Pol. J. Environ. Stud. Vol. XX, No. X (XXXX), 1-21
DOI: *10.15244/pjoes/203348* ONLINE PUBLICATION DATE:

Original Research

Hybrid Transformer-TimesNet Model for Accurate Prediction of Industrial Air Pollutants: A Case Study of the Xinyang Industrial Zone, China

Siyuan He[1]*, Guanwei Jang[2], Yuhao Liu[1]**

[1]Guangdong University of Science and Technology, Dongguan, China
[2]Shaoguan University, Shaoguan, China

Received: 22 November 2024
Accepted: 25 March 2025

Abstract

Air pollution in industrial zones poses significant threats to environmental and public health, especially in developing regions, highlighting the necessity for accurate forecasting to guide pollution management. This study presents an Optuna-enhanced hybrid Transformer-TimesNet model aimed at improving time series forecasting of six critical pollutants (SO_2, NO_2, CO, O_3, PM_{10}, and $PM_{2.5}$) in Xinyang Industrial Zone, Xiamen, China. Utilizing air quality data from 2019 to 2023, the model combines the Transformer's strength in capturing long-range dependencies with TimesNet's expertise in handling complex temporal patterns. Advanced preprocessing techniques were employed to address both linear and non-linear data components, and Optuna was used for hyperparameter tuning, enhancing model stability and predictive accuracy. Comparative experiments demonstrated the hybrid model's superior performance against traditional statistical methods, machine learning models, and deep learning approaches, evaluated through metrics such as MAE, RMSE, SMAPE, and R^2. The model's capability to accurately capture long-term pollutant trends underscores its reliability and validity as a predictive tool for policymakers and environmental managers. These results contribute to theoretical advancements in environmental monitoring and offer practical solutions for public health protection and pollution mitigation, demonstrating the potential of hybrid deep learning models in addressing complex forecasting challenges.

Keywords: hybrid model, time series forecasting, industrial air pollutants, transformer, TimesNet

Introduction

Air pollution, driven by industrialization, transportation, and rapid economic growth, has become one of the most severe environmental challenges worldwide. The World Health Organization (WHO) reports that air pollution continues to pose a major threat to public health, contributing to millions of premature deaths annually and deteriorating the quality of life in urban areas (WHO, 2021). Industrial zones, in particular, are significant sources of harmful pollutants, including sulfur dioxide (SO_2), nitrogen dioxide (NO_2),

*e-mail: hesiyuan@gdust.edu.cn
**e-mail: steve.jang@sgu.edu.cn

carbon monoxide (CO), ozone (O_3), and particulate matter (PM10, PM2.5), all of which are linked to severe health conditions such as respiratory and cardiovascular diseases (WHO, 2018). In developing countries like China, industrial activities are often concentrated in specific areas, exacerbating local air quality issues (AirVisual, 2020). As urbanization and industrial expansion continue, particularly in emerging economies like China, effective air quality monitoring and pollutant forecasting have become increasingly urgent.

Over the past three decades, China's rapid industrialization has led to severe air pollution in numerous cities [1]. For instance, industrial areas in Xiamen, such as the Xinyang Industrial Zone, are economic hubs but also major sources of pollution. The high pollution levels in Xinyang have raised concerns about the impact on the environment and community health impact. Although stricter emissions controls and industrial regulations have been implemented in recent years, achieving accurate and real-time pollutant forecasting remains essential for effective environmental management.

Accurate forecasting of industrial pollutants is crucial for guiding environmental policies, protecting public health, and reducing pollution. Time series forecasting is a valuable tool for predicting pollutant concentrations, enabling proactive air quality management [2]. However, conventional statistical methods like Autoregressive Integrated Moving Average (ARIMA) models and machine learning models such as Categorical Boosting (CatBoost) often fail to capture the complex temporal dynamics and non-linear characteristics inherent in industrial pollution data [3]. These limitations highlight the need for advanced methods to manage long-sequence data and adapt to rapidly changing environmental conditions.

This study addresses these challenges by proposing a hybrid model that combines the Transformer architecture with the TimesNet framework. The objective is to enhance the accuracy and robustness of time series forecasting for industrial pollutants. By leveraging the strengths of both models, the hybrid Transformer-TimesNet model can effectively capture long-range dependencies and complex temporal patterns in industrial pollution data. The model was applied to data collected from monitoring stations in Xinyang Industrial Zone between 2019 and 2024, covering six key pollutants (SO_2, NO_2, CO, O_3, PM_{10}, and $PM_{2.5}$) monitored on an hourly basis.

The significance of this research lies in its potential to improve the accuracy of environmental time series forecasting, offering a more reliable tool for predicting pollutant concentrations in industrialized regions. This hybrid model can provide policymakers and environmental agencies with actionable insights to mitigate the adverse impacts of air pollution, facilitating more informed decision-making in air quality management and public health protection. The findings of this study contribute to advancing the field of environmental monitoring and offer practical solutions for pollution reduction, with implications for other industrialized regions globally.

The remaining structure of this paper is as follows: Section 2 provides a literature review, Section 3 details the model architecture and methods, Section 4 introduces the dataset and experimental setup, Section 5 discusses the research results, and Section 6 presents conclusions and future research directions.

Literature Review

Accurate and timely pollutant concentration predictions are crucial for environmental protection agencies, policymakers, and urban planners. Time series forecasting techniques have proven valuable for predicting future pollutant levels based on historical data [4]. However, despite significant advancements in this field, many existing models struggle to handle the unique challenges posed by industrial pollutant data, particularly in capturing long-term trends in highly industrialized areas.

Traditional statistical models, widely applied in pollutant concentration prediction, primarily rely on time series analysis and linear regression. A classic time series approach, autoregressive Integrated Moving Average (ARIMA) models are frequently used to capture linear trends and seasonal patterns. While ARIMA is advantageous in terms of simplicity and interpretability, it is less adaptable to industrial pollutants' high variability and non-linear characteristics [5]. Croston's method, which is specialized for intermittent time series forecasting, has shown strong performance in specific industrial scenarios with sporadic pollution patterns, although it is less commonly applied [6].

Other regression models such as Linear Regression (LR), Ridge Regression, Bayesian Ridge, and Elastic Net are also commonly employed in pollutant forecasting. These models are generally suited to low-dimensional, linearly characterized data; however, the non-linear relationships often observed in industrial environments can reduce predictive accuracy [7]. Lasso Regression, Least Angle Regression, and Elastic Net models improve overfitting resistance by incorporating regularization terms. Lasso models, for example, are commonly used in feature selection for pollutant prediction, while Elastic Net, which combines L1 and L2 regularization, is effective in multi-variable, sparse data settings [8]. Orthogonal Matching Pursuit (OMP), a greedy algorithm used for sparse signal recovery, effectively extracts key features from sparse time series data, making it well-suited for certain industrial pollutant scenarios with high-dimensional data [9]. Huber Regression improves robustness by reducing sensitivity to outliers, though it still faces limitations in complex industrial pollutant forecasting [10]. While these models can perform short-term forecasting under relatively stable conditions, they lack flexibility

in handling the non-linear, highly variable conditions typical of industrial pollution [11].

As data dimensions and scale increase, the limitations of traditional statistical models have led researchers to adopt non-linear machine learning methods [12]. Tree-based models, such as Random Forest (RF) and Extra Trees (ET), have demonstrated exceptional performance in handling high-dimensional, non-linear data related to industrial pollutants [13]. These models are particularly effective in managing diverse pollutant characteristics and exhibit strong adaptability in contexts with complex variable interactions. However, their ability to model long-term dependencies is limited. Boosting-based models, including Gradient Boosting Regressor (GBR), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, are widely used for pollutant forecasting. They enhance prediction accuracy by sequentially refining multiple weak learners [14]. LightGBM, which uses a histogram-based decision tree algorithm, performs exceptionally well with high-dimensional and large datasets [15]. CatBoost, designed for categorical feature encoding and balanced gradient handling, is particularly suitable for complex industrial pollutant data with non-linear characteristics [16]. XGBoost, an extension of gradient boosting optimized with second-order derivatives, achieves high speed and accuracy, making it suitable for complex pollutant forecasting tasks [17]. AdaBoost, a classic boosting method, enhances short-term pollutant fluctuation forecasting performance by aggregating weak classifiers [18]. Yet, these models often face challenges when dealing with long sequences of pollutant data and complex temporal dependencies. The k-Nearest Neighbors (KNN) algorithm is advantageous in capturing spatial similarities among pollutant features, though it faces limitations when applied to long time-series data [19].

With their robust feature extraction capabilities, deep learning models have brought breakthroughs in time series forecasting. Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN) are prominent architectures that excel in capturing temporal dependencies and spatial features [20]. CNNs are particularly effective for identifying local patterns in industrial pollutant data, making them well-suited for air pollution forecasting [21]. LSTM networks are capable of modeling non-linear time series data with long-term dependencies and have shown remarkable success in air pollution forecasting [22]. For example, Aggarwal & Toshniwal (2021) employed LSTM to predict air quality in 15 regions in India, achieving high accuracy and showcasing LSTM's strength in handling complex time dependencies [23]. However, these models struggle with handling multiple scales of temporal patterns and long-range dependencies, which are critical for accurately predicting the volatile pollution levels in industrial zones. RNN variants like BiLSTM and GRU further enhance sequence modeling but remain limited in capturing multi-scale temporal information [24].

Given the increasing demand for complex time series forecasting, researchers have explored hybrid deep learning models for improved prediction accuracy. Luo et al. (2021) combined LSTM with XGBoost to effectively predict COVID-19 transmission, demonstrating deep learning's capability for capturing complex time series patterns [25]. Ahmed et al. (2024) developed a hybrid model combining ConvLSTM, SVM, and BiGRU for air quality prediction, achieving precise Air Quality Index (AQI) forecasts [26]. Liang et al. (2020) applied AdaBoost, ANN, Random Forest, Stacking Ensemble, and SVM for AQI-level forecasting with strong results [27]. Tsokov et al. (2022) used a CNN-LSTM hybrid model for spatiotemporal forecasting of $PM_{2.5}$ and air pollution levels at specific sites [28], while Wu et al. (2023) introduced a deep learning-based Res-GCN-BiLSTM hybrid model, integrating Residual Neural Networks (ResNet), Graph Convolutional Networks (GCN), and Bidirectional LSTM (BiLSTM) for short-term regional NO_2 and O_3 forecasting [29]. Kim et al. (2021) combined 3D-CNN with BiLSTM for atmospheric pollutant forecasting, proving deep learning's effectiveness in capturing spatial and temporal trends of air pollutants [30]. Their findings highlighted the effectiveness of deep learning in capturing both temporal and spatial variations in air pollutant dynamics, surpassing traditional methods in performance. However, despite these significant advancements, challenges remain in improving the deep learning models' predictive accuracy and generalization capability in time series forecasting, particularly for diverse and complex datasets.

Initially designed for natural language processing tasks, the Transformer model has recently gained increasing attention in time series forecasting due to its self-attention mechanism [31]. This mechanism enables the model to effectively capture long-term dependencies, offering significant advantages over RNN-based models [32]. For example, Wu et al. (2020) employed a Transformer-based machine learning model to predict influenza trends [33], while Zeng et al. (2023) demonstrated that Transformer models can effectively handle long-span time series data, leveraging their self-attention mechanism to simultaneously focus on multiple critical points in the series, thereby improving prediction accuracy [34]. Furthermore, Liang et al. (2023) introduced AirFormer, an innovative Transformer model designed to predict national air quality in China with unprecedented spatial granularity, covering thousands of locations, thereby enhancing forecasting performance for complex pollutant time series [35]. Zhang and Zhang (2023) utilized a Sparse Attention Transformer Network (STN) to model air quality by learning long-term dependencies and intricate relationships from $PM_{2.5}$ time series data [36]. Unlike traditional sequential models, Transformers can simultaneously process all points in a sequence, making them particularly well-suited

for handling long-sequence data. This capability provides distinct advantages in addressing the complexity of industrial pollution patterns [37].

Building on the success of Transformers, TimesNet further optimizes time feature processing by capturing multi-scale temporal patterns [38]. By incorporating multi-scale processing, TimesNet is particularly effective at modeling the fluctuating pollutant levels in industrial zones, where both gradual trends and sudden peaks must be considered. This model is a significant improvement over the traditional Transformer model, as it enhances its ability to handle the volatility inherent in industrial pollutant forecasting.

Our hybrid model, combining Transformer and TimesNet architectures, offers enhanced accuracy in pollutant forecasting. This hybrid model leverages the Transformer's self-attention mechanism to handle long-term dependencies, while TimesNet captures fine-grained, multi-scale temporal variations. In industrial zones like Xinyang, pollutant levels are affected by both large-scale industrial activity and short-term events (e.g., equipment failures or production surges), for which the hybrid model provides a robust predictive solution. The hybrid model has demonstrated significant improvements in predictive accuracy on complex datasets, effectively addressing both long- and short-term temporal dependencies, which is crucial for accurate forecasting in dynamic industrial environments [39]. In Xinyang, where pollutant levels exhibit both gradual trends and sudden peaks, this hybrid model surpasses traditional models in predictive capability.

This study proposes a hybrid Transformer-TimesNet model that leverages the strengths of both architectures to overcome the limitations of previous approaches. The Transformer model's self-attention mechanism is ideal for handling long-term dependencies and capturing global patterns in pollutant data, while TimesNet's multi-scale temporal processing enhances the model's ability to capture both gradual trends and sudden spikes in pollutant concentrations. The hybrid model effectively manages long-term dependencies and short-term fluctuations in industrial pollutant data, a challenge that traditional statistical and machine learning models cannot solve. Combining Transformer's global attention mechanism with TimesNet's multi-scale temporal processing allows for more accurate modeling of both long-term trends and short-term events, such as production surges or equipment failures. The hybrid model is tailored to handle real-world industrial pollution data's complex and noisy nature, particularly in zones like Xinyang, where both large-scale industrial activities and transient events influence pollutant levels.

While existing models have made significant strides in air quality forecasting, they often struggle to capture the complexities of industrial pollutant data. By combining the strengths of the Transformer and TimesNet architectures, this hybrid model offers a more robust solution to the unique challenges posed by industrial zones. This approach is particularly well-suited for areas like Xinyang, where pollutant levels exhibit both long-term trends and sudden peaks. Through this innovation, this study aims to improve the accuracy and reliability of pollutant forecasting, offering a powerful tool for environmental monitoring and public health protection.

Methods

Transformer Mode

The Transformer model is built on key components, including Self-Attention, Multi-Head Attention, and Positional Encoding. The self-attention mechanism captures dependencies between any positions in a sequence, while the multi-head attention mechanism enhances feature representation by attending to different subspaces of the data in parallel. Positional encoding introduces sequential information to the model, addressing the inherent lack of temporal or positional awareness in Transformer architectures. The standard Transformer framework consists of an encoder and a decoder, though only the encoder is typically employed in time series forecasting tasks. The encoder leverages self-attention mechanisms and feed-forward neural networks to extract features from temporal data [40].

To clarify the functionality of the Transformer in time series forecasting, the following sections detail its core mechanisms and mathematical formulations:

Input Representation

For time series data, assume we have an input sequence $X \in R^{T \cdot D}$, where T is the number of time steps, and D is the feature dimension. The input sequence is projected into a higher-dimensional space through a linear transformation:

$$X_{emb} = X \cdot W_{emb} + b_{emb} \tag{1}$$

where $W_{emb} \in R^{T \cdot d_{model}}$ and $b_{emb} \in R^{d_{model}}$ are trainable weight and bias parameters and d_{model} represents the model dimension.

Positional Encoding

Since the Transformer lacks inherent sequence order information, positional encoding is added to introduce position awareness to the input sequence. The positional encoding is typically implemented using sine and cosine functions:

$$\mathrm{PE}_{(t,2i)} = \sin\left(\frac{t}{1000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

$$PE_{(t,2i+1)} = cos(\frac{t}{1000^{\frac{2i}{d_{model}}}}) \quad (3)$$

where t denotes the position index and i represents the dimension index. The positional encoding vector PE is added to the embedded input data:

$$Z = X_{emb} + PE \quad (4)$$

Self-Attention Mechanism

In the self-attention mechanism, the input Z is projected into three vectors: Query (Q), Key (K), and Value (V):

$$Q = Z \cdot W_Q \quad (5)$$

$$K = Z \cdot W_K \quad (6)$$

$$V = Z \cdot W_V \quad (7)$$

where W_Q, W_K, and W_V are trainable projection matrices. The attention scores are computed by taking the dot product of the query and key vectors, followed by scaling and applying the softmax function:

$$Attention(Q,K,V) = Softmax\left(\frac{Q \cdot K^t}{\sqrt{d_k}}\right) \cdot V \quad (8)$$

where d_K is the dimension of the key vector.

Multi-Head Attention

The multi-head attention mechanism allows for parallel computation across multiple self-attention heads, capturing different feature subspaces. For each head h, the output is:

$$head_h = Attention(Q_h, K_h, V_h) \quad (9)$$

The outputs of all heads are concatenated and passed through a linear transformation:

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h) \cdot W_O \quad (10)$$

where W_O is the output projection matrix.

Feed-forward Neural Network

The output of the multi-head attention layer is passed through a feed-forward neural network, typically consisting of two linear layers with a *ReLU* activation function:

$$FFN(x) = max(0, x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (11)$$

Residual Connections and Layer Normalization

Each sublayer (such as the self-attention layer and the feed-forward neural network) is followed by residual connections and layer normalization:

$$LayerNorm(x + SubLayer(x)) \quad (12)$$

where the *SubLayer*(x) could be either the self-attention layer or the feed-forward neural network and *LayerNorm* refers to the layer normalization operation.

Model Output

For time series forecasting, the output of the Transformer model represents the predicted sequence, indicating the values for future time steps. Let the final output of the Transformer be Z_{out}. The forecasted values are obtained by projecting this output through a linear layer:

$$\hat{y} = Z_{out} \cdot W_y + b_y \quad (13)$$

This comprehensive mechanism enables the Transformer to excel in capturing long-term dependencies, making it well-suited for complex time series forecasting tasks.

TimesNet Model

TimesNet is a deep learning model specifically designed for time series forecasting, focusing on efficiently handling complex and long-sequence time series data. This model integrates multiple deep learning techniques to enhance learning capacity and prediction accuracy. Key components of TimesNet include the Temporal Convolutional Block, Global Average Pooling, and a Fully Connected Layer [41].

Input Segmentation

TimesNet decomposes the time series into different temporal blocks, where each block represents a periodic segment of the time series. Suppose the sequence is divided into K blocks, each of length L, then the input can be represented as:

$$X = \{X_1, X_2, \dots, X_K\}, X_i \in R^{L \cdot d} \quad (14)$$

Where L is the length of each block, and d is the feature dimension of the time series.

Temporal Convolutional Blocks

The temporal convolutional block uses multiple one-dimensional convolutional layers to capture dependencies across different time steps. For a single convolutional layer, the output feature H can be represented as:

$$H = ReLU[Conv1D(X, W_{conv}) + b_{conv}] \quad (15)$$

where W_{conv} is the convolutional kernel, b_{conv} is the bias term, and *ReLU* is the activation function.

Global Average Pooling

After feature extraction by the temporal convolutional block, TimesNet applies global average pooling to aggregate information along the time dimension. Let the output from the temporal convolutional block be $H \in R^{T \times M}$, where T is the number of time steps after convolution, and M is the feature dimension. Global average pooling computes the mean along the time dimension, producing a global feature vector G:

$$G = \frac{1}{T}\sum_{t=1}^{T} H_t \quad (16)$$

This reduces the temporal information into a single global feature vector while retaining important characteristics of the input sequence.

Fully Connected Layer

The global feature vector is then passed through a fully connected layer to map it to the final output layer, which generates the time series predictions. If the goal is to predict the next time step value, the output can be represented as:

$$\hat{y} = G \cdot W_{fc} + b_{fc} \quad (17)$$

where W_{fc} and b_{fc} are the weights and biases of the fully connected layer and $\hat{y}$ represents the predicted value for the next time step.

Model Fusion Approach

This research proposes a feature-level fusion method to integrate Transformer and TimesNet, retaining and leveraging the unique features extracted by each model to maximize their strengths. Specifically, features from both models are concatenated, allowing the combined model to capture both local and global information and enhance predictive accuracy.

In this fusion process, TimesNet extracts local temporal features from the time series data, while the Transformer model provides global contextual information. The features are concatenated along the feature dimension. Assuming the feature vector output from TimesNet is $Z_{TimesNet}$ and from Transformer is $Z_{Transformer}$, the fused feature vector Z_{fusion} is represented as:

$$Z_{fusion} = Concat(Z_{TimesNet}, Z_{Transformer}, dim = -1) \quad (18)$$

The concatenated feature vector Z_{fusion} is then input into a fully connected layer W_{fc} to produce the final prediction:

$$\hat{y} = Z_{fusion} \cdot W_{fc} + b_{fc} \quad (19)$$

This feature fusion approach effectively combines the strengths of both models, capturing both local temporal patterns and global contextual information, thereby improving the prediction accuracy for pollutant concentration forecasting. Fig. 1 presents the architecture of the proposed model, including the Transformer processing module, the TimesNet processing module, and the model fusion process.

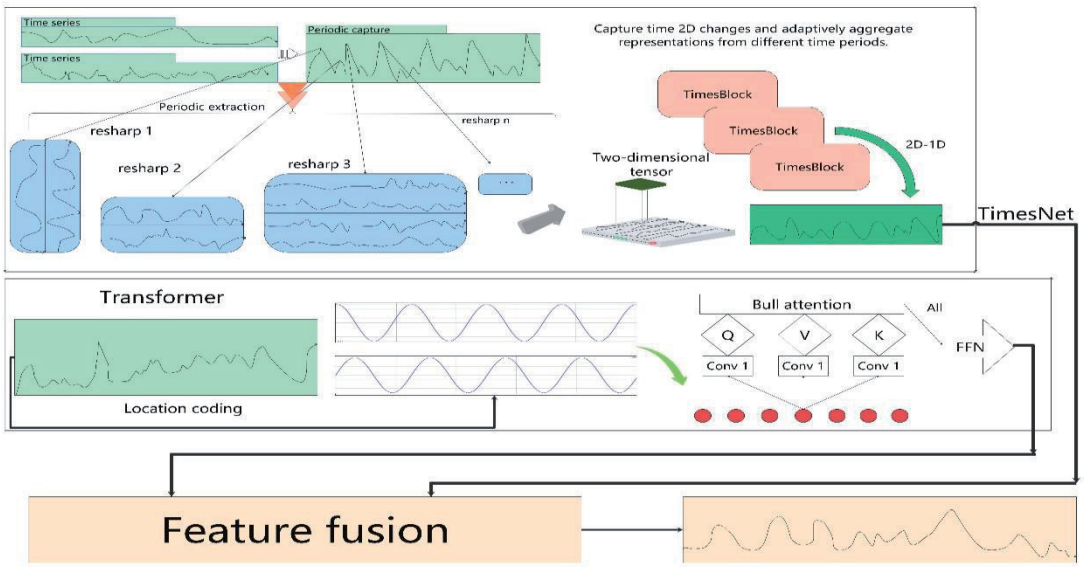

Fig. 1. Mechanism and Workflow of the Transformer-TimesNet Hybrid Model.

Architecture Optimization

The core architecture design of the model is as follows:

TimesNet: Defined with a TimesBlock component, which includes a 1D convolutional layer to extract local features from time series data and introduces non-linearity through a ReLU activation function. It then applies global average pooling to reduce dimensionality, followed by a fully connected layer to generate the final output.

Transformer Model: The Transformer first embeds the input into a higher-dimensional space, then processes these embedded features through multiple encoder-decoder layers, ultimately producing predictions via a fully connected layer.

Fusion Model: This uses a concatenation fusion approach, where the outputs of the two models are concatenated along the feature dimension to form a new feature vector. This vector is then passed through a fully connected layer to yield the final prediction [42].

To optimize model performance, hyperparameters for both sub-models are tuned to ensure structural and functional harmony. Table 1 outlines the key hyperparameters and their search ranges. Specifically, TimesNet hyperparameters include hidden_size and num_blocks, which determine hidden layer size and convolutional block count, which are critical for enhancing their ability to handle time series data. Transformer hyperparameters include hidden_size (kept consistent with TimesNet for feature dimension alignment) and num_heads, which determine the parallelism of the multi-head attention mechanism. The learning rate is also optimized for the overall model's training effectiveness.

This study employs the AdamW optimizer, an Adam optimization algorithm variant that independently controls weight decay to mitigate overfitting, making it particularly suited for complex models like the Transformer-TimesNet fusion. Unlike the traditional Adam optimizer, which applies L2 regularization coupled with Adam's adaptive learning rates, AdamW decouples weight decay from the gradient update, which can significantly enhance generalization in large, complex models [43, 44].

The Optuna framework is used to further enhance model performance and optimize hyperparameters. Optuna is an efficient, automated optimization framework that uses Bayesian optimization to search the hyperparameter space by building a probabilistic model and selecting new hyperparameter combinations in each iteration based on an acquisition function, typically Expected Improvement (EI) or Upper Confidence Bound (UCB) [45].

Table 1. Potential Hyperparameter Ranges.

Hyperparameter	Search Range
Hidden_size	32 to 128
Num_blocks (TimesNet)	1 to 5
Num_heads (Transformer)	2, 4, 8
Learning_rate	0.0005 to 0.01

Within Optuna, each hyperparameter combination generates an objective function value to minimize (or maximize) the function to optimize model performance. To speed up the search, Optuna implements an early stopping mechanism that halts optimization if the objective function shows no significant improvement over a set number of iterations [46].

For multi-objective optimization, since various hyperparameter combinations may impact multiple performance metrics, Optuna applies the Pareto optimality concept to find a balance among the objectives, optimizing each while maintaining overall performance. We defined the target objective function through extensive experiments based on hyperparameters such as hidden layer size, convolution block count, attention headcount, and learning rate. We selected the optimal configuration to minimize test set loss [47].

This architecture and hyperparameter optimization strategy successfully built an efficient hybrid model that leverages Transformer and TimesNet for accurate forecasting. The hyperparameter tuning improved the model's accuracy in long-term predictions and allowed it to excel in pollutant concentration forecasting. Optuna's effective hyperparameter optimization, combined with the weight decay mechanism of AdamW, further enhanced the model's robustness and generalization ability. Future studies could explore other optimization algorithms and model fusion strategies to further enhance industrial pollutant forecasting performance.

Materials

Spatial Scope and Time Span

This study focuses on the Xinyang Industrial Zone in Xiamen City as the research area. The zone is situated south of Maluan Bay, north of Caijianwei Mountain, extending east to Wengcuo and west to Haixin Highway, approximately 11.3 km from downtown Xiamen. The terrain slopes downward from south to north, with a slightly elevated central east-west corridor flanked by lower-lying areas on both sides, as shown in Fig. 2.

The pollutant concentration data used in this research, including SO_2, NO_2, CO, O_3, PM_{10}, and $PM_{2.5}$, were obtained from the air quality monitoring data center of the environmental monitoring station in Xinyang Industrial Zone. The data, with an hourly granularity, spans the period from September 1, 2020, to January 7, 2024, encompassing a total of 26586 air quality records.

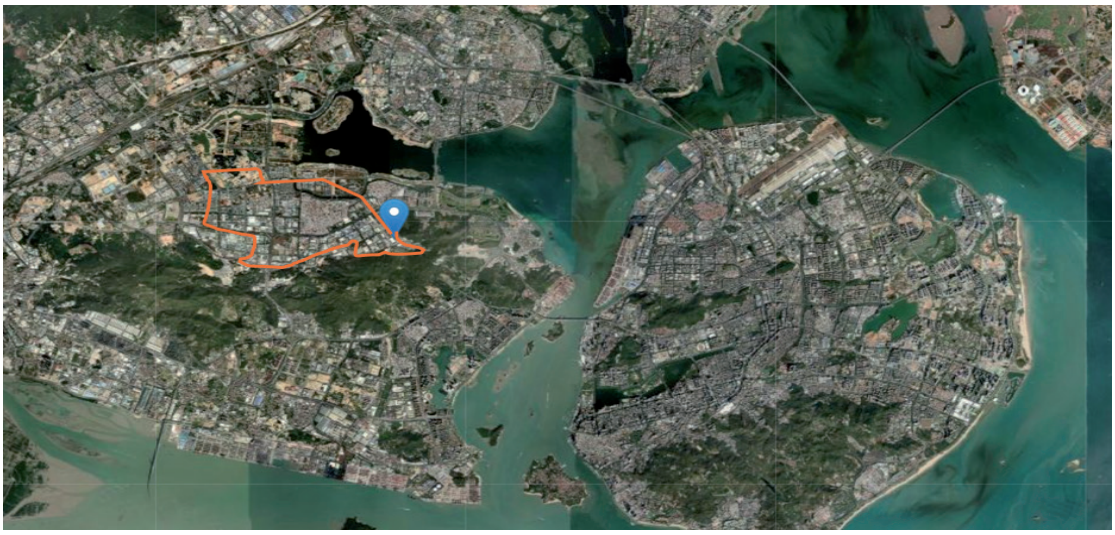

Fig. 2. Geographic Location of Xinyang Industrial Zone.

This study aims to predict the concentrations of these pollutants and evaluate the performance of the proposed Transformer-TimesNet hybrid model by comparing it against other models. To ensure comparability across models, a consistent training and testing dataset split was used: the first 80% of the data was utilized for model training, while the remaining 20% was reserved for testing.

Table 2. Summary of anomalous values in the dataset.

Feature	Missing Values	Negative Values
SO_2	0	310
NO_2	8	405
CO	15	1248
O_3	0	160
PM_{10}	0	139
$PM_{2.5}$	0	246

Data Preprocessing

Environmental data is highly susceptible to external factors, such as weather conditions, equipment malfunctions, or sudden pollution events, which can introduce anomalies. Table 2 presents an overview of the anomalous values identified in the dataset. In this dataset, a value of -1 indicates missing data, which is clearly inconsistent with real-world conditions. Therefore, missing values were replaced with 0, and a descriptive statistical analysis was conducted. The results are summarized in Table 3.

Table 3 presents each variable's sample size, mean, standard deviation, minimum, maximum, and quartiles. The descriptive statistical analysis reveals significant outliers in the dataset. For instance, while the mean of CO is 3.04, the maximum value reaches 11059.8, which is clearly an anomaly. To mitigate the impact of these outliers on model training, this study employs Support Vector Machine (SVM) methods to address them by replacing the outliers with the mean value. SVM effectively identifies and isolates outliers by

Table 3. Descriptive Statistics of the Dataset.

Pollutant	SO_2	NO_2	CO	O_3	PM_{10}	$PM_{2.5}$
Count	26586	26586	26586	26586	26586	26586
Mean	4.04645	21.76	3.03572	52.0312	40.6862	18.3441
Std	2.16587	13.4463	137.342	32.7039	22.5754	9.94619
Min	0	0	0	0	0	0
25%	3	12	0.4	26	25	11
50%	4	19	0.5	49	36	17
75%	5	28	0.6	73	53	24
Max	97	116	11059.8	218	222	153

constructing a maximized boundary, preventing these anomalies from negatively affecting the model. This approach is particularly suitable for datasets with long-tailed distributions or outliers, as it helps smooth the data and reduce the interference of outliers on model performance. The core principle of SVM for outlier detection is constructing a boundary (hyperplane or dividing surface in higher-dimensional space) to separate normal data from anomalies [48].

During the anomaly detection process, we employed the One-Class Support Vector Machine (One-Class SVM) algorithm and applied data normalization to

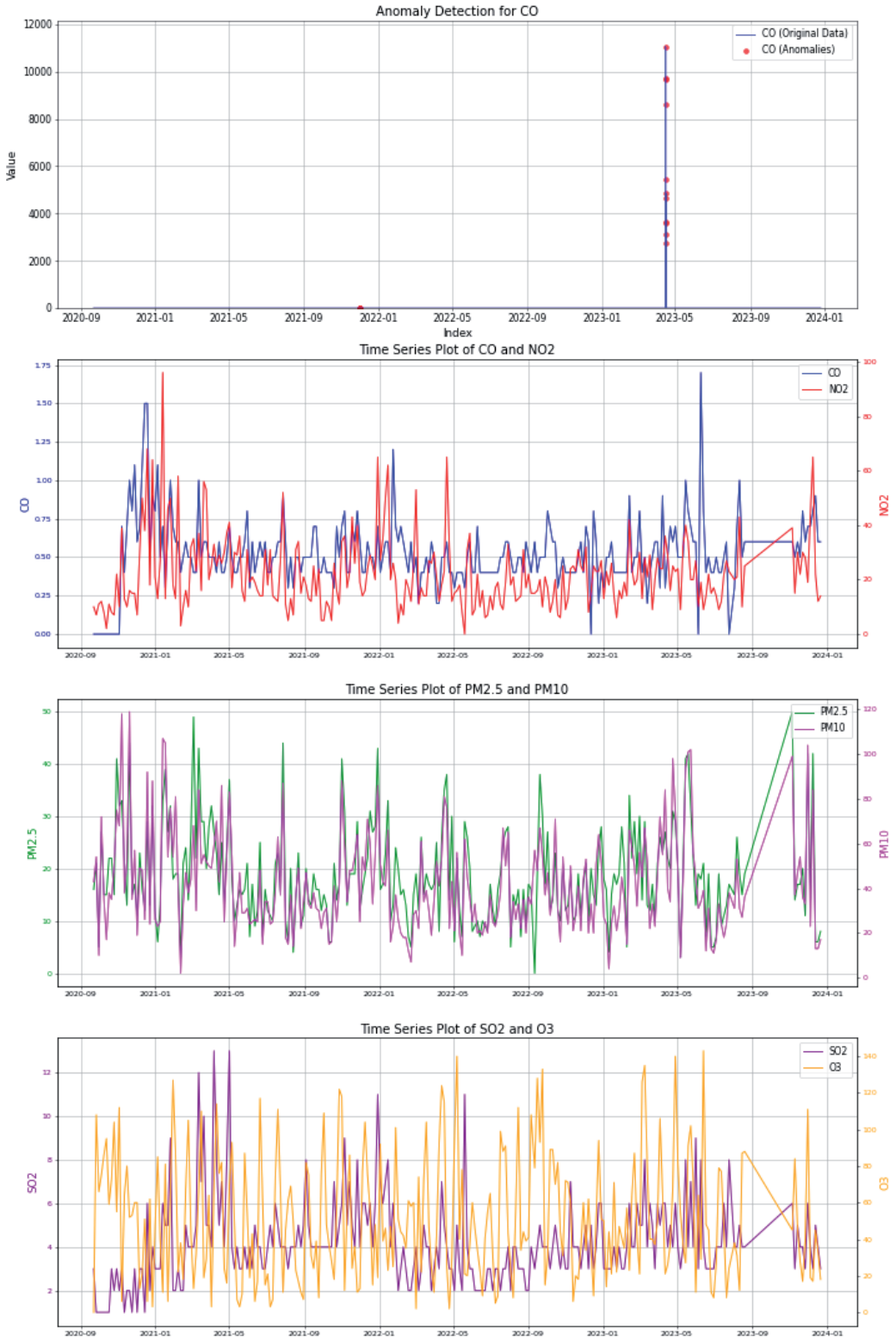

Fig. 3. SVM-Based Anomaly Detection Results and Time Series Plot of Pollutant Concentrations.

ensure consistency. The radial basis function (RBF) was used as the kernel, where the gamma parameter, which controls the influence range of the kernel, was set to 0.1. The nu parameter, which defines the proportion of data points classified as anomalies, was set to 0.05. After training the model, we predicted the status of each data point, classifying them as either normal (1) or anomalous (-1). The data were then inverse-transformed to restore the original scale [49]. Based on the SVM detection results, anomalous values were replaced with the mean of the normal data points, thereby minimizing the impact of anomalies on subsequent analyses and ensuring data consistency.

Since descriptive statistics show that only the CO column contains significant outliers, the SVM outlier detection method is applied to this column, and outliers are replaced with the mean value, as shown in Fig. 3.

Due to the large volume of data, directly plotting all time series data would result in overly dense charts that are difficult to interpret. Therefore, we applied downsampling to reduce the number of plotted data points, preserving the main trends and patterns in the time series and making the charts clearer. The downsampled data retains key long-term trends, seasonal changes, and periodic patterns. Reducing the noise and detail helps better identify long-term trends and seasonal fluctuations. Fig. 3 displays the time series graphs of the concentrations of various pollutants. It can be observed that some pollutants exhibit similar trends within the industrial zone, such as NO_2 and $PM_{2.5}$, as well as PM_{10} and $PM_{2.5}$.

Seasonal Fluctuation Study

A detailed annual seasonal analysis of pollutant concentrations was conducted, and the results are presented in Fig. 4. The SO_2 concentration exhibits a relatively stable annual average, but its seasonal averages

Fig. 4. Annual Mean Seasonal Analysis of Pollutant Concentrations, Monthly Average Concentrations, and Seasonal Fluctuation Rates (2023).

are higher in spring and winter. NO_2 concentrations are higher in spring and summer, with a lower peak value in summer. PM_{10}'s distribution of annual mean values is similar to that of NO_2, but its peak values are mainly concentrated in the autumn. CO concentrations show a decreasing trend in peak values over time. O_3 concentrations have higher mean values in summer and winter, with peak values concentrated in autumn. The maximum concentration of $PM_{2.5}$ occurs in the summer.

The observed patterns suggest that pollutant concentrations exhibit consistent periodic fluctuations throughout the year. Since the fluctuation trends across different years are similar, this study focuses on the 2023 pollutant concentration data for detailed volatility analysis, as shown in Fig. 4. SO_2 and CO show low volatility and remain relatively stable, indicating that the concentration changes for these pollutants are minimal throughout the year. In contrast, other pollutants exhibit varying degrees of volatility, with PM_{10} and O_3 showing more significant fluctuations. Furthermore, NO_2 exhibits substantial fluctuations in early to mid-August, likely related to increased industrial activity during that period.

Table 4 displays the concentration limits for various pollutants set by the Ministry of Ecology and Environment of China. These limits reflect the country's air quality management policies and emphasize the strict monitoring of major pollutants (Ministry of Ecology and Environment of China, 2016). These limits provide important reference points for further analysis of pollution levels in this study area.

Based on the concentration limits in Table 4 and the frequency distribution histogram in Fig. 5, it is clear to discern the pollution levels of each pollutant. SO_2 concentrations mostly fall within the 0-25 μg/m^3 range, not reaching the Level 1 pollution threshold. NO_2 concentrations predominantly range from 0-100 μg/m^3, staying below the Level 1 pollution warning level. PM_{10} and $PM_{2.5}$, evaluated based on 24-hour average concentrations, show relatively high levels in the industrial zone, with concentrations meeting the Level 1 pollution threshold. CO concentrations are mostly within the 0-2 mg/m^3 range, with few instances exceeding 3 mg/m^3, indicating that CO concentrations are well-controlled. O_3 concentrations primarily fall within the 0-200 μg/m^3 range, with some instances exceeding the Level 1 pollution threshold.

Results and Discussion

Evaluation Metrics

This study adopted a series of commonly used evaluation metrics, including *MAE*, *MASE*, R^2, *RMSE*, *RMSSE,* and *SMAPE* [50]. These metrics provide a comprehensive quantitative basis for evaluating the model's prediction performance, ensuring the scientific rigor and reliability of the results. Each metric performs differently across various pollutants. To present a clearer comparison of model performance, both quantitative results and fitting graphs are provided to visually illustrate the prediction accuracy of the models.

Table 4. Environmental Air Pollutant Concentration Limits.

Pollutant	Average Time	Concentration Limit		Unit
		Level 1	Level 2	
SO_2	Annual Average	20	60	μg/m^3
	24-hour Average	50	150	
	1-hour Average	150	500	
NO_2	Annual Average	40	40	
	24-hour Average	80	80	
	1-hour Average	200	200	
CO	24-hour Average	4	4	mg/m^3
	1-hour Average	10	10	
O_3	Daily Max 8-hour Average	100	160	μg/m^3
	1-hour Average	160	200	
PM_{10}	Annual Average	40	70	
	24-hour Average	50	150	
$PM_{2.5}$	Annual Average	15	35	
	24-hour Average	35	75	

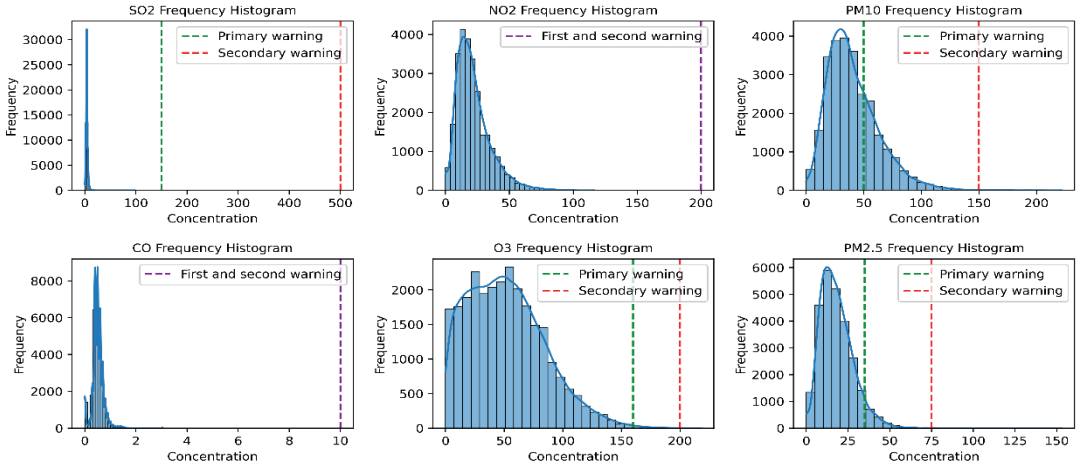

Fig. 5. Frequency Histogram of Pollutant Concentrations.

Mean Absolute Error (*MAE*): The average absolute difference between the actual observed values and the predicted values across *n* samples are calculated as:

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{20}$$

Where y_i represents the actual values and $\hat{y}_i$ represents the predicted values.

Mean Absolute Scaled Error (*MASE*): This is a normalized version of *MAE* used to compare the forecasting performance across different time series. The lower the *MASE*, the better the improvement in prediction relative to a baseline model:

$$MASE = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|}{\frac{1}{n-m} \sum_{i=m+1}^{n} |y_i - y_{i-m}|} \tag{21}$$

Coefficient of Determination (R^2): R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables. The value of R^2 ranges between 0 and 1, where a value closer to 1 indicates a stronger explanatory ability of the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{22}$$

Root Mean Squared Error (*RMSE*): *RMSE* is the square root of the mean squared error, which quantifies the standard deviation of prediction errors:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{23}$$

Root Mean Squared Scaled Error (*RMSSE*): This scaled version of *RMSE* is used to compare prediction performance across different datasets. A smaller *RMSSE* indicates a greater improvement in prediction accuracy compared to a baseline model. Its advantage lies in standardizing comparisons across varying dataset sizes and variability. The formula is:

$$RMSSE = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\frac{1}{n-m} \sum_{i=m+1}^{n} |y_i - y_{i-m}|} \tag{24}$$

Symmetric Mean Absolute Percentage Error (*SMAPE*): *SMAPE* measures the relative error between the predicted and observed values, which is particularly effective when observed values are near zero. Lower values indicate better predictive accuracy. *SMAPE* performs well in time series forecasting and scenarios requiring relative error assessments. The formula is:

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{25}$$

Model Training

The specific steps of the training process are as follows:

Epoch Setup: Set the training cycles to 30 epochs to ensure the model converges after sufficient iterations.

Model Training Mode: Enable the model's training mode to prepare it for the training process.

Loss Accumulation: Use epoch_loss to accumulate the loss from each batch and continuously monitor the training progress.

Data Format Conversion: Convert each batch of data into a format suitable for the TimesNet and Transformer models.

Gradient Zeroing: Clear the gradients in the optimizer before each optimization step to avoid cumulative effects.

Forward Propagation: Pass the input data through the hybrid model to compute the predicted output.

Loss Calculation: Compute the loss function value based on the predicted output and true labels.

Table 5. Optimal Hyperparameter Settings.

Hyperparameter	Search Result
Hidden_size (hidden layer size)	64
Num_blocks (TimesNet blocks)	3
Num_heads (attention heads)	2
Learning_rate (learning rate)	0.001058

Backward Propagation: Use loss.backward() to calculate the gradients for updating the model weights.

Parameter Update: Use optimizer.step() to adjust the model parameters based on the computed gradients.

The optimal hyperparameters obtained from training are shown in Table 5. This parameter configuration effectively balances the model's learning capacity and stability, improving prediction accuracy.

Model Comparison

To improve the accuracy of pollutant concentration prediction, this study constructs separate TimesNet, Transformer, and RNN models and uses the PyCaret library to incorporate various traditional statistical and machine learning models for comparison and evaluation. This helps highlight the hybrid model's predictive advantages.

RNN is a deep learning model for handling sequential data and capturing temporal dependencies in air quality monitoring data. The hourly pollutant concentration depends on the previous values, so RNN effectively identifies these sequential dependencies. The Transformer model can capture long-term dependencies and global features from different time points in the sequence. TimesNet, a deep learning model designed specifically for time series forecasting, further optimizes time feature processing by extracting multi-scale temporal patterns and is suitable for complex time series data, capturing dependencies at different time scales.

Multiple statistical and machine learning models were constructed using PyCaret for benchmarking to comprehensively evaluate model performance. This comparative approach between deep learning models, traditional statistical methods, and machine learning frameworks provides a robust basis for assessing the proposed hybrid model's superior performance.

Performance of the Transformer-TimesNet Hybrid Model

Table 6 shows the evaluation metrics of the Transformer-TimesNet hybrid model for predicting pollutant concentrations.

SO_2: R^2 is 0.6057, indicating that the model captures the main trend of SO_2 changes. The *MAE* is 0.5703, and the *SMAPE* is 0.1233, indicating low prediction errors.

NO_2: R^2 is 0.7239, showing a good fit for NO_2. The *MAE* is 4.3955, and the *RMSE* is 6.5545, with relatively low error.

CO: R^2 is 0.7136, indicating good model performance with low error.

O_3: R^2 is 0.8635, indicating the model's high adaptability to complex time series data.

PM_{10}: R^2 is 0.7815, which reflects a good prediction of PM_{10} variation, although errors (*MAE*: 7.6261, *RMSE*: 10.9775) are relatively high.

$PM_{2.5}$: R^2 is 0.7970, and *SMAPE* is 0.1891, showing good performance with minimal error.

Performance of Deep Learning Models

Table 7 presents the evaluation results of RNN, Transformer, and TimesNet models for pollutant concentration prediction. The following is a detailed analysis of each model's performance:

RNN Model Performance

The RNN model demonstrates stable performance in predicting most pollutant concentrations, effectively capturing short-term trends in time series data. The R^2 values for most pollutants are above 0.6, indicating that the model explains a substantial portion of the variance in the data and exhibits strong fitting capabilities.

SO_2: With an R^2 of 0.6123, the model explains 61.23% of the variance in SO_2 concentration, with *MAE* = 0.6585, *RMSE* = 0.8731, and *SMAPE* = 0.1208, indicating high prediction accuracy.

Table 6. Evaluation of the Transformer-TimesNet Hybrid Model on Pollutant Concentration Prediction.

Model: Transformer-TimesNet	MAE	MASE	R^2	RMSE	RMSSE	SMAPE
SO_2	0.57028	1.13629	0.60565	0.88054	0.94343	0.12325
NO_2	4.39547	0.99247	0.7239	6.55454	0.9173	0.1851
CO	0.05998	1.26037	0.71358	0.10551	0.9042	0.1415
O_3	8.13991	0.90167	0.86352	12.5324	0.88607	0.23398
PM_{10}	7.62612	0.97896	0.78153	10.9775	0.93629	0.18719
$PM_{2.5}$	3.09063	1.03604	0.79704	4.4887	0.97472	0.18907

Table 7. Evaluation of RNN, Transformer, and TimesNet Models for Pollutant Concentration Prediction.

Model: RNN	MAE	MASE	R^2	RMSE	RMSSE	SMAPE
SO_2	0.65852	1.11285	0.61229	0.8731	0.93546	0.12083
NO_2	4.31879	0.97516	0.71108	6.70499	0.93835	0.18415
CO	0.05557	1.16756	0.71405	0.10542	0.90345	0.13639
O_3	9.46619	0.93782	0.84382	13.4068	0.94789	0.24252
PM_{10}	9.25754	0.93165	0.79244	10.6998	0.91261	0.17716
$PM_{2.5}$	6.10028	1.03928	0.79254	4.53816	0.98546	0.18903
Model: Transformer	MAE	MASE	R^2	RMSE	RMSSE	SMAPE
SO_2	2.09265	4.16962	-2.0445	2.44661	2.62136	0.59956
NO_2	9.78553	2.20952	-0.2976	14.2097	1.98863	0.42479
CO	0.13557	2.84856	0.18399	0.17809	1.52618	0.28627
O_3	33.5959	3.72148	-0.7136	44.4085	3.13976	0.78687
PM_{10}	34.3017	4.40331	-1.9268	40.1791	3.42694	1.22049
$PM_{2.5}$	16.3477	5.48007	-2.5311	18.7226	4.0656	1.42923
Model: TimesNet	MAE	MASE	R^2	RMSE	RMSSE	SMAPE
SO_2	0.88525	1.76385	0.21193	1.24477	1.33368	0.19618
NO2	7.83147	1.7683	0.32417	10.2548	1.43514	0.33383
CO	0.08804	1.84978	0.55744	0.13115	1.12394	0.19157
O_3	23.7157	2.62703	0.22133	29.9352	2.11648	0.52825
PM_{10}	12.2553	1.57321	0.51568	16.3444	1.39404	0.29654
$PM_{2.5}$	5.36558	1.79865	0.48922	7.12078	1.54628	0.30583

NO_2: An R^2 of 0.7111 shows strong prediction performance, with a low *SMAPE* of 0.1842, demonstrating effective forecasting.

CO: With R^2 = 0.7141, the model accurately predicts CO concentrations, as indicated by the low *MAE* (0.0556) and *RMSE* (0.1054).

O_3: The model's R^2 value of 0.8438 reflects its strong ability to explain O_3 concentration variations.

PM_{10} & $PM_{2.5}$: The R^2 values of 0.7924 and 0.7925 indicate satisfactory prediction performance with relatively low error metrics.

Transformer Model Performance

Despite its theoretical advantages with self-attention mechanisms, the Transformer model performs relatively poorly in this task. Most pollutants have negative R^2 values, suggesting that the model fails to effectively capture time series dependencies, even under ideal conditions.

SO_2 & NO_2: The Transformer model struggles to predict SO_2 and NO_2, with R^2 values of -2.0445 and -0.2976 and very high *SMAPE* values (0.5996 and 0.4248), indicating large prediction errors.

CO: The model performs slightly better for CO, with R^2 = 0.1840 and *SMAPE* = 0.2863, but still lags behind RNN in performance.

O_3 & PM_{10}: The Transformer model shows poor results in predicting O_3 and PM_{10}, with R^2 values of -0.7136 and -1.9268, respectively, and very high *SMAPE* scores.

$PM_{2.5}$: Similarly, the prediction of $PM_{2.5}$ also underperforms, with R^2 = -2.5311 and *SMAPE* = 1.4292.

TimesNet Model Performance

TimesNet performs well for some pollutants, especially CO, but struggles with others, such as NO_2 and O_3. Its advantage lies in extracting local time series features, though it faces limitations when capturing long-term dependencies.

SO_2: The model's R^2 of 0.2119 can explain only 21.19% of SO^2 variations. However, it demonstrates moderate prediction accuracy with *MAE* = 0.8853 and *SMAPE* = 0.1962.

NO_2 & O_3: The model's performance is weaker for NO_2 (R^2 = 0.3242, *SMAPE* = 0.3338) and O_3 (R^2 = 0.2213, *SMAPE* = 0.5283).

CO: The model performs well for CO, with R^2 = 0.5574, *MAE* = 0.0880, and *SMAPE* = 0.1916, indicating a good fit and minimal prediction error.

PM_{10} & $PM_{2.5}$: The model's performance for PM_{10} (R^2 = 0.5157, *SMAPE* = 0.2965) and $PM_{2.5}$ (R^2 = 0.4892, *SMAPE* = 0.3058) is moderate but still reflects some prediction errors.

(1) Comparison of Transformer-TimesNet Hybrid Model

The RNN model performs well in predicting PM_{10}, CO, and $PM_{2.5}$, effectively capturing these pollutants' temporal trends and short-term dependencies. Particularly in short-term forecasting, RNN handles the concentration variations of these pollutants well, with relatively high R^2 values, indicating good model fit. However, the RNN model exhibits larger errors when predicting pollutants like O_3. This is especially evident when attempting to capture complex long-term dependencies, limiting its prediction accuracy for long-term time series data. Despite this, RNN's performance remains acceptable in multi-feature time series forecasting, especially for pollutants with short-term concentration fluctuations, where it successfully captures most pollutant concentration changes.

The Transformer model, which employs the self-attention mechanism, is theoretically capable of capturing global information, making it suitable for handling time series data with complex dependencies. Transformer excels in learning long-term dependencies and global patterns. However, its performance is suboptimal when used alone for pollutant concentration forecasting. The model underperforms in predicting most pollutants, such as SO^2 and PM_{10}, suggesting that Transformer has limitations in capturing local temporal dependencies. While the self-attention mechanism allows for long-distance dependencies, it struggles with local dependencies in time series data, which hampers its effectiveness in pollutant concentration prediction compared to other models.

TimesNet focuses on extracting local temporal features, which results in a strong performance for data with significant local dependencies, such as CO, providing a good fit and low error. However, TimesNet's performance is moderate when predicting pollutants like NO_2 and PM_{10}, particularly poor for NO_2, where the model's fit is inadequate. Additionally, TimesNet struggles with long-term predictions, especially on larger datasets. Its ability to extract local features makes it highly effective for short-term predictions of certain pollutants, but its performance is constrained when dealing with complex temporal relationships, particularly long-term dependencies.

Although the individual performances of Transformer and TimesNet are subpar, the Transformer-TimesNet hybrid model combines TimesNet's local temporal feature extraction with the Transformer's ability to capture global information, effectively compensating for the weaknesses of both models. TimesNet first processes the input data to extract local temporal features, and Transformer further captures global patterns and long-term dependencies. This synergy allows TimesNet to provide local modeling capabilities, while Transformer's self-attention mechanism enhances the model's understanding of long-term dependencies and global patterns, achieving a complementary relationship between local and global features. This hybrid approach is better equipped to handle complex time series data, particularly pollutants with both local and long-term dependencies. The hybrid model demonstrates excellent performance in pollutant concentration forecasting, particularly for O_3 and $PM_{2.5}$. Compared to Transformer or TimesNet alone, the hybrid model captures key patterns in the time series data more effectively through multi-level feature processing. By integrating the strengths of both models, the Transformer-TimesNet hybrid model improves prediction accuracy for pollutant concentrations. When compared to individual deep learning models, the hybrid model offers a more comprehensive ability to capture both local and global patterns, providing greater adaptability.

Performance of Traditional Statistical Models and Machine Learning Models

The PyCaret library was used to automatically construct multiple traditional statistical methods and machine learning prediction models compared to traditional statistical methods and machine learning models. Traditional statistical methods generally require stationary data, so we first tested the stationarity of the data. Using unit root tests (such as the ADF test), we checked the stationarity of the data. The p-value in the ADF test indicates the likelihood of the null hypothesis being true, where the null hypothesis states that the data series contains a unit root, meaning the data is non-stationary. By plotting the autocorrelation function (ACF) and partial autocorrelation function (PACF) for each indicator (shown in Fig. 6), and based on the test results, the p-values for all statistical indicators were less than 0.05, rejecting the null hypothesis, indicating that the data is stationary [51]. Therefore, we proceeded with modeling based on stationary time series data, and the results were as follows, with the best models selected for comparison.

During the model construction process, we compared the prediction performance of multiple models, including deseasonalized and detrended models. The concentrations of SO_2, NO_2, CO, O_3, PM_{10}, and $PM_{2.5}$ were predicted, and the prediction accuracy of the models was evaluated. The models tested include Decision Tree, Light Gradient Boosting Machine (LightGBM), Gradient Boosting Machine (GBR), Orthogonal Matching Pursuit (OMP), CatBoost Regressor, Random Forest (RF), Linear Regression (LR), Ridge Regression, Extreme Random Trees (ET), Bayesian Ridge Regression (BR), AdaBoost, Lasso Regression, Lasso Least Angle Regression (LLAR), Croston method, Elastic Net (EN), Extreme Gradient

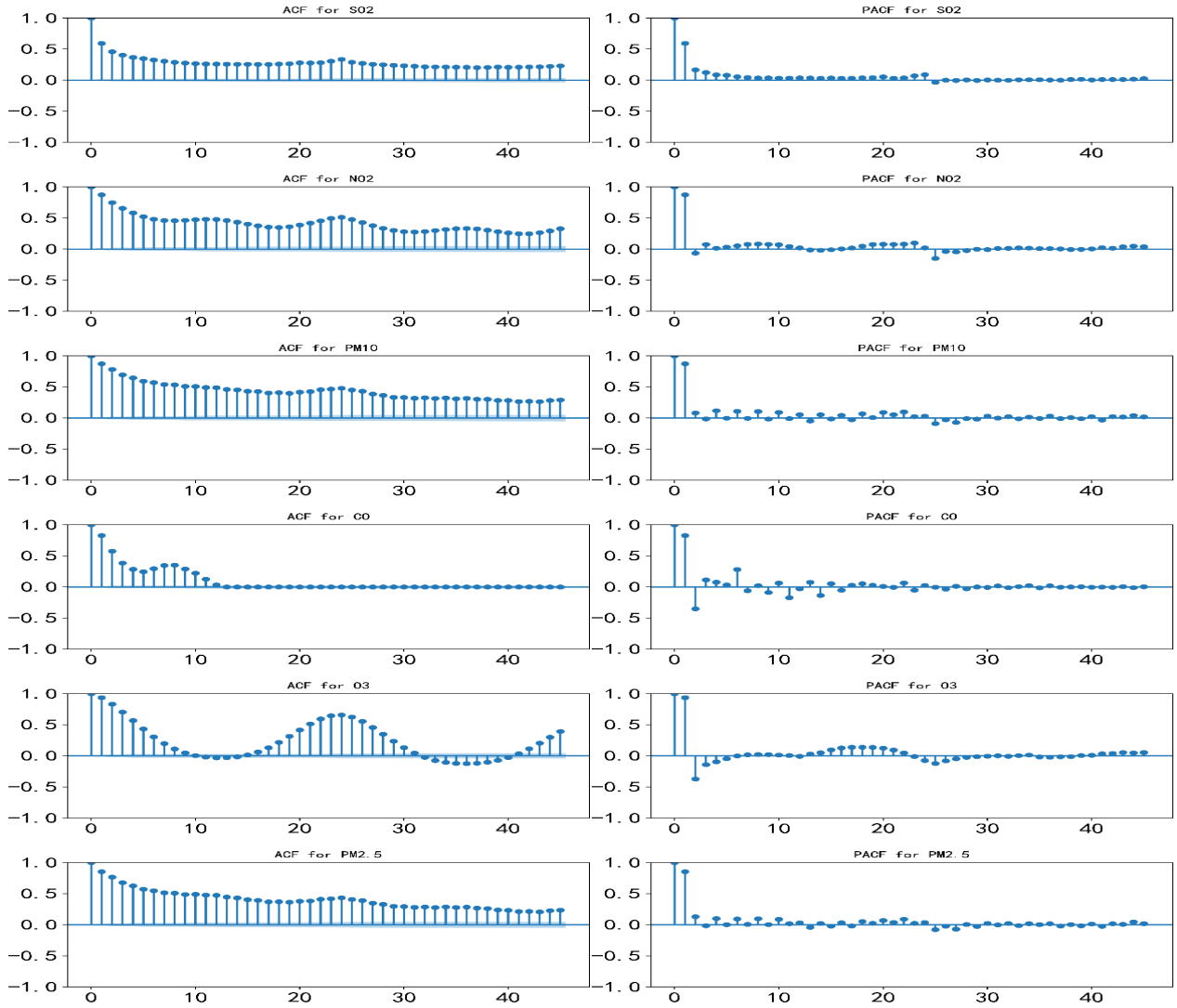

Fig. 6. ACF and PACF Plots of Various Prediction Indicators.

Boosting (XGBoost), K-Nearest Neighbors (KNN), Huber Regression, and ARIMA time series model. Ultimately, we selected the best-performing models from the multiple models for comparison. The specific results are shown in Table 8, with the optimal indicators highlighted in bold.

Based on the results of PyCaret's multi-model construction and selection, the performance of the best models for different pollutants varies significantly. The optimal model for SO_2 is Linear Regression, but its R^2 value is only 0.0011, indicating limited explanatory power. The optimal model for NO_2 was the Croston model, with *MAE*, *MASE*, R^2, and *RMSE* values of 4.382, 0.572, -0.0001, and 5.37, respectively. However, the negative R^2 suggests poor generalization ability. For CO, the best model was KNN, with *MAE*, *MASE*, R^2, and *RMSE* values of 0.11, 1.301, -0.515, and 0.1228, respectively, performing well in *MAE* and *SMAPE* but with a low R^2. The best model for O_3 was LightGBM, with *MAE*, *MASE*, R^2, and *RMSE* values of 6.834, 0.364, 0.812, and 7.791, respectively, indicating good performance in capturing the variation in O_3 concentration. For PM_{10}, the best model was the CatBoost Regressor, with *MAE*, *MASE*, R^2, and *RMSE* values of 12.869, 0.899, -0.144, and 16.3046, showing poor performance with a negative R^2. For $PM_{2.5}$, the best model was the Decision Tree, with *MAE*, *MASE*, R^2, and *RMSE* values of 9.069, 1.38, -0.96, and 11.915, indicating poor performance with a negative R^2.

The study results show that compared to the models selected by PyCaret, the hybrid model based on Transformer and TimesNet outperforms key evaluation metrics like R^2. While certain traditional statistical methods and machine learning models (such as CatBoost and LightGBM) perform well in specific metrics (e.g., the Croston model for NO_2 outperforms the hybrid model in *MASE* and *RMSSE*), their low R^2 values indicate poor generalization. In contrast, the hybrid model demonstrates stronger trend-capturing ability and generalization performance in forecasting CO, PM_{10}, and $PM_{2.5}$ concentrations. For O_3 prediction, although traditional models perform well in *MAE* and *RMSE*,

Table 8. Predictive Indicator Evaluation of Pollutant Concentrations After Model Selection Using PyCaret.

Predictive Indicator	Model	MAE	MASE	R^2	RMSE	RMSSE	SMAPE
SO_2	LR	**1.1832**	**1.2389**	**0.0011**	**1.4472**	**0.9911**	0.2198
	Ridge Regression	**1.1832**	**1.2389**	**0.0011**	**1.4472**	**0.9911**	0.2198
	BR	1.1834	1.2391	0.0003	1.4478	0.9915	0.2199
	ET	1.1909	1.2469	-0.0444	1.4825	1.0152	**0.219**
NO_2	Croston	**4.3823**	**0.5723**	**-0.0001**	**5.37**	**0.4944**	**0.1807**
CO	KNN	**0.1101**	**1.301**	**-0.5153**	**0.1228**	**0.9786**	0.1779
	CatBoost Regressor	0.111	1.3124	-0.5528	0.1243	0.9912	**0.1769**
	Decision Tree	0.113	1.3356	-1.0764	0.1377	1.0979	0.1782
	EN	0.1175	1.3889	-1.3775	0.1534	1.2232	0.1897
O_3	LightGBM	**6.834**	**0.3647**	**0.8124**	**7.971**	**0.3057**	0.9911
	CatBoost Regressor	8.4341	0.4501	0.6973	10.1192	0.3881	1.0742
PM_{10}	CatBoost Regressor	**12.8691**	**0.8996**	-0.1441	16.3046	0.8217	0.2018
	LightGBM	12.8924	0.9012	**-0.1065**	**16.0682**	**0.8098**	**0.2008**
$PM_{2.5}$	Decision Tree	**9.0692**	**1.3869**	-0.9615	11.9156	1.3196	**0.3239**
	GBR	9.4498	1.4451	**-0.3939**	**10.2785**	**1.1384**	0.3522

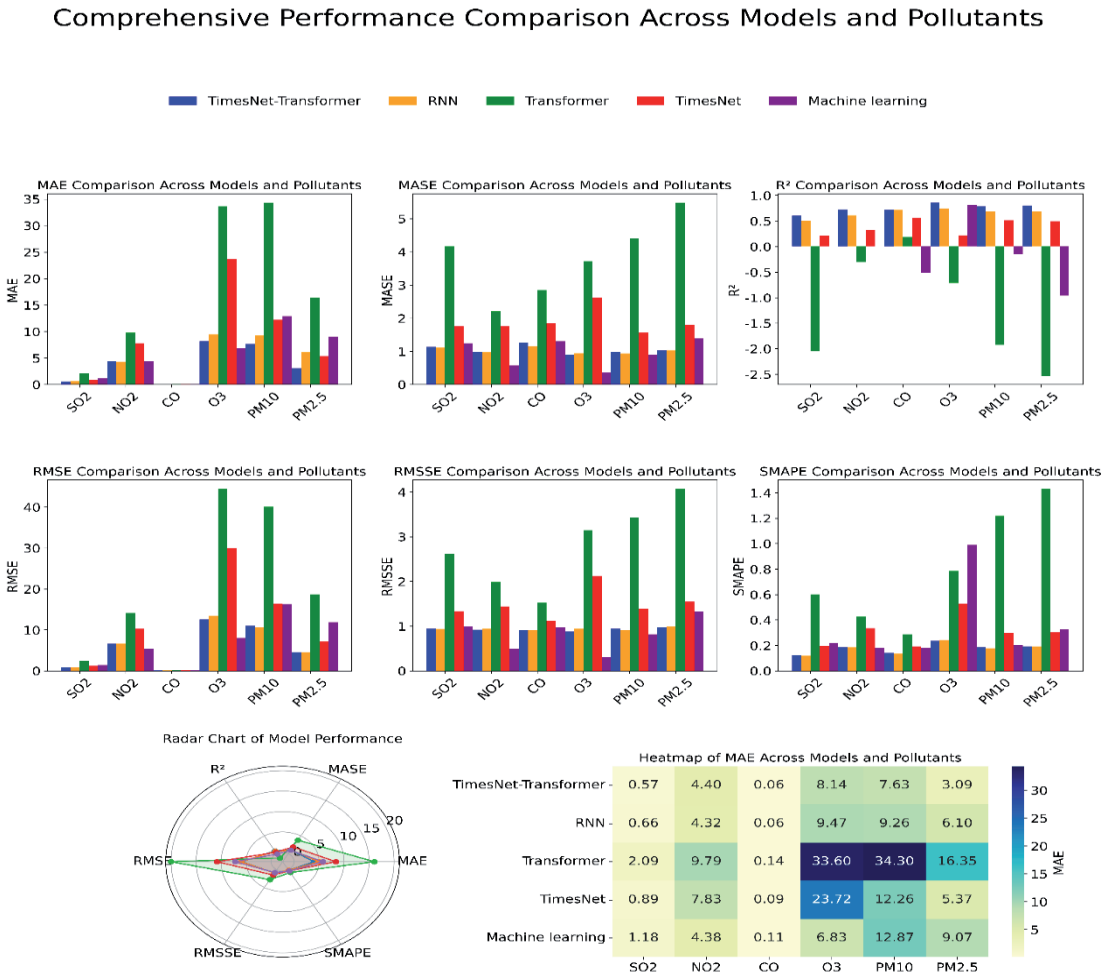

Fig. 7. Visualization of the predictive performance of different models.

the hybrid model excels in R^2, offering a better explanation of the complex variations in O_3, especially in long-term fluctuations.

Furthermore, while PyCaret provides automated modeling and optimization tools, it requires constructing multiple models and selecting the best, which can be time-consuming, especially for large datasets. In comparison, the hybrid model achieves a better balance between time cost and performance. Specifically, the hybrid model improves R^2 by 21.6%, 2%, 29.06%, and 40% for SO_2, PM_{10}, CO, and $PM_{2.5}$ predictions, respectively. Although traditional models for NO_2 and O_3 slightly outperform the hybrid model in some metrics, the hybrid model stands out in terms of R^2, offering superior performance overall.

Fig. 7 presents multiple visualizations illustrating the overall performance of various models in predicting pollutant concentrations, including SO_2, NO_2, CO, O_3 PM_{10}, and $PM_{2.5}$. Based on multiple evaluation metrics such as *MAE*, *MASE*, R^2, *RMSE*, R*MSSE*, and *SMAPE*, the Transformer-TimesNet hybrid model consistently demonstrates lower prediction errors across most pollutants, particularly excelling in *MAE* and *RMSE*. This highlights its superior accuracy and stability in pollutant concentration forecasting. In contrast, standalone models such as RNN, Transformer, and machine learning approaches (e.g., CatBoost and LightGBM) exhibit suboptimal performance for certain pollutants, such as PM_{10} and $PM_{2.5}$. The heatmap and radar chart further illustrate the disparities in model performance across different evaluation metrics, reinforcing the advantages of the Transformer-TimesNet model in multi-pollutant forecasting.

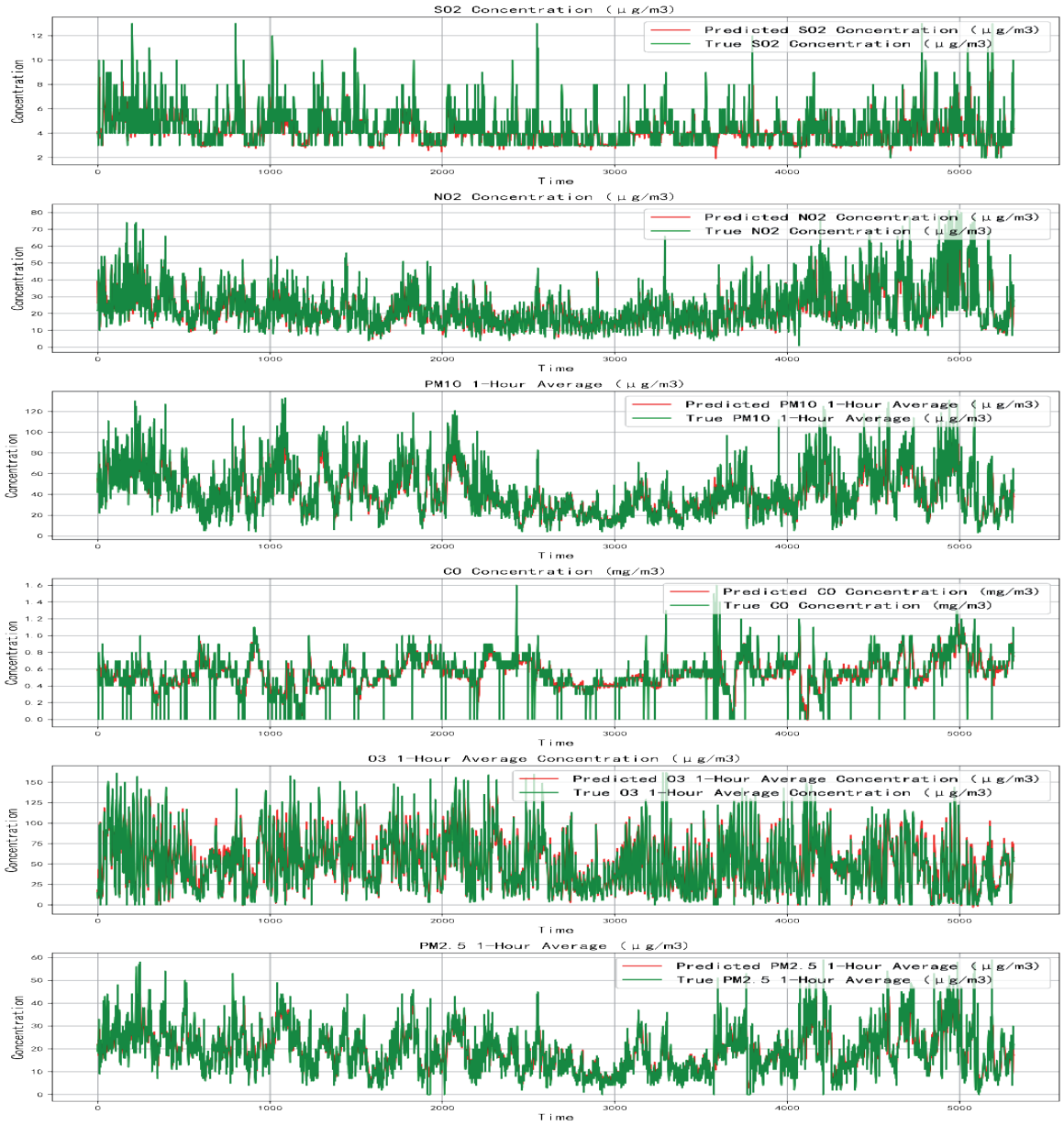

Fig. 8. Performance of Pollutant Concentrations on Test Set.

Model Prediction Results

The model demonstrates overall excellent performance, particularly in predicting NO_2 and PM_{10} concentrations (as shown in Fig. 8). The results for SO_2, NO_2, CO, and PM_{10} show that the model effectively captures the trends and seasonality of the data with high accuracy in predicting peak values. In particular, the model exhibits strong trend-capturing ability in the long-term predictions of NO_2 and PM_{10}, accurately reflecting the fluctuations of these pollutants at different time scales.

However, the prediction of SO_2 concentration is slightly less accurate than other pollutants. While the model successfully captures the overall periodic changes of SO_2, its accuracy in predicting extreme values is somewhat lacking. This may be due to the sharp fluctuations in SO_2 levels and the sporadic industrial emissions, especially during peak periods, which cause significant volatility in the data, making the prediction more challenging. Additionally, SO_2 concentrations are strongly influenced by short-term industrial activities, which often result in sudden variations that are challenging for traditional forecasting models to fully capture. These inherent characteristics of SO_2 dynamics highlight the need for further refinement in the modeling approach to better address short-term and high-volatility scenarios.

For O_3 and $PM_{2.5}$ predictions, despite the influence of multiple external factors (e.g., photochemical reactions and atmospheric circulation), the model still exhibits strong trend-capturing capabilities. In particular, $PM_{2.5}$ shows high robustness in capturing cyclical fluctuations, indicating that the model can handle complex time-series fluctuations and provide reliable predictions. Notably, for longer time scales, the model accurately reflects the changing trends of $PM_{2.5}$, suggesting that it can effectively manage the complexity of pollutant concentrations impacted by multiple factors.

While the hybrid Transformer-TimesNet model demonstrates substantial accuracy and trend detection advantages, its generalizability to other regions and pollutants remains a critical concern. Differences in local emission sources, meteorological conditions, and regulatory frameworks may limit the model's scalability without extensive retraining on region-specific datasets. Future research should explore transfer learning techniques to adapt the model to diverse environmental settings, ensuring broader applicability.

From a policy perspective, the predictive insights derived from this model have significant implications for urban air pollution mitigation strategies. In developing economies, where industrial emissions are a primary contributor to air pollution, targeted regulatory interventions – such as stricter emission standards and real-time monitoring – can leverage such forecasting models to implement proactive measures. For example, integrating predictive analytics with vertical decentralization in environmental governance (i.e., balancing local and central regulatory authority) can optimize pollution control policies, particularly in industrial zones where decentralized regulation has shown mixed effectiveness in reducing enterprise-level emissions [52].

Overall, while the Transformer-TimesNet model demonstrates superior performance in pollutant forecasting, addressing its scalability, volatility handling, and integration with policy-driven variables remains crucial. Future research should explore interdisciplinary approaches that combine predictive modeling, environmental economics, and regulatory frameworks to enhance air pollution management.

Conclusions

This study developed and validated a hybrid Transformer-TimesNet model for predicting industrial pollutant levels in the Xinyang industrial zone. The model significantly outperforms widely used forecasting models (including standalone deep learning models, traditional statistical methods, and machine learning models) on various evaluation metrics. This demonstrates a superior ability to capture long-term trends in pollutant concentrations. The model significantly improved prediction accuracy, with all performance metrics showing varying degrees of improvement compared to the baseline models.

Theoretical Implications

This research enhances time series forecasting in environmental monitoring by integrating Transformer and TimesNet models. Unlike traditional approaches that often struggle with capturing non-linear patterns, our hybrid model effectively learns multi-scale dependencies in air pollution data. The model's success underscores the transformative potential of advanced deep learning techniques in environmental monitoring and management, further contributing to the expanding body of research on AI-driven solutions in environmental science.

Managerial and Policy Implications

The improved forecasting accuracy has significant implications for environmental regulation and urban planning. Accurate predictions enable policymakers to implement more effective pollution control measures, optimize industrial emission strategies, and enhance public health interventions [53]. The model's application in early warning systems can help mitigate health risks associated with air pollution, especially in industrialized and developing regions.

Research Limitations and Future Directions

While promising, this study has certain limitations. The model was tested exclusively in Xinyang, and its applicability to other industrial zones requires further investigation. Future research should focus on validating the model across diverse regions, incorporating real-time monitoring data, and optimizing computational efficiency for practical deployment. Additionally, extending the model to predict a broader range of pollutants, such as volatile organic compounds (VOCs) and greenhouse gases (e.g., CO_2 and CH_4), would further enhance its effectiveness in environmental monitoring.

Additionally, incorporating external factors such as meteorological conditions and economic activities could improve prediction accuracy. Explainable AI techniques should also be explored to enhance model transparency for policymakers. Cross-regional validation and integration with geospatial analysis could support more data-driven urban planning and environmental management.

In summary, this study highlights the potential of advanced hybrid deep learning models for predicting industrial air pollutants, offering valuable insights for both theoretical research and practical applications. Future research should enhance the model's adaptability, interpretability, and scalability, ensuring its broader application in environmental monitoring and public health management.

Acknowledgments

This research was funded by the Researchers Supporting Project, number GKY-2024BSQDW-6, Guangdong University of Science and Technology, Dongguan, China; Project number 2024STY118, China Business Statistics Society; and Project number GD2023SKFC18, Guangdong Federation of Social Sciences.

Conflict of Interest

The authors declare no conflict of interest.

References

1. CHEN S., OLIVA P., ZHANG P. Air pollution and mental health: evidence from China. AEA Papers and Proceedings, **114**, 423, **2024**.
2. KAUR J., PARMAR K.S., SINGH S. Autoregressive models in environmental forecasting time series: a theoretical and application review. Environmental Science and Pollution Research, **30** (8), 19617, **2023**.
3. LUO J., GONG Y. Air pollutant prediction based on ARIMA-WOA-LSTM model. Atmospheric Pollution Research, **14** (6), 101761, **2023**.
4. NATH P., SAHA P., MIDDYA A.I., ROY S. Long-term time-series pollution forecast using statistical and deep learning methods. Neural Computing and Applications, 1, **2021**.
5. GARG N., SONI K., SAXENA T.K., MAJI S. Applications of Autoregressive integrated moving average (ARIMA) approach in time-series prediction of traffic noise pollution. Noise Control Engineering Journal, **63** (2), 182, **2015**.
6. GREKOV A.N., VYSHKVARKOVA E.V., MAVRIN A.S. Forecasting and Anomaly Detection in BEWS: Comparative Study of Theta, Croston, and Prophet Algorithms. Forecasting, **6** (2), 343, **2024**.
7. MÉNDEZ M., MERAYO M.G., NÚÑEZ M. Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, **56** (9), 10031, **2023**.
8. SETHI J.K., MITTAL M. An efficient correlation based adaptive LASSO regression method for air quality index prediction. Earth Science Informatics, **14** (4), 1777, **2021**.
9. OBIDALLAH W.J. Artificial intelligence modeling and simulation of membrane-based separation of water pollutants via ozone Process: Evaluation of separation. Thermal Science and Engineering Progress, **51**, 102627, **2024**.
10. JI Y., ZHI X., WU Y., ZHANG Y., YANG Y., PENG T., JI L. Regression analysis of air pollution and pediatric respiratory diseases based on interpretable machine learning. Frontiers in Earth Science, **11**, 1105140, **2023.**
11. ZHANG X., DING C., WANG G. An Autoregressive-Based Kalman Filter Approach for Daily PM2. 5 Concentration Forecasting in Beijing, China. Big Data, **12** (1), 19, **2024**.
12. KOTHANDARAMAN D., PRAVEENA N., VARADARAJKUMAR K., MADHAV RAO B., DHABLIYA D., SATLA S., ABERA W. Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology, **2022**, 5086622, **2022**.
13. HE S., WU J., WANG D., HE X. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere, **290**, 133388, **2022**.
14. ESSAMLALI I., NHAILA H., EL KHAILI M. Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review. Sustainability, **16** (3), 976, **2024**.
15. KE G., MENG Q., FINLEY T., WANG T., CHEN W., MA W., YE Q., LIU T.Y. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, **30**, **2017**.
16. DOROGUSH A.V., ERSHOV V., GULIN A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv, **1810**, 11363, **2018**.
17. MA J., YU Z., QU Y., XU J., CAO Y. Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: a case study of Shanghai. Aerosol and Air Quality Research, **20** (1), 128, **2020**.
18. WANG J., LU Y., XIN C., YOO C., LIU H. Kernel PLS with AdaBoost ensemble learning for particulate matters forecasting in subway environment. Measurement, **204**, 111974, **2022**.
19. CHAO B., GUANG QIU H. Air pollution concentration fuzzy evaluation based on evidence theory and the K-nearest neighbor algorithm. Frontiers in Environmental Science, **12**, 1243962, **2024**.

20. LIAO Q., ZHU M., WU L., PAN X., TANG X., WANG Z. Deep learning for air quality forecasts: a review. Current Pollution Reports, **6**, 399, **2020**.
21. BROWNLEE J. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery, **2018**.
22. ULPIANI G., DUHIRWE P.N., YUN G.Y., LIPSON M.J. Meteorological influence on forecasting urban pollutants: Long-term predictability versus extreme events in a spatially heterogeneous urban ecosystem. Science of the Total Environment, **814**, 152537, **2022**.
23. AGGARWAL A., TOSHNIWAL D. A hybrid deep learning framework for urban air quality forecasting. Journal of Cleaner Production, **329**, 129660, **2021**.
24. SARAVANAN D., KUMAR K.S. Improving air pollution detection accuracy and quality monitoring based on bidirectional RNN and the Internet of Things. Materials Today: Proceedings, **81**, 791, **2023**.
25. LUO J., ZHANG Z., FU Y., RAO F. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results in Physics, **27**, 104462, **2021**.
26. AHMED A.A.M., JUI S.J.J., SHARMA E., AHMED M.H., RAJ N., BOSE A. An advanced deep learning predictive model for air quality index forecasting with remote satellite-derived hydro-climatological variables. Science of The Total Environment, **906**, 167234, **2024**.
27. LIANG Y.C., MAIMURY Y., CHEN A.H.L., JUAREZ J.R.C. Machine learning-based prediction of air quality. Applied Sciences, **10** (24), 9151, **2020**.
28. TSOKOV S., LAZAROVA M., ALEKSIEVA-PETROVA A. A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. Sustainability, **14** (9), 5104, **2022**.
29. WU C.L., SONG R.F., ZHU X.H., PENG Z.R., FU Q.Y., PAN J. A hybrid deep learning model for regional O_3 and NO_2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network. Environmental Pollution, **320**, 121075, **2023**.
30. KIM J., WANG X., KANG C., YU J., LI P. Forecasting air pollutant concentration using a novel spatiotemporal deep learning model based on clustering, feature selection and empirical wavelet transform. Science of the Total Environment, **801**, 149654, **2021**.
31. ZHOU H., ZHANG S., PENG J., ZHANG S., LI J., XIONG H., ZHANG W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, **35** (12), 11106, **2021**.
32. REZA S., FERREIRA M.C., MACHADO J.J., TAVARES J.M.R. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Systems with Applications, **202**, 117275, **2022**.
33. WU N., GREEN B., BEN X., O'BANION S. Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv, **2001**, 08317, **2020**.
34. ZENG A., CHEN M., ZHANG L., XU Q. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, **37** (9), 11121, **2023**.
35. LIANG Y., XIA Y., KE S., WANG Y., WEN Q., ZHANG J., ZHENG Y., ZIMMERMANN R. Airformer: Predicting nationwide air quality in china with transformers. In Proceedings of the AAAI Conference on Artificial Intelligence, **37** (12), 14329, **2023**.
36. ZHANG Z., ZHANG S. Modeling air quality $PM_{2.5}$ forecasting using deep sparse attention-based transformer networks. International Journal of Environmental Science and Technology, **20** (12), 13535, **2023**.
37. AL-QANESS M.A., DAHOU A., EWEES A.A., ABUALIGAH L., HUAI J., ABD ELAZIZ M., HELMI A.M. ResInformer: residual transformer-based artificial time-series forecasting model for PM2. 5 concentration in three major Chinese cities. Mathematics, **11** (2), 476, **2023**.
38. WU H., HU T., LIU Y., ZHOU H., WANG J., LONG M. Timesnet: Temporal 2d-variation modeling for general time series analysis. arXiv preprint arXiv, **2210**, 02186, **2022**.
39. ZHANG X., YANG K., ZHENG L. Transformer Fault Diagnosis Method Based on TimesNet and Informer. Actuators, **13** (2), 74, **2024**.
40. ZHOU T., MA Z., WEN Q., WANG X., SUN L., JIN R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In International Conference on Machine Learning, 27268, **2022**.
41. ZUO C., WANG J., LIU M., DENG S., WANG Q. An ensemble framework for short-term load forecasting based on timesnet and tcn. Energies, **16** (14), 5330, **2023**.
42. SALMAN D., DIREKOGLU C., KUSAF M., FAHRIOGLU M. Hybrid deep learning models for time series forecasting of solar power. Neural Computing and Applications, **1**, **2024**.
43. LOSHCHILOV I. Decoupled weight decay regularization. arXiv preprint arXiv, **1711**, 05101, **2017**.
44. KINGMA D.P. Adam: A method for stochastic optimization. arXiv preprint arXiv, **1412**, 6980, **2014**.
45. AKIBA T., SANO S., YANASE T., OHTA T., KOYAMA M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623, **2019**.
46. SNOEK J., LAROCHELLE H., ADAMS R.P. Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, **25**, **2012**.
47. EMMERICH M.T., DEUTZ A.H. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. Natural Computing, **17**, 585, **2018**.
48. TAX D.M., DUIN R.P. Support vector data description. Machine Learning, **54**, 45, **2004**.
49. KHAN S.S., MADDEN M.G. One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review, **29** (3), 345, **2014**.
50. HUANG Y., ZHOU Z., WANG Z., ZHI X., LIU X. TimesNet-$PM_{2.5}$: Interpretable TimesNet for Disentangling Intraperiod and Interperiod Variations in PM2. 5 Prediction. Atmosphere, **1** (11), 1604, **2023**.
51. CHEUNG Y.W., LAI K.S. Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics, **13** (3), 277, **1995**.
52. WU Y., HU J., IRFAN M., HU M. Vertical decentralization, environmental regulation, and enterprise pollution: an evolutionary game analysis. Journal of Environmental Management, **349**, 119449, **2024**.
53. ELBAZ K., HOTEIT I., SHABAN W.M., SHEN S.L. Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM. Chemosphere, **313**, 137636, **2023**.