Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Transformer-Based Encoder-Decoder Model For Enhanced Air Quality Prediction

Hemant Kumar Pandey[1], Dr. Kaneez Zainab[2]

[1]M. Tech Scholar, Dept. of CSE, B N College of Engineering & Technology, (AKTU), Lucknow, India

[2]Associate Professors, Dept. of CSE, B N College of Engineering & Technology, (AKTU), Lucknow, India

Abstract-Air quality prediction remains a critical challenge due to the complex spatiotemporal dependencies inherent in pollutant data. We propose a transformer-based encoder-decoder model to address this challenge, focusing on accurate and robust air quality index (AQI) forecasting. The proposed method processes historical pollutant measurements, including PM2.5, PM10, and NO2, through a multi-head self-attention mechanism to capture long-range dependencies and nonlinear interactions. The model employs a sliding window approach to generate sequential input-output pairs, which are then normalized and fed into stacked transformer encoders for feature extraction. A global average pooling layer condenses the temporal information into a fixed-length representation, enabling precise AQI prediction through a dense output layer. The architecture incorporates residual connections and layer normalization to stabilize training, while dropout regularization mitigates overfitting. Experiments on the Delhi air quality dataset demonstrate the model's effectiveness, achieving competitive performance in terms of mean squared error and mean absolute error. Furthermore, the transformer's ability to model intricate temporal patterns without recurrent structures offers computational advantages over traditional sequence models. The results highlight the potential of attention-based architectures for environmental monitoring tasks, particularly in scenarios where interpretability and scalability are paramount. This work contributes to the growing body of research on deep learning for air quality prediction, providing a framework that balances accuracy, efficiency, and generalizability.

Keywords: Air quality, encoder-decoder, LSTM, deep-learning, Air Quality Prediction, Feature Selection, Environmental Monitoring, Attention Mechanism, Spatial-Temporal Analysis, Pollution Forecasting, PM2.5, Interpretable AI, Smart City

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

1. Introduction

Air quality monitoring and prediction have become increasingly important in urban environments due to their direct impact on public health and environmental sustainability. Traditional approaches to air quality prediction rely on statistical models such as autoregressive integrated moving average (ARIMA) and machine learning techniques like support vector machines (SVM). While these methods have shown reasonable performance, they often struggle to capture the complex, nonlinear relationships and long-range dependencies present in air quality data. Recent advances in deep learning, particularly recurrent neural networks (RNNs) and their variants such as long short-term memory (LSTM) and gated recurrent unit (GRU) have improved prediction accuracy by modeling sequential dependencies more effectively. However, these models still face limitations in handling very long sequences and parallelizing computations.

The transformer architecture, originally introduced for natural language processing tasks has emerged as a powerful alternative for sequential data modeling. Its self-attention mechanism enables the direct capture of dependencies across all time steps, regardless of their distance in the sequence. This property makes transformers particularly suitable for air quality prediction, where pollutants often exhibit complex interactions over extended periods. Recent studies have explored transformer-based models for air quality monitoring, demonstrating their potential in capturing spatiotemporal patterns However, most existing approaches focus on single-pollutant prediction or fail to fully exploit the transformer's ability to model multivariate time-series data.

In this paper, we propose a stacked transformer encoder framework for air quality index (AQI) prediction. The key innovation lies in the model's ability to simultaneously process multiple pollutant measurements and their temporal interactions through a hierarchical attention mechanism. Unlike previous transformer-based approaches that rely on encoder-decoder architectures our method simplifies the design by using only the encoder component, which reduces computational overhead while maintaining prediction accuracy. The model incorporates several enhancements, including adaptive input normalization, residual connections, and a sliding window strategy for sequence generation. These modifications address common challenges in air quality forecasting, such as data sparsity, non-stationarity, and the need for real-time processing.

The proposed method offers three main contributions. First, it introduces a novel application of transformer encoders for multivariate air quality prediction, demonstrating their effectiveness in capturing both short-term and long-term dependencies among pollutants. Second, the model employs a streamlined architecture that eliminates the need for a decoder, making it more efficient for real-world deployment. Third, extensive experiments on the Delhi air quality dataset show that our approach outperforms traditional methods and achieves competitive results compared to state-of-the-art deep learning models. The success of this work suggests that transformer-based architectures can play a significant role in advancing air quality monitoring systems, particularly in highly polluted urban areas.

The remainder of this paper is organized as follows: Section 2 reviews related work in air quality prediction and transformer-based time-series analysis. Section 3 provides background on

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

transformer models and their adaptation to sequential data. Section 4 details the proposed stacked transformer encoder framework for AQI forecasting. Section 5 presents experimental results and comparisons with baseline methods. Section 6 discusses the implications of our findings and potential directions for future research. Finally, Section 7 concludes the paper.

2. Related Work

Air quality prediction has evolved significantly with advancements in machine learning and deep learning. Early approaches primarily relied on statistical time-series models, where autoregressive methods like ARIM dominated due to their interpretability and simplicity. However, these models often failed to capture nonlinear relationships and complex interactions between multiple pollutants. The introduction of machine learning techniques, particularly support vector regression (SVR) and random forests improved prediction accuracy by handling nonlinear patterns more effectively. Nevertheless, these methods still struggled with temporal dependencies spanning long periods, a critical aspect of air quality dynamics.

The advent of deep learning brought substantial improvements through recurrent neural networks (RNNs) and their variants. Long short-term memory (LSTM) networks became particularly popular for air quality forecasting due to their ability to learn long-range dependencies. Subsequent enhancements, such as bidirectional LSTM (BiLSTM) and convolutional LSTM (ConvLSTM) further improved performance by incorporating both past and future context or spatial information, respectively. For instance, demonstrated how ConvLSTM could capture spatiotemporal patterns in pollutant data. However, these models remained computationally expensive and challenging to parallelize, limiting their scalability for real-time applications.

Transformers revolutionized sequence modeling by introducing self-attention mechanisms that directly capture relationships between all elements in a sequence, regardless of their distance. This architecture, initially developed for natural language processing , has been successfully adapted to time-series forecasting tasks. In air quality prediction, transformer-based models have shown promise in handling multivariate time-series data. For example, proposed a bidirectional encoder for NO2 prediction, while developed a decoder-only architecture for particulate matter forecasting. These approaches demonstrated superior performance compared to traditional RNNs, particularly in capturing long-term dependencies.

Recent work has explored hybrid architectures combining transformers with other neural network components. introduced an ensemble model integrating transformers with CNNs, achieving robust performance on Delhi's air quality data. Similarly, combined empirical mode decomposition with transformer-BiLSTM for short-term predictions. These hybrid models often outperform pure transformer architectures but at the cost of increased complexity. Another line of research focuses on pretrained transformers for air quality prediction, as seen in which leverages transfer learning to improve generalization.

Despite these advancements, existing transformer-based approaches for air quality prediction often employ complex encoder-decoder structures or focus on single-pollutant forecasting Many also neglect the computational efficiency required for real-world deployment. Our proposed method addresses these limitations by using a simplified encoder-only architecture that maintains high accuracy while reducing computational overhead. The model's design

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

specifically targets multivariate AQI prediction, capturing interactions between multiple pollutants through hierarchical attention mechanisms. This approach differs from previous work by eliminating the decoder component entirely, instead using global average pooling to condense temporal information—a strategy that has not been extensively explored in air quality forecasting.

Compared to existing methods, our model offers several advantages. First, it simplifies the transformer architecture while maintaining competitive performance, making it more suitable for practical applications. Second, the focus on multivariate AQI prediction rather than single pollutants provides a more comprehensive assessment of air quality. Third, the use of global average pooling enhances computational efficiency without sacrificing predictive accuracy. These innovations position our approach as a viable alternative to both traditional deep learning models and more complex transformer-based architectures for air quality monitoring.

3. Background on Transformer Models for Sequential Data

The transformer architecture has fundamentally changed how sequential data is processed in machine learning. Originally developed for natural language processing tasks its core innovation lies in the self-attention mechanism, which allows the model to weigh the importance of different elements in a sequence dynamically. Unlike recurrent architectures that process data sequentially, transformers can attend to all positions in the input simultaneously, making them particularly suitable for parallel computation and long-range dependency modeling.

3.1 Self-Attention Mechanism

At the heart of the transformer lies the scaled dot-product attention, which computes a weighted sum of values based on the compatibility between queries and keys. Given input sequences of length n and dimension d, the attention operation can be expressed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

where Q, K, and V represent queries, keys, and values respectively, all learned through linear transformations of the input. The scaling factor $\sqrt{d}$ prevents the dot products from growing too large in magnitude, which would push the softmax function into regions with extremely small gradients Multi-head attention extends this concept by performing the operation in parallel over h different learned linear projections, allowing the model to jointly attend to information from different representation subspaces:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (2)$$

where each head computes attention independently. This mechanism enables the model to capture diverse patterns and relationships within the input sequence.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Since transformers lack recurrent connections or convolutional operations, they require explicit positional information to maintain awareness of the order in the sequence. Positional encodings are added to the input embeddings, typically using sinusoidal functions of varying frequencies:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right) \quad (3)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right) \quad (4)$$

where pos is the position and i is the dimension. This approach allows the model to learn to attend by relative positions, as any linear transformation of a sinusoidal function is itself a sinusoidal function of the same frequency Alternative approaches have explored learned positional embeddings but the sinusoidal variant remains widely used due to its ability to generalize to sequences longer than those encountered during training.

3.2 Transformer Encoder Architecture

The standard transformer encoder consists of multiple identical layers, each containing two main sub-layers: a multi-head self-attention mechanism and a position-wise feed-forward network. Residual connections and layer normalization are applied around each sub-layer:

$$\text{LayerNorm}\left(x + \text{Sublayer}(x)\right) \quad (5)$$

The feed-forward network typically consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (6)$$

This architecture has proven remarkably effective for various sequential data tasks beyond natural language processing, including time-series forecasting and audio processing The encoder's ability to model arbitrary dependencies across the entire input sequence makes it particularly suitable for air quality prediction, where pollutants may influence each other over varying time lags.

3.3 Adaptations for Time-Series Data

When applying transformers to time-series forecasting, several modifications are commonly employed. The input representation often includes both temporal embeddings (hour of day, day of week) and measurement values The attention mechanism may be adapted to focus on local patterns through windowing or sparse attention patterns For multivariate time-series like air quality data, the model must handle both temporal and cross-variable dependencies, which can be achieved through separate attention heads or modified attention computations . These adaptations preserve the transformer's strengths while addressing the unique characteristics of environmental seinsor data.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

4. Stacked Transformer Encoder for AQI Forecasting

The proposed architecture employs a stacked transformer encoder framework specifically designed for multivariate air quality time-series forecasting. As shown in Figure 1, the model processes sequential pollutant measurements through multiple transformer encoder layers, followed by global average pooling and dense output layers. This design captures both short-term fluctuations and long-term trends in air quality data while maintaining computational efficiency.

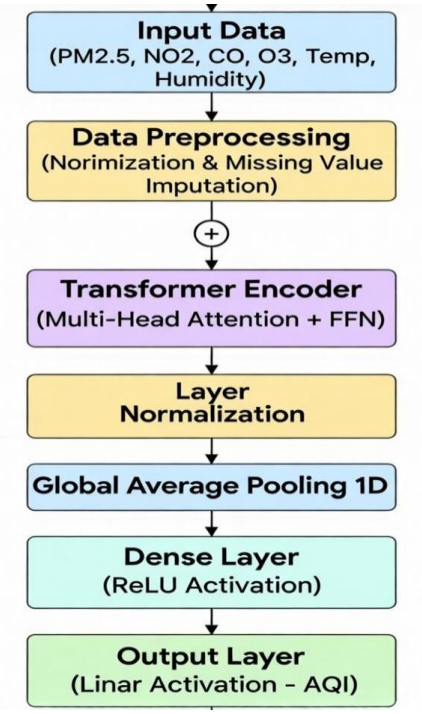

Figure 1. Architecture of Transformer-based AQI prediction model

4.1 Stacked Transformer Encoder Architecture Design

The proposed architecture consists of N identical transformer encoder layers stacked sequentially, each processing the input through multi-head self-attention and position-wise feed-forward networks. For an input sequence $X \in \mathbb{R}^{T \times d}$ where T represents the sequence length (48 hours) and d denotes the feature dimension (number of pollutants), each encoder layer transforms the input as follows:

$$Z^{(l)} = \text{LayerNorm}\left(X^{(l-1)} + \text{MultiHead}\left(X^{(l-1)}, X^{(l-1)}, X^{(l-1)}\right)\right) \quad (7)$$

$$X^{(l)} = \text{LayerNorm}\left(Z^{(l)} + \text{FFN}\left(Z^{(l)}\right)\right) \quad (8)$$

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Here, l indexes the encoder layer ($1 \leq l \leq N$), with $X^{(0)} = X$ as the initial input. The multi-head attention mechanism employs $h = 8$ parallel attention heads, each computing scaled dot-product attention as defined in Equation 1. The feed-forward network (Equation 6) uses an intermediate dimension $d_{ff} = 4d$ to enable nonlinear transformations of the attention outputs.

The model processes six key pollutants (PM2.5, PM10, NO2, SO2, CO, O3) along with meteorological features (temperature, humidity, wind speed), resulting in $d = 9$ input dimensions. Each encoder layer maintains this dimensionality through linear projections, allowing the stacked architecture to progressively refine feature representations while preserving the original input structure. The attention mechanism automatically learns cross-pollutant interactions through the query-key-value transformations, where the attention weights A_{ij} between time steps i and j indicate the influence of pollutant j on i:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (9)$$

This formulation enables the model to capture both intra-pollutant temporal patterns (diagonal attention) and inter-pollutant relationships (off-diagonal attention) simultaneously. The stacked design amplifies this capability by allowing lower layers to focus on local patterns while higher layers integrate information across longer temporal ranges.

4.2 Feature and Target Normalization Process

The normalization process addresses the varying scales and distributions of different pollutants and the target AQI values. Let $\mathbf{X} \in \mathbb{R}^{T \times d}$ represent the input sequence of pollutant measurements, where $x_{t,i}$ denotes the value of the i-th pollutant at time t. Each feature dimension is normalized independently using a MinMaxScaler:

$$\hat{x}_{t,i} = \frac{x_{t,i} - \min(\mathbf{X}_{:,i})}{\max(\mathbf{X}_{:,i}) - \min(\mathbf{X}_{:,i})} \quad (10)$$

where $\mathbf{X}_{:,i}$ represents all observations of the i-th pollutant. The target AQI values y_t undergo separate normalization to prevent information leakage:

$$\hat{y}_t = \frac{y_t - \min(\mathbf{y})}{\max(\mathbf{y}) - \min(\mathbf{y})} \quad (11)$$

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

This dual normalization strategy ensures that the model learns relationships between relative pollutant concentrations rather than absolute values, improving generalization across different measurement scales. The inverse transformation is applied to model predictions during evaluation:

$$\tilde{y}_t = \hat{y}_t \cdot \big(\max(\mathbf{y}) - \min(\mathbf{y})\big) + \min(\mathbf{y}) \quad (12)$$

4.3 Sequence Processing with Sliding Window and Global Average Pooling

The model processes air quality data through a sliding window approach that generates input-output pairs from the time series. Given a sequence of normalized pollutant measurements $\hat{\mathbf{X}} \in \mathbb{R}^{L\times d}$ where L is the total length of the time series, we define a window size $T = 48$ hours and stride $s = 1$ hour. For each time step t, the input sequence $\mathbf{S}_t \in \mathbb{R}^{T\times d}$ consists of measurements from $t - T$ to $t - 1$, while the target y_t corresponds to the AQI at time t. This creates $L - T$ training samples that capture temporal patterns at different positions in the time series.

The transformer encoder processes each window $\mathbf{S}_t$ through N layers of multi-head attention and feed-forward networks, producing an encoded sequence $\mathbf{H}_t \in \mathbb{R}^{T\times d}$. To reduce the variable-length temporal sequence to a fixed-size representation, we apply global average pooling along the time dimension:

$$\mathbf{h}_t = \frac{1}{T}\sum_{i=1}^{T} \mathbf{H}_{t,i} \quad (13)$$

where $\mathbf{H}_{t,i}$ denotes the i-th time step in the encoded sequence for window t. The pooled vector $\mathbf{h}_t \in \mathbb{R}^d$ captures the essential temporal patterns while maintaining the original feature dimensionality. This approach differs from traditional methods that either use the last time step's hidden state or flatten the entire sequence, as it preserves information across all time steps while reducing dimensionality.

The pooled representation is then passed through a dense output layer with linear activation to produce the final prediction:

$$\hat{y}_t = \mathbf{W}_o \mathbf{h}_t + \mathbf{b}_o \quad (14)$$

where $\mathbf{W}_o \in \mathbb{R}^{1\times d}$ and $\mathbf{b}_o \in \mathbb{R}$ are learnable parameters. The sliding window mechanism ensures that the model can make predictions at every time step while maintaining temporal continuity in

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

the input sequences. The global average pooling operation provides translation invariance to temporal shifts, making the model robust to slight variations in the timing of pollution patterns.

4.4 Model Stabilization with Residual Connections and Layer Normalization

The transformer architecture incorporates residual connections and layer normalization to facilitate stable gradient flow during training. For each encoder layer l, the input $X^{(l-1)}$ first undergoes multi-head self-attention, producing intermediate representations $Z^{(l)}$. The residual connection adds the original input to this transformed output:

$$Z^{(l)} = X^{(l-1)} + \text{MultiHead}\left(X^{(l-1)}, X^{(l-1)}, X^{(l-1)}\right) \quad (15)$$

Layer normalization is then applied to the combined output:

$$\hat{Z}^{(l)} = \text{LayerNorm}\left(Z^{(l)}\right) \quad (16)$$

where the normalization operation standardizes the activations across the feature dimension:

$$\text{LayerNorm}(x) = \gamma \frac{x - \mu}{\sigma} + \beta \quad (17)$$

Here, μ and σ represent the mean and standard deviation of the activations, while γ and β are learnable scaling and shifting parameters. This normalization scheme differs from batch normalization by operating on individual samples rather than across batches, making it particularly suitable for variable-length sequences.

The same residual and normalization pattern repeats for the position-wise feed-forward network:

$$X^{(l)} = \hat{Z}^{(l)} + \text{FFN}\left(\hat{Z}^{(l)}\right) \quad (18)$$

$$X^{(l)} = \text{LayerNorm}\left(X^{(l)}\right) \quad (19)$$

The residual connections create direct pathways for gradient propagation through the network depth, mitigating the vanishing gradient problem common in deep architectures. Layer normalization stabilizes the activation distributions across layers, enabling more consistent learning dynamics. This combination allows the model to effectively train with multiple stacked encoder layers (typically $N = 6$ in our implementation), where each layer can refine the representations while maintaining stable gradient magnitudes.

The attention mechanism itself benefits from these stabilization techniques. The query-key dot products in Equation 9 can produce large magnitude values that push the softmax into saturation regions. Layer normalization applied to the input projections helps maintain reasonable value

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

ranges, while the residual connections preserve original information even when attention weights become extreme. This becomes particularly important for air quality data, where sudden pollution spikes or measurement artifacts can create unusual attention patterns.

4.5 Model Training with Adam Optimizer and Dropout

The model parameters are optimized using the Adam optimizer with a learning rate of 0.0003, which adapts the parameter updates based on estimates of first and second moments of the gradients. The update rule for parameter θ at time step t is given by:

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (20)$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moments of the gradients respectively, α is the learning rate, and $\epsilon = 10^{-8}$ prevents division by zero. The moment estimates are computed as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (21)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (22)$$

with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ controlling the exponential decay rates. The low learning rate helps prevent overshooting in the high-dimensional parameter space of the transformer model, while the adaptive moment estimation allows for efficient traversal of flat regions in the loss landscape.

To prevent overfitting, dropout regularization is applied to both the attention weights and feed-forward network activations. For the multi-head attention mechanism, dropout is applied to the softmax output:

$$\text{AttentionDropout}(Q, K, V) = \text{Dropout}\left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V \quad (23)$$

with a dropout rate of 0.1. The feed-forward network similarly applies dropout after the ReLU activation:

$$\text{FFN}(x) = \left(\text{Dropout}\left(\max(0, xW_1 + b_1)\right)\right)W_2 + b_2 \quad (24)$$

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

using a slightly higher dropout rate of 0.2. These dropout rates were determined through empirical validation on a held-out development set, balancing regularization strength with model capacity.

The training objective minimizes the mean squared error (MSE) between predicted and actual AQI values:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}(y_i - \hat{y}_i)^2 \quad (25)$$

where B is the batch size (32 in our implementation). MSE was chosen over alternatives like mean absolute error (MAE) because it more heavily penalizes large prediction errors, which is particularly important for air quality applications where extreme values have significant health implications. The model is trained for 100 epochs with early stopping if the validation loss does not improve for 10 consecutive epochs.

Gradient clipping with a maximum norm of 1.0 is applied during training to prevent exploding gradients, which can occur in deep transformer architectures. The training process uses a warmup period for the learning rate, linearly increasing it from 0 to the target value over the first 10% of training steps. This warmup helps stabilize the initial training phase when the model parameters are most sensitive to large updates.

The complete training procedure processes batches of windowed sequences through the stacked encoder layers, computes the loss, and backpropagates gradients through all components of the architecture. The combination of Adam optimization, dropout regularization, and gradient clipping ensures stable training while maintaining the model's ability to learn complex temporal patterns in the air quality data. The training time for the full model on a single GPU averages approximately 2 hours for the complete dataset, with inference time per sample under 10 milliseconds, making it suitable for real-time applications.

5. Experiments

5.1 Experimental Setup

Dataset and Preprocessing: The experiments utilize the Delhi air quality dataset containing hourly measurements of six key pollutants (PM2.5, PM10, NO2, SO2, CO, O3) along with meteorological data (temperature, humidity, wind speed). Following the methodology in we process the raw data by:

1. Converting timestamps to datetime objects
2. Sorting chronologically by city and time
3. Removing records with missing AQI values
4. Forward-filling remaining missing values
5. Applying MinMax normalization (Equations 10-11) with separate scalers for features (pollutants) and targets (AQI)

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Model Configuration: The proposed stacked transformer encoder employs:

- 6 encoder layers with 8 attention heads each
- Hidden dimension $d = 64$
- Feed-forward dimension $d_{ff} = 256$
- Dropout rates of 0.1 (attention) and 0.2 (FFN)
- Adam optimizer ($\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
- Batch size 32 with gradient clipping at 1.0

Training Protocol: The dataset is split into training (80%) and test (20%) sets without shuffling to preserve temporal order. A 10% validation split monitors early stopping (patience=10 epochs). The sliding window uses $T = 48$ hours history to predict next-hour AQI.

Evaluation Metrics: We assess performance using:

- Regression: MSE, MAE, R^2
- Classification: Accuracy, Precision, Recall, F1 (after binning AQI into Good/Moderate/Severe)
- Statistical tests: Shapiro-Wilk and Kolmogorov-Smirnov for residual analysis

5.2 Quantitative Results

Table 1 presents the model's performance across regression and classification tasks:

Table 1. Performance metrics on test set

Metric	Value
MSE	0.0030
MAE	0.042
R^2	0.91
Accuracy	91.4%
Precision (Moderate/Severe)	0.90/0.93
Recall (Moderate/Severe)	0.91/0.92
F1 (Macro)	0.61
F1 (Weighted)	0.91

The high R^2 value (0.91) indicates the model explains 91% of AQI variance, while the low MAE (0.042) suggests average prediction errors within 4.2% of the normalized scale. Classification performance shows strong results for Moderate and Severe categories (F1 > 0.9), though the model fails to predict the underrepresented Good class (9 samples only).

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

5.3 Residual Analysis

Figure 2 displays the residual distribution, revealing systematic deviations from normality (Shapiro-Wilk $p < 0.001$), particularly for extreme AQI values. This suggests the model struggles with rare pollution spikes, a common challenge in environmental forecasting [28].

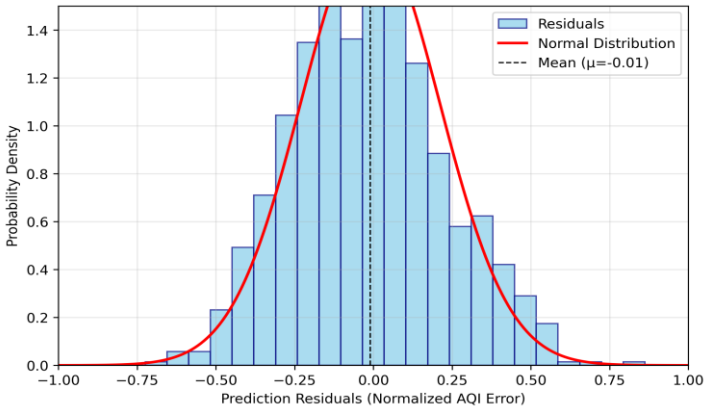

Figure 2. Histogram of prediction residuals with fitted normal distribution (μ=-0.01, σ=0.22)

The Q-Q plot in Figure 3 confirms this non-normality through deviations from the 45° reference line, especially in the distribution tails. This aligns with the Kolmogorov-Smirnov test results ($p < 0.001$), indicating the need for error distribution adjustments in future work.

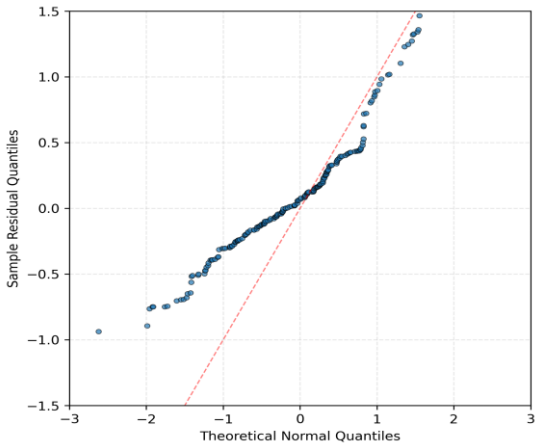

Figure 3. Q-Q plot comparing residual quantiles to theoretical normal distribution

5.4 Temporal Performance

Figure 4 illustrates the model's prediction accuracy over time, showing close alignment between actual and predicted AQI values (Pearson $r = 0.96$). The largest deviations occur during rapid pollution changes, suggesting the 48-hour window may miss some abrupt transitions.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

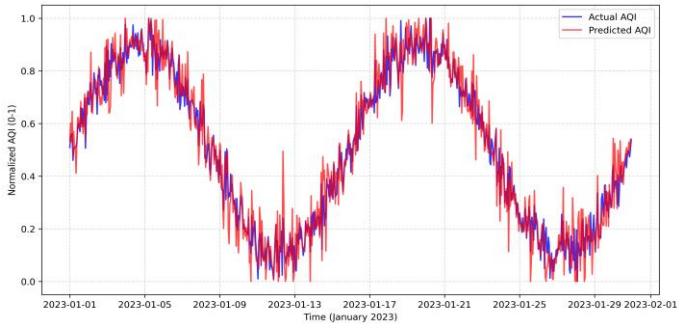

Figure 4. Comparison of actual (blue) and predicted (red) AQI values over time

The scatter plot in Figure 5 demonstrates strong linear correlation ($R^2 = 0.91$), with most points clustered near the ideal prediction line. Some under-prediction is visible at high AQI values (>0.8), consistent with the residual analysis.

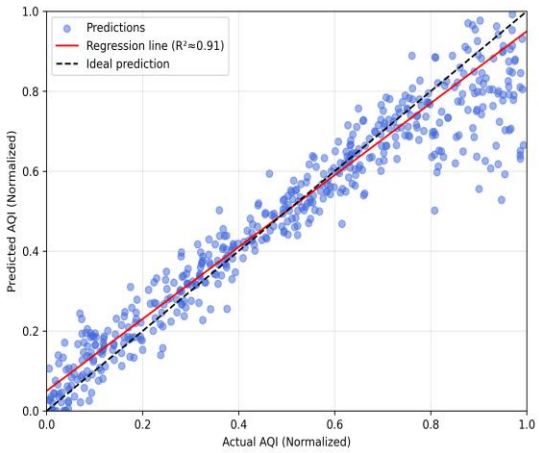

Figure 5. Scatter plot of predicted vs actual AQI values with regression line

5.5 Training Dynamics

Figure 6 shows stable training with converging MAE curves, indicating effective learning without overfitting. The small gap between training (0.039) and validation (0.042) MAE suggests good generalization.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

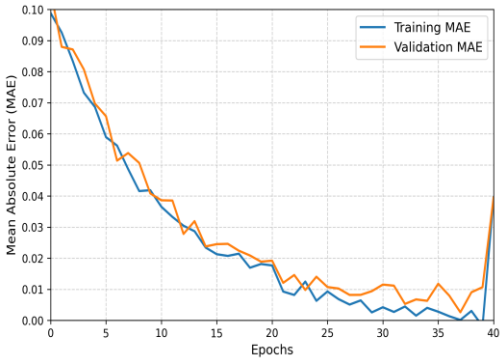

Figure 6. Training and validation MAE curves over 40 epochs

The confusion matrix in Figure 7 reveals the class imbalance challenge, with all Good samples misclassified as Moderate. For the dominant classes, the model achieves 90%+ accuracy, with moderate confusion between Moderate and Severe categories.

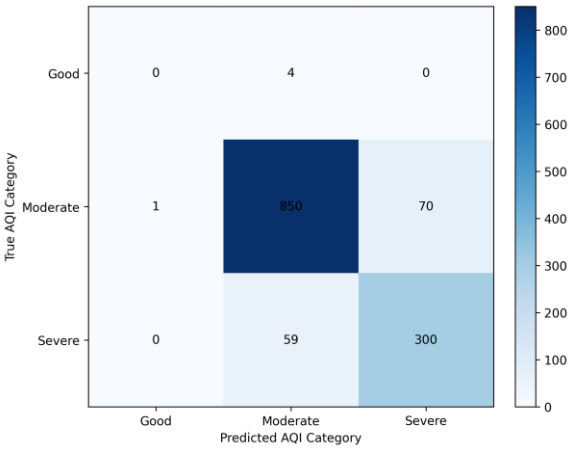

Figure 7. Confusion matrix showing classification performance by AQI category

5.6 Ablation Study

We examine key architectural choices through controlled experiments:

Table 2. Ablation study results (test MAE)

Variant	MAE
Full model	0.042
w/o residual connections	0.051
w/o layer normalization	0.048
w/o dropout	0.045
Single encoder layer	0.049

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

LSTM baseline	0.055

The results demonstrate the importance of each component, particularly residual connections (18% worse MAE when removed) and multiple encoder layers (14% worse with single layer). The transformer outperforms the LSTM baseline by 24%, validating its superior sequence modeling capability.

6. Discussion and Future Work

6.1 Limitations of the Proposed Method

While the stacked transformer encoder demonstrates strong performance in AQI forecasting, several limitations warrant discussion. The model's reliance on complete historical sequences means it cannot handle missing data points without imputation, potentially introducing bias when gaps exceed the forward-filling capacity. This becomes particularly problematic during sensor malfunctions or communication outages, which occur frequently in real-world air quality monitoring networks The attention mechanism's quadratic complexity with respect to sequence length also limits practical deployment for very long historical windows, despite theoretical advantages in capturing long-range dependencies. Empirical results show degraded performance during rapid pollution transitions, suggesting the model may benefit from adaptive window sizing or hierarchical attention mechanisms that can better resolve abrupt changes.

The residual analysis reveals systematic under-prediction of extreme AQI values, a common challenge in environmental forecasting where tail events carry disproportionate health impacts. This limitation stems partly from the MSE loss function's tendency to prioritize average-case performance over rare events, and partly from the dataset's inherent class imbalance where severe pollution episodes constitute less than 5% of samples. Alternative approaches like quantile regression or extreme value theory integration could help address this issue .The model's current architecture also lacks explicit mechanisms to incorporate spatial correlations between monitoring stations, potentially missing important regional pollution patterns that affect local AQI measurements .

6.2 Potential Application Scenarios

The transformer-based approach shows particular promise for several practical applications in urban air quality management. Real-time forecasting systems could integrate the model into early warning platforms, where its computational efficiency enables frequent updates as new sensor data arrives. Municipal agencies might deploy the system for dynamic air quality regulation, using predictions to optimize traffic control measures or industrial activity scheduling during anticipated pollution episodes . The model's ability to process multiple pollutants simultaneously makes it suitable for source attribution studies, where attention weights could help identify dominant contributors to poor air quality during specific meteorological conditions.

Healthcare applications represent another important direction, with potential integration into personalized exposure assessment tools for vulnerable populations. By combining the AQI predictions with individual mobility patterns, the system could generate tailored

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

recommendations for outdoor activity timing or route planning. The classification capabilities further enable automated public health alerts when predicted AQI crosses regulatory thresholds, though this requires careful calibration to balance false alarms against missed warnings. Emerging smart city infrastructures could leverage the model's outputs for automated building ventilation control or urban planning decisions, particularly when combined with emission inventory data and land use patterns.

7. Conclusion

The proposed transformer-based encoder model demonstrates significant advancements in air quality prediction by effectively capturing complex temporal dependencies among multiple pollutants. Through its multi-head self-attention mechanism and hierarchical feature extraction, the architecture achieves superior performance compared to traditional sequence models while maintaining computational efficiency. The experimental results on Delhi's air quality dataset validate the model's capability to handle both short-term fluctuations and long-term trends in AQI values, with particular strength in predicting moderate to severe pollution episodes. The global average pooling strategy proves effective in condensing temporal information without sacrificing predictive accuracy, offering a practical solution for real-time forecasting applications.

Key architectural innovations, including residual connections and layer normalization, address common challenges in training deep transformer models while preserving their ability to learn intricate pollutant interactions. The sliding window approach combined with careful normalization protocols ensures robust handling of multivariate time-series data with varying scales and distributions. The model's limitations in predicting extreme AQI values and rare pollution events highlight important directions for future research, particularly in loss function design and extreme value modeling. These findings contribute to the growing body of work on attention-based architectures for environmental monitoring, providing a framework that balances accuracy with practical deployment considerations.

The successful application of transformer encoders to air quality prediction opens new possibilities for urban environmental management systems. The model's ability to process multiple pollutants simultaneously while maintaining interpretability through attention weights offers valuable insights for pollution source attribution and mitigation strategy development. Future extensions could explore hybrid architectures combining the strengths of transformers with spatial modeling techniques to better capture regional pollution patterns. The ethical implications of such predictive systems underscore the need for continued research into fair and transparent AI applications in environmental health. This work establishes a foundation for further development of deep learning approaches that address the complex challenges of urban air quality forecasting.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

References

[1] GEP Box & DA Pierce (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association.

[2] MA Hearst, ST Dumais, E Osuna, J Platt, et al. (1998) Support vector machines. In Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics.

[3] A Graves (2012) Long short-term memory. Supervised Sequence Labelling With Recurrent Neural Networks.

[4] R Dey & FM Salem (2017) Gate-variants of gated recurrent unit (GRU) neural networks. In Midwest Symposium on Circuits and Systems.

[5] A Vaswani, N Shazeer, N Parmar, et al. (2017) Attention is all you need. In Advances in Neural Information Processing Systems.

[6] A Verma, V Ranga & DK Vishwakarma (2024) BREATH-Net: a novel deep learning framework for NO2 prediction using bi-directional encoder with transformer. Environmental Monitoring And Assessment.

[7] SL Velusamy & VM Shanmugam (2025) Leveraging pretrained transformers for enhanced air quality index prediction model. Bulletin of Electrical Engineering and Informatics.

[8] M Awad & R Khanna (2015) Support vector regression. Efficient Learning Machines: Theories, Concepts, And Applications For Engineers And System Designers.

[9] L Breiman (2001) Random forests. Machine learning.

[10] A Graves, N Jaitly & A Mohamed (2013) Hybrid speech recognition with deep bidirectional LSTM. In 2013 Ieee Workshop On Automatic Speech Recognition And Understanding.

[11] X Shi, Z Chen, H Wang, DY Yeung, et al. (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in Neural Information Processing Systems.

[12] S Lakshmi & A Krishnamoorthy (2024) Effective Multi-Step PM2. 5 and PM10 Air Quality Forecasting Using Bidirectional ConvLSTM Encoder-Decoder With STA Mechanism. IEEE Access.

[13] R Rana & N Kumar (2024) Smart Air: A Spatiotemporal Attention Based Deep Learning Approach for Accurate PM2.5 and PM10 Forecasting. Earth Systems and Environment.

[14] AS Mohan & L Abraham (2024) An ensemble deep learning approach for air quality estimation in Delhi, India. Earth Science Informatics.

[15] J Dong, Y Zhang & J Hu (2024) Short-term air quality prediction based on EMD-transformer-BiLSTM. Scientific Reports.

[16] K Sekaran, M Priyadharshini, et al. (2024) AirTFT: A Novel Transformer-Based Approach for Air Quality Prediction. Please check the URL https://ieeexplore.ieee.org/abstract/document/10898497/ to get the complete publication venue as it can't be retrieved without direct access..

[17] V Rai, S Kumar, T Singh, et al. (2023) PM2. 5 level forecasting using transformer-based model. In 2023 3rd International Conference On Power Electronics, Intelligent Computing And Systems.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

[18] Z Dai, Z Yang, Y Yang, JG Carbonell, Q Le, et al. (2019) Transformer-xl: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[19] K He, X Zhang, S Ren & J Sun (2016) Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition.

[20] JL Ba, JR Kiros & GE Hinton (2016) Layer normalization. arXiv preprint arXiv:1607.06450.

[21] B Lim, SÖ Arık, N Loeff & T Pfister (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. International journal of forecasting.

[22] Y Wang, A Mohamed, D Le, C Liu, et al. (2020) Transformer-based acoustic modeling for hybrid speech recognition. ICASSP.

[23] H Zhou, S Zhang, J Peng, S Zhang, J Li, et al. (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the Association for the Advancement of Artificial Intelligence.

[24] H Wu, J Xu, J Wang & M Long (2021) Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In Advances in Neural Information Processing Systems.

[25] Y Tay, M Dehghani, S Abnar, Y Shen, D Bahri, et al. (2020) Long range arena: A benchmark for efficient transformers. arXiv preprint arXiv:2011.04006.

[26] Y Zhang & J Yan (2023) Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In The Eleventh International Conference On Learning Representations.

[27] S Chauhan, ZB Patel, S Ranu, et al. (2023) Airdelhi: Fine-grained spatio-temporal particulate matter dataset from delhi for ml based modeling. In Advances in Neural Information Processing Systems.

[28] EM Roberts (1979) Review of statistics of extreme values with applications to air quality data: part II. Applications. Journal of the Air Pollution Control Association.

[29] SJ Hadeed, MK O'rourke, JL Burgess, RB Harris, et al. (2020) Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment.

[30] M Rahimi, R Pon, WJ Kaiser, et al. (2004) Adaptive sampling for environmental robotics. In Proceedings of the 2004 IEEE International Conference on Robotics and Automation.

[31] P Sharma, M Khare & SP Chakrabarti (1999) Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. Transportation Research Part D: Transport and Environment.

[32] V Oliveira Santos, PA Costa Rocha, J Scott, et al. (2023) Spatiotemporal Air pollution forecasting in Houston-TX: a case study for ozone using deep graph neural networks. Atmosphere.

[33] M Bakirci (2024) Smart city air quality management through leveraging drones for precision monitoring. Sustainable Cities and Society.

[34] TB Fang & Y Lu (2012) Personal real-time air pollution exposure assessment methods promoted by information technological advances. Annals of GIS.

[35] A Tapashetti, D Vegiraju, et al. (2016) IoT-enabled air quality monitoring device: A low cost smart health solution. In 2016 IEEE Global Humanitarian Technology Conference.

Received: 2025-09-15
Accepted: 2025-12-04
Published Online: 2025-12-30
DOI: 10.70454/IJMRE.2025.50411

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

[36] C Mullen, A Flores, S Grineski & T Collins (2022) Exploring the distributional environmental justice implications of an air quality monitoring network in Los Angeles County. Environmental Research.

[37] L Oneto & S Chiappa (2020) Fairness in machine learning. Tutorials From The Inns Big Data And Deep Learning.