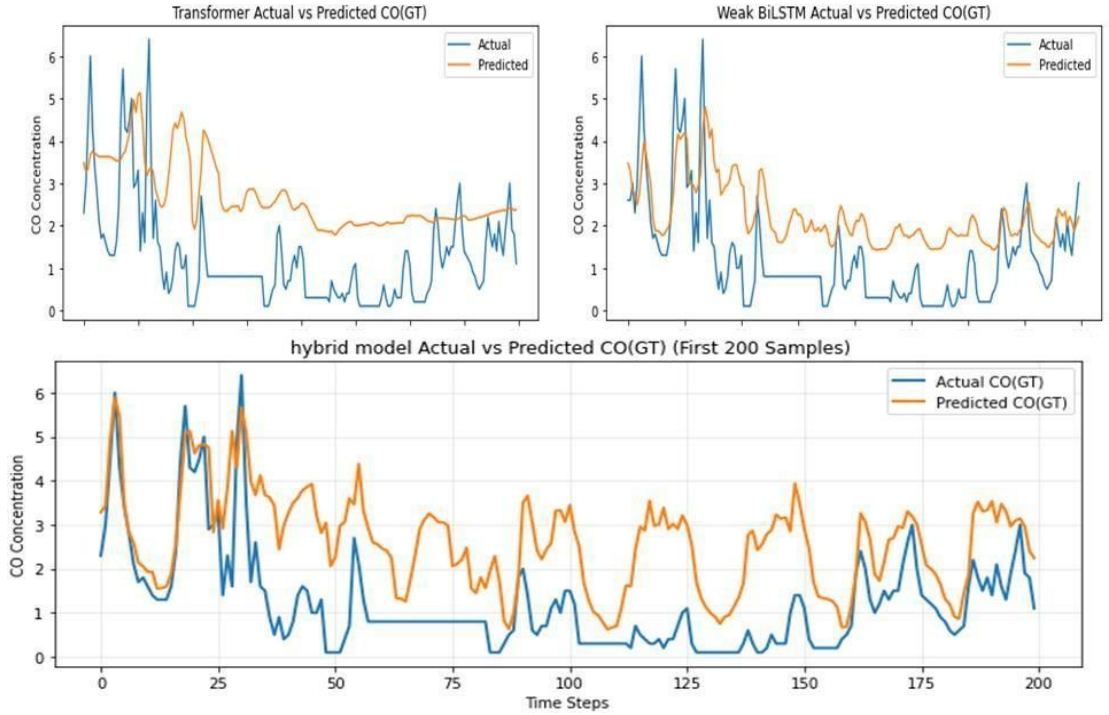


Figure 4. A comparison between the actual and expected CO(GT) values for models

5.3 Training Convergence and Stability

The training and validation loss for three models as illustrated in Figure 5. The training and validation loss curves show a smooth and consistent convergence throughout the training period. The convergence of these curves indicates that the model is well-optimized and does not suffer from over-allocation. This behavior demonstrates the ability of the proposed hybrid architecture to generate useful representations that efficiently match current data, compared to baseline models.

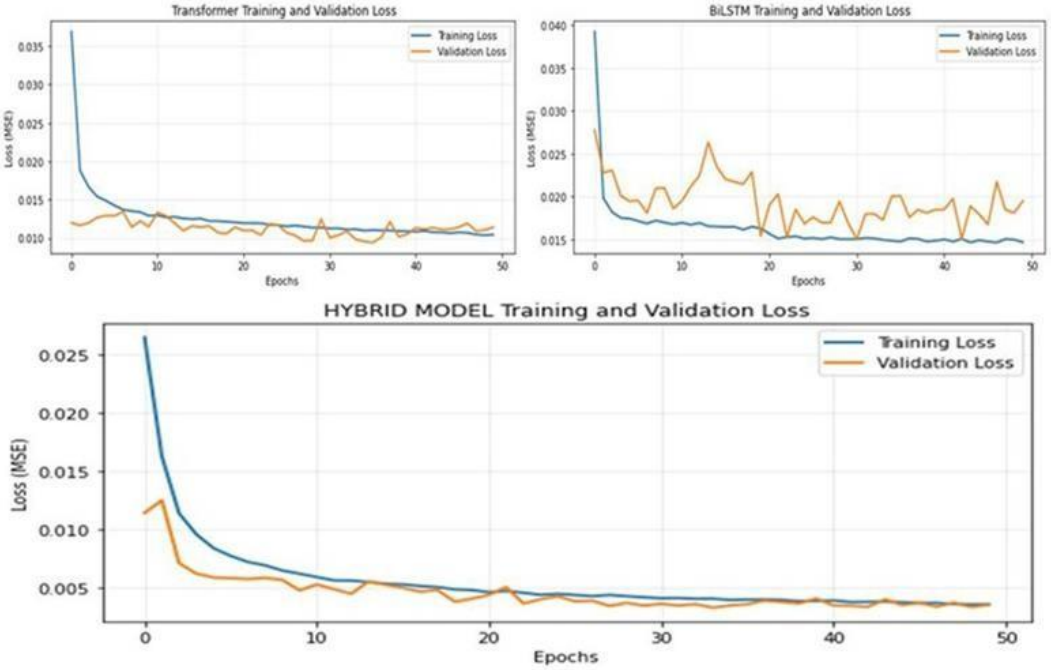


Figure 5. Training and validation loss for models

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

5.4 Multivariate Prediction and Generalization

The proposed model can accurately forecast multiple air quality variables simultaneously, including carbon monoxide, ethane, nitrogen oxides, nitrogen dioxide, and significant weather elements, as shown by the actual versus predicted time-series graphs for a number of contaminants. For several contaminants, Figure 6 illustrates how the anticipated trajectories closely match the observed trends in terms of both relative size and time structure. The model's capacity to predict more than one thing at a time is demonstrated by its consistent performance across several factors. Additionally, it works well in real-world situations where assessing air quality across several factors is necessary.

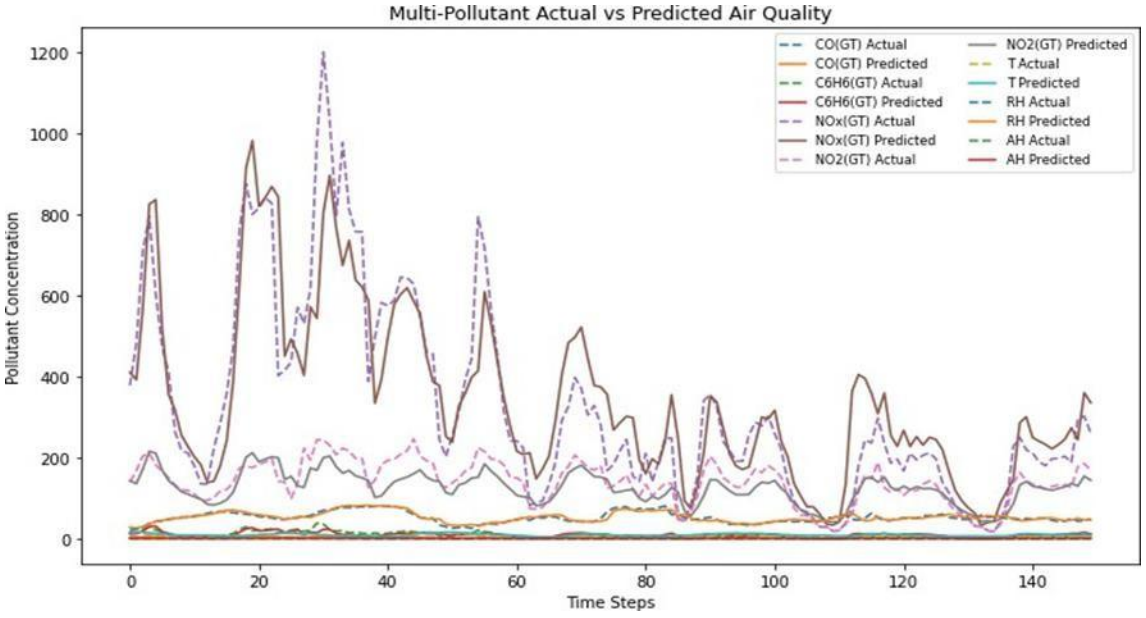


Figure 6. Actual and predicted values of multiple air quality parameters across time

5.5 Scatter Analysis of Observed and Predicted

The dispersion plots for the actual and expected concentrations of carbon monoxide (CO), ethane (C6H6), nitrogen oxides (NOx), and nitrogen dioxide (NO2) are displayed in Figure 7.All of the contaminants have strong linear connections, as these charts show. The majority of sites are situated near the diagonal reference line, suggesting that there is little systematic bias and that the forecasts are very consistent. These findings are supported by the coefficients of determination (R^2). The explanatory power of all pollutants ranged from moderate to strong, with nitrogen oxides (NOx) having the highest R2 value. The greater dispersion observed at high concentration levels is due to the intrinsic variability in the occurrence of large pollutants, rather than to limitations in model stability.

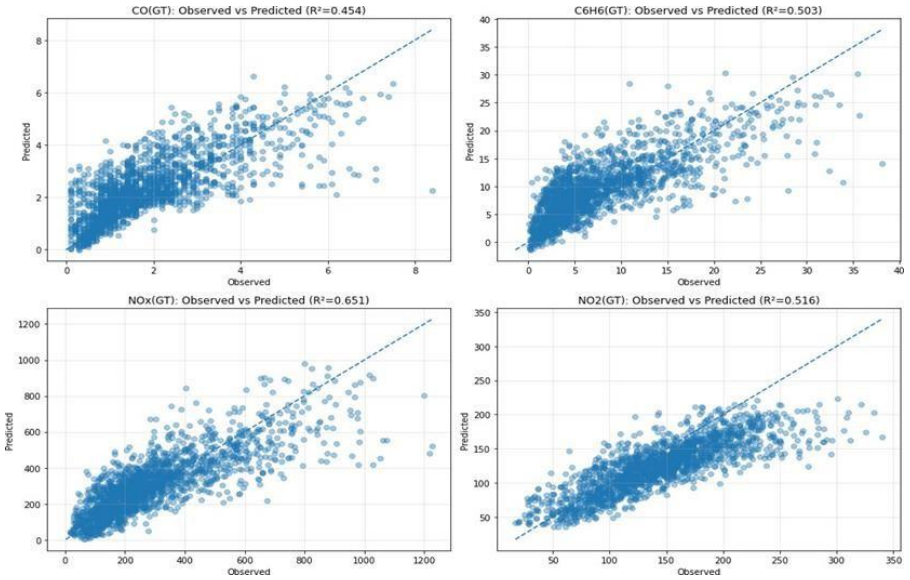


Figure 7. Relationship between observed and expected values for several air quality parameters

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

5.6 Error Distribution and Robustness

The characteristics of pollutants include abnormal distributions with positive skewness and some outliers. This indicates that air pollution occurs in short bursts. In contrast, climate variables exhibit more stable and symmetrical patterns. These characteristics highlight the inherent variability and uncertainty in air quality data, underscoring the feasibility of applying a flexible, nonlinear hybrid deep learning model to obtain consistent and reliable predictions. Figure 8 shows how air quality and climate variables are spread out.

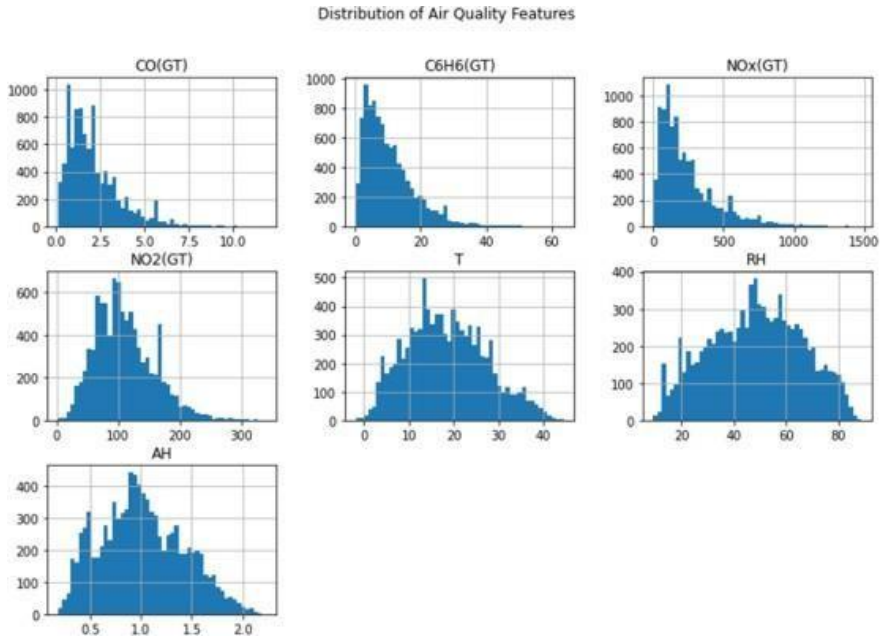


Figure 8. Histograms of the distribution of air quality features

5.7 Feature Relationships and Data Characteristics

The correlation heatmap indicates that important air pollutants like CO, NOx , and NO2 are very dependent on each other. It also illustrates that there are strong links between weather variables and pollutant concentrations. These results provide strong support for employing a multivariate learning technique and elucidate the reasons attention-based feature modeling enhances predictive accuracy. The feature distribution plots also demonstrate that some pollutant variables have skewed and non-Gaussian patterns. This shows how important it is to have proposed hybrid deep learning model that can accurately represent difficult and diverse data distributions, as seen in Figure 9.

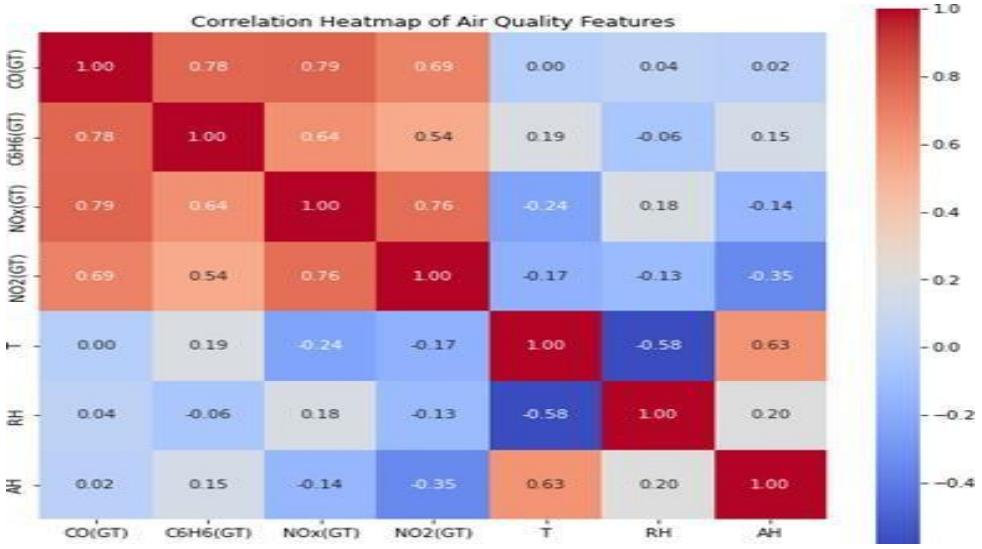


Figure 9. A heatmap showing how air quality features are related to each other

Academia Open
Vol. 11 No. 1 (2026): June
DOI: 10.21070/acopen.11.2026.13837

5.8 Changes over time in air quality contaminants and weather variables

Pollutants have sudden changes and surprising peaks, whereas temperature and humidity follow more stable, steady patterns. This shows that we need models that can capture both short-term changes and long-term trends. Figure 10 depicts how air pollution levels and weather conditions change over time.

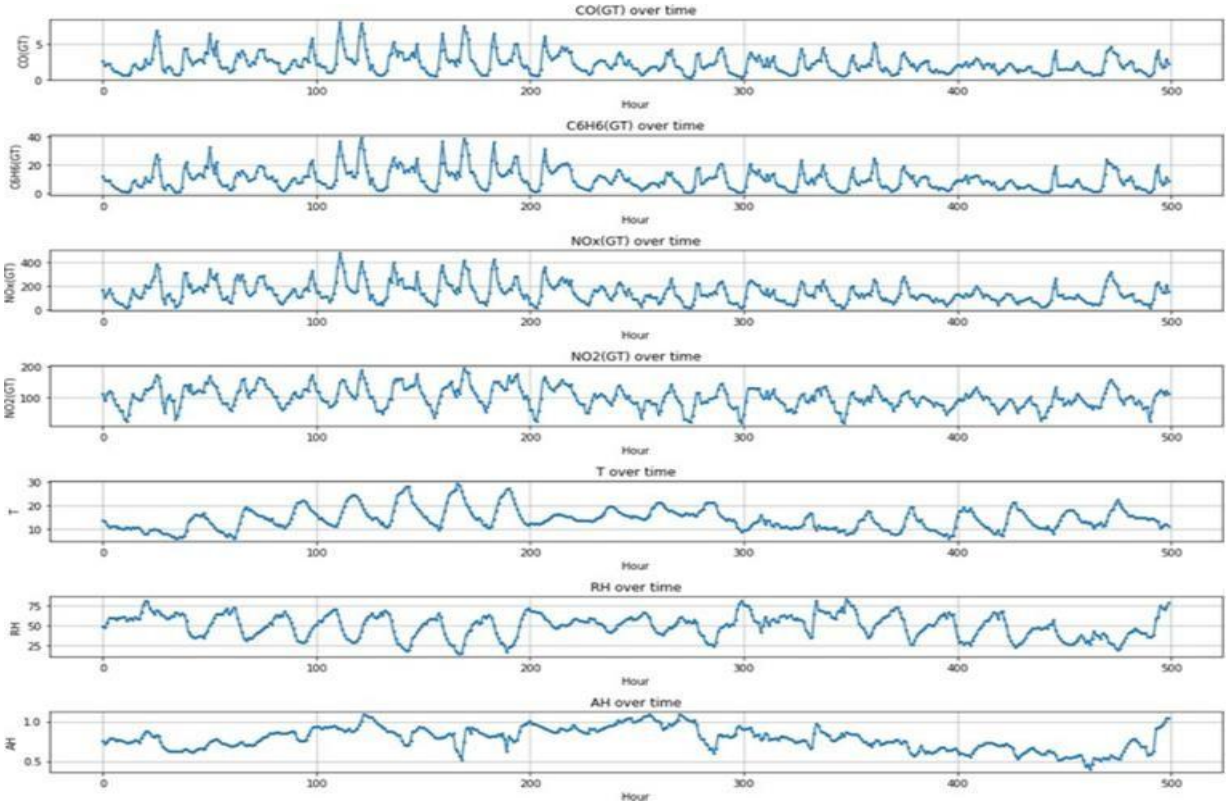


Figure 10. Air pollution and climate variables in time series

6. Discussion

We use three standard automatic evaluation metrics, namely RMSE, MAE, and R^2, to assess our approach. We tested air quality using our proposed model and baseline models. The experimental results show that the proposed hybrid model Transformer–BiLSTM–Attention, clearly outperforms individual models such as BiLSTM and Transformer in terms of accuracy, stability, and ability to handle sudden fluctuations. It also flexibly handles the inherent temporal complexity of real-world air quality data. The Transformer layer captures long-term changes, while the BiLSTM layer helps understand the fine temporal details in both directions, giving the model a better understanding of the overall data context. Graphical results, such as time series comparisons, dispersion analysis, loss curves, residual distributions, and correlation maps, support the model's reliability and demonstrate its consistent performance. This study presents an advanced framework for air quality forecasting. Therefore, these two deep learning models, along with our proposed model, are used as baselines for comparing and evaluating our methodology. Furthermore, we compared the proposed model with previous studies and it demonstrated its superiority and high performance. There are some limitations in this work. These aspects do not diminish the value of the results; rather, they open the door to future studies that can strengthen and develop the model. The main shortcomings of the proposed model are as follows:

- The model's generalizability is diminished when it is dependent on a station with a restricted geographic range.
- Extreme weather conditions or seasonal variations might not be fully reflected in a short time frame.

7. Conclusion and Future work

This research introduces a hybrid deep-learning model to enable more accurate predictions of multiple variable air quality utilizing state-of-the-art hybrid deep learning techniques. pointing out its shortcomings and paving the way for advancements in this field. Our approach is able to effectively and uniquely capture the complex temporal dependencies and nonlinear interactions present in real-world air quality data. The experimental results conducted using the Air Quality UCI dataset demonstrate that the hybrid framework performs better than the baseline models based on prediction accuracy, training stability, and robustness from earlier research on air quality data. By emphasizing significant historical time points,

the incorporation of attention mechanisms facilitates comprehension of the model. Applications that monitor the environment and assist users in making decisions would greatly benefit from this. In order to give spatial and temporal predictions, future research could expand on this paradigm by combining geographic data from numerous monitoring sites. Potential enhancements include real-time deployment scenarios, multi-stage forecasting, and transfer learning between cities. Forecast accuracy would be improved, and the development of comprehensive environmental early warning systems would be advanced by including external factors such as traffic flow data, industrial emissions inventories, and urban energy consumption patterns.

References

[1] D. Bhardwaj and P. R. Ragiri, "A deep learning approach to enhance air quality prediction: Comparative analysis of LSTM, LSTM with attention mechanism, and BiLSTM," in *Proc. 2024 IEEE Region 10 Symp. (TENSYMP)*, Sep. 2024, pp. 1–8.

[2] L. Zhang, P. Liu, L. Zhao, G. Wang, W. Zhang, and J. Liu, "Air quality predictions with a semi-supervised bidirectional LSTM neural network," *Atmospheric Pollution Research*, vol. 12, no. 1, pp. 328–339, 2021.

[3] S. Liu and Y. Hu, "Air quality prediction based on factor analysis combined with Transformer and CNN-BiLSTM-attention models," *Scientific Reports*, vol. 15, no. 1, Art. no. 20014, 2025.

[4] L. N. T. My, V. Nguyen, and T. Vo, "An efficient denoising Transformer-based architecture for long-ranged time-series air quality prediction," *Concurrency and Computation: Practice and Experience*, vol. 37, nos. 27–28, Art. no. e70450, 2025.

[5] W. Zhao, Q. Zhang, T. Shu, and X. Du, "AirTrace-SA: Air pollution tracing for source attribution," *Information*, vol. 16, no. 7, Art. no. 603, 2025.

[6] C. Tang, D. Zhen, L. Zhang, F. Zhao, and Y. Wei, "ST-OzoneNet: A spatio-temporal deep learning model for accurate prediction of ozone concentration," *Earth Science Informatics*, vol. 18, no. 4, pp. 1–22, 2025.

[7] M. Yu, A. Masrur, and C. Blaszczak-Boxe, "Predicting hourly PM2.5 concentrations in wildfire-prone areas using a spatiotemporal Transformer model," *Science of the Total Environment*, vol. 860, Art. no. 160446, 2023.

[8] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Deep air quality forecasting using hybrid deep learning framework," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2412–2424, 2019.

[9] J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, "Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM," *IEEE Access*, vol. 7, pp. 107897–107907, 2019.

[10] R. Sivasubramanian, R. C. N., and S. Kumar, "Attention-based spatio-temporal graph neural network for multi-pollutant urban air quality prediction," in *Proc. 2025 3rd Int. Conf. Data Science and Network Security (ICDSNS)*, Jul. 2025, pp. 1–5.

[11] W. Cao, W. Qi, and P. Lu, "Air quality prediction based on time series decomposition and convolutional sparse self-attention mechanism Transformer model," *IEEE Access*, 2024.

[12] A. Xu and X. Jiang, "Construction of deep learning air quality forecasting system with multi-source data fusion," in *Proc. 2023 3rd Int. Conf. Electronic Information Engineering and Computer Communication (EIECC)*, Dec. 2023, pp. 184–188.

[13] J. Zhang, Z. Luo, and Z. Yang, "Research on air quality prediction based on LSTM-Transformer with adaptive temporal attention mechanism," in *Proc. 2023 2nd Int. Conf. Artificial Intelligence and Intelligent Information Processing (AIIIP)*, Oct. 2023, pp. 320–323.

[14] Z. He, Q. Guo, Z. Zhang, G. Feng, S. Qiao, and Z. Wang, "Forecasting daily ambient PM2.5 concentrations in Qingdao City using deep learning and hybrid interpretable models and analysis of driving factors using SHAP," *Toxics*, vol. 14, no. 1, Art. no. 44, 2025.

[15] R. Zou *et al.*, "PD-LL-Transformer: An hourly PM2.5 forecasting method over the Yangtze River Delta urban agglomeration, China," *Remote Sensing*, vol. 16, no. 11, Art. no. 1915, 2024.

[16] Z. Wang, K. Jia, W. Zhang, and C. Zhang, "PM2.5 concentration prediction in the cities of China using multi-scale feature learning networks and Transformer framework," *Sustainability*, vol. 17, no. 19, Art. no. 8891, 2025.

[17] H. Chen, M. Guan, and H. Li, "Air quality prediction based on integrated dual LSTM model," *IEEE Access*, vol. 9, pp. 93285–93297, 2021.

[18] X. Mo, H. Li, and L. Zhang, "Design of a regional and multistep air quality forecast model based on deep learning and domain knowledge," *Frontiers in Earth Science*, vol. 10, Art. no. 995843, 2022.

[19] A. T. Nguyen, D. H. Pham, B. L. Oo, Y. Ahn, and B. T. Lim, "Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization," *Journal of Big Data*, vol. 11, no. 1, Art. no. 71, 2024.

[20] C. Wang and M. Zhu, "A novel hybrid model based on deep learning and autoregressive for air quality prediction," in *Proc. 2023 Int. Joint Conf. Neural Networks (IJCNN)*, Jun. 2023, pp. 1–7.

[21] Y. Liu, M. Tee, L. Lu, F. Zhou, and B. Lu, "High-precision urban air quality prediction using a LSTM-Transformer hybrid architecture," *Int. J. Advanced*

Computer Science and Applications, vol. 16, no. 4, 2025.

[22] UCI Machine Learning Repository, "Air Quality Dataset." [Online]. Available: https://archive.ics.uci.edu/dataset/360/air+quality

[23] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On-field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[24] M. Alabadla *et al*., "Systematic review of using machine learning in imputing missing values," *IEEE Access*, vol. 10, pp. 44483–44502, 2022.

[25] G. Narkhede, M. Poonawala, A. Sonawane, A. Hiwale, and A. R. Singh, "Air pollution prediction with advanced preprocessing and deep ensemble learning," *Atmospheric Pollution Research*, Art. no. 102610, 2025.

[26] A. Shoari Nejad, *Statistical and Machine Learning Models for Multivariate Sensor Data with Application to Environmental Monitoring*, Ph.D. dissertation, National Univ. of Ireland Maynooth, 2023.

[27] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: A survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575–1637, 2024.

[28] Q. Wen *et al*., "Transformers in time series: A survey," *arXiv preprint arXiv:2202.07125*, 2022.

[29] N. Hassan, A. S. M. Miah, and J. Shin, "A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition," *Applied Sciences*, vol. 14, no. 2, Art. no. 603, 2024.

[30] D. Soydaner, "Attention mechanism in neural networks: Where it comes from and where it goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13371–13385, 2022.

[31] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Information Sciences*, vol. 585, pp. 609–629, 2022.

[32] S. Saha, H. S. Makkar, V. B. Sukumaran, and C. R. Murthy, "On the relationship between mean absolute error and age of incorrect information in the estimation of a piecewise linear signal over noisy channels," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2576–2580, Nov. 2022.

[33] M. Berggren, "Coefficients of determination measured on the same scale as the outcome: Alternatives to R^2 that use standard deviations instead of explained variance," *Psychological Methods*, 2024.