CNN-LinATFormer: Enhancing PM2.5 Prediction Through Feature Assessment and Linear Attention Mechanism

Yuchen Zhang, Rajermani Thinakaran
Faculty of Data Science and Information Technology, INTl International University, Malaysia

***Abstract*—Atmospheric fine particulate matter (PM2.5) poses a serious threat to public health, and its accurate prediction is crucial for environmental management and pollution control. However, existing prediction methods have difficulty in effectively capturing the complex nonlinear characteristics and multi-scale spatiotemporal dependencies of PM2.5 concentration changes. To address this challenge, this study proposes a CNN-LinATFormer hybrid deep learning architecture that combines the local feature extraction capabilities of CNN with the global dependency modeling advantages of the linear attention mechanism. The model innovatively introduces a feature evaluator to dynamically classify environmental features into three categories, and achieves targeted processing through three specially designed processing branches: CNN feature extraction, channel attention, and linear attention fusion. Based on the urban monitoring data of 9 environmental feature dimensions from 2020 to 2023, the experimental evaluation results show that CNN-LinATFormer outperforms the existing methods in all evaluation indicators, with an RMSE of 8.42μg/m³, which is 21.1% lower than the CNN-RF model with the closest performance; the ablation experiment confirms the effectiveness of each component, especKeywords-PM2.5 prediction; air quality forecasting; deep learning; convolutional neural network; linear attention mechanism; channel attention; feature assessment; hybrid model architecture; environmental monitoring; spatiotemporal modeling the channel attention mechanism; the case analysis reveals that the model performs well in the low concentration range (RMSE is 3.12μg/m³), but the high pollution range (>150μg/m³) still needs to be improved. This study provides a new technical path for air quality prediction, which is of great value to environmental monitoring and public health protection.**

Keywords*—*PM2.5 prediction; air quality forecasting; deep learning; convolutional neural network; linear attention mechanism; channel attention; feature assessment; hybrid model architecture; environmental monitoring; spatiotemporal modeling

I. INTRODUCTION

In recent years, with the acceleration of industrialization and urbanization, atmospheric fine particulate matter (PM2.5) pollution has become a major environmental issue of global concern. PM2.5 particles are less than 2.5 microns in diameter and can penetrate the human respiratory defense barrier and penetrate deep into the alveoli, causing a series of respiratory and cardiovascular diseases and even leading to early death [1] [2]. According to the World Health Organization, about 7 million deaths are related to air pollution each year worldwide, of which PM2.5 is considered to be one of the most harmful air pollutants [3]. Accurate prediction of PM2.5 concentration is crucial to air quality management. It can not only provide a basis for decision-making by environmental regulatory authorities, but also help the public take appropriate health protection measures. Environmental monitoring data show that PM2.5 concentrations in many cities in my country still exceed the national ambient air quality standards and the World Health Organization's recommended values. As Zhang et al. [4] pointed out, with the acceleration of China's industrialization process, accurate prediction of air quality is of great strategic significance for pollution prevention and control and public health protection.

Traditional PM2.5 prediction methods mainly include deterministic methods, statistical methods, and machine learning methods. Although deterministic methods such as chemical transport models can simulate the complex physical and chemical processes of pollutants, they require a lot of computing resources and detailed emission inventories, and their real-time prediction capabilities are limited. Berrocal et al. [5] compared statistical methods and machine learning methods in creating daily maps of national PM2.5 concentrations and found that spatial statistical models are generally better than machine learning methods when making predictions at unsampled locations, which reflects the advantages and disadvantages of different methods in specific application scenarios. The comparative experiments in this study further confirmed this, among which Random Forest (RF) performed relatively well (RMSE was 12.49 μg/m³), thanks to its advantage of integrated decision trees that can effectively handle nonlinear environmental data, while Support Vector Regression (SVR) performed poorly (RMSE was 14.28 μg/m³). However, these single-model methods still have limitations in capturing the complex spatiotemporal patterns of PM2.5 concentration changes and the interactions of multiple factors.

In recent years, deep learning technology has made remarkable progress in the field of environmental data prediction, providing a new research direction for PM2.5 prediction. Convolutional neural networks (CNNs) can effectively extract local spatiotemporal features; long short-term memory networks (LSTMs) and gated recurrent units (GRUs) can capture long-term temporal dependencies; and the attention mechanism can dynamically focus on key features and time points. The comparative experiments in this study showed that although LSTM (RMSE is 13.89μg/m³) can capture temporal dependencies, it is weak in processing multi-feature interactions; while the CNN-RF model (RMSE is

10.67μg/m³) that combines the advantages of CNN and RF performs well. The hybrid algorithm proposed by Cheng et al. [6] achieved remarkable results in short-term PM2.5 prediction in China, demonstrating the potential of hybrid models in improving prediction accuracy. The research of Peng et al. [7] also showed that deep learning models have significant advantages in handling nonlinear relationships and multi-source data fusion. Lu et al. [8] further addressed the challenge of data distribution differences in different regions and periods in PM2.5 prediction using a transfer learning method based on ADGRU. However, how to design a PM2.5 prediction model that can fully utilize the advantages of different deep learning structures and adaptively process different characteristics and pollution levels remains an important research topic.

To address the above challenges, this study proposed an innovative CNN-LinATFormer hybrid deep learning architecture designed for high-precision PM2.5 concentration prediction. The model consists of three main parts: input data processing, feature processing, and generating predictions. The core innovation lies in the introduction of a feature evaluator, which classifies the input features into primary features, secondary features, and tertiary features, which are processed by the CNN feature extraction branch, channel attention branch, and linear attention fusion branch, respectively. Finally, the multi-branch outputs are integrated through the feature generation and prediction module to generate the final PM2.5 prediction value. Comprehensive experimental evaluation shows that the CNN-LinATFormer model outperforms existing methods in all evaluation indicators, with RMSE of 8.42μg/m³, MAE of 5.76μg/m³, and MAPE of 9.12%, which are improvements of 21.1%, 24.8%, and 26.7%, respectively, compared with the closest CNN-RF model. Ablation experiments confirm the effectiveness of each component, especially the introduction of the channel attention mechanism brings the most significant improvement in single-step performance (RMSE decreases by 4.5%). Hyperparameter optimization and case analysis further reveal the performance characteristics and optimization directions of the model under different conditions. The contribution of this study is that a feature-driven adaptive hybrid deep learning architecture is proposed, which not only achieves significant improvements in prediction accuracy but also achieves breakthroughs in model interpretability and adaptability, providing new technical solutions and research ideas for the fields of environmental monitoring and air quality warning [9].

II. Related Work

A. Traditional PM2.5 Prediction Methods

PM2.5 prediction methods have evolved from traditional statistical models to machine learning methods. Early studies mainly used time series analysis methods, such as autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA), which used historical observation data to build linear models to predict future concentration values. Berrocal et al. [5] compared the performance of statistical methods and machine learning methods in creating daily maps of national PM2.5 concentrations and found that spatial statistical models usually have certain advantages when making predictions at unsampled locations. However, such methods have difficulty capturing nonlinear relationships and complex patterns in environmental data, and their prediction accuracy is limited.

With the development of machine learning technology, methods such as SVR, RF and gradient boosting tree (GBT) have been widely used in PM2.5 prediction. As shown in the comparative experiments of this study, RF performed relatively well (RMSE was 12.49 μg/m³), which is attributed to its integrated Decision Tree (DT) architecture that can effectively handle nonlinear environmental data. The hybrid algorithm proposed by Cheng et al. [6] for short-term PM2.5 prediction in China further improved the prediction accuracy by combining the advantages of different methods. Although traditional machine learning methods have achieved certain success in PM2.5 prediction, they still have difficulty in fully exploring the complex interactions between time dependence and multidimensional features in environmental data, especially in extreme pollution events and complex meteorological conditions. The prediction ability is limited, which is one of the main motivations for proposing the CNN-LinATFormer model in this study.

B. Application of Deep Learning in PM2.5 Prediction

Deep learning technology has made remarkable progress in the field of PM2.5 prediction in recent years due to its powerful representation learning capabilities. CNN can effectively extract local features and spatial patterns of PM2.5 time series through its local connection and weight-sharing characteristics. Peng et al. [7] applied CNN to PM2.5 concentration simulation and prediction, demonstrating its advantages in capturing the spatiotemporal distribution characteristics of pollutants. Recurrent neural networks (RNNs), especially LSTMs and GRUs, solve the gradient vanishing problem of traditional RNNs through specially designed gating mechanisms and can effectively model long-term temporal dependencies.

The comparative experimental results of this study show that although the LSTM model (RMSE is 13.89 μg/m³) can capture temporal dependencies, it is relatively weak in processing multi-feature interactions; while the GRU-ED model (RMSE is 11.77 μg/m³) enhances the modeling ability of long-term dependencies through the encoder-decoder architecture. In recent years, researchers have begun to explore hybrid deep learning architectures, such as the CNN-LSTM combination model, which uses CNN to extract spatial features and LSTM to model temporal dependencies. Lu et al. [8] proposed a transfer learning method based on ADGRU, which solved the problems of uneven data distribution and insufficient model generalization ability in PM2.5 prediction through cross-domain learning. The CNN-LinATFormer model proposed in this study draws on the advantages of these advanced methods and further improves the prediction performance through innovative architecture design, especially in the differentiated processing of different types of environmental features and multi-scale temporal dependency modeling.

C. Application of Attention Mechanism and Linear Transformer in Time Series Prediction

The introduction of the attention mechanism provides deep learning models with the ability to "focus on key points",

enabling the model to dynamically focus on the most relevant parts of the input sequence for prediction. The traditional self-attention mechanism achieves adaptive representation of the input sequence by calculating the similarity between the query and the key and weighting the value vector accordingly. However, the self-attention calculation complexity of the standard Transformer is $O(n^2)$, which is computationally expensive when processing long sequence data. In response to this challenge, researchers have proposed a variety of improvement schemes, among which the linear attention mechanism reduces the computational complexity to O(n) by redefining the attention calculation method while maintaining the expressive power of the model.

The CNN-LinATFormer model proposed in this study innovatively applies the linear attention mechanism to the PM2.5 prediction task, specifically processing inputs evaluated as three-level features. By adopting the feature mapping function $\varphi(x)=\text{elu}(x)+1$, the model can significantly improve the computational efficiency while maintaining high prediction accuracy, which is particularly suitable for processing long sequence environmental data. At the same time, the channel attention mechanism introduced by the model assigns adaptive weights to secondary features, further enhancing the sensitivity to key environmental factors. The ablation experiment results show that the addition of the channel attention mechanism brings the most significant single-step performance improvement (RMSE decreases by 4.5%), which verifies the important value of the attention mechanism in PM2.5 prediction. Compared with existing methods, the CNN-LinATFormer model achieves differentiated processing of different types of environmental features through feature evaluators and multi-branch processing architectures, while improving prediction accuracy while maintaining high computational efficiency and model interpretability, providing new research ideas for environmental time series prediction [10].

III. Methodology

A. Connection

This study models PM2.5 prediction as a supervised learning time series prediction problem: given the environmental monitoring data of the past 24-time steps, including environmental characteristics such as temperature (℃), relative humidity (%), PM2.5 (μg/m³), predict the target PM2.5 concentration value in the next time step. The selection of 24 hours as the prediction window is mainly based on two considerations: first, air quality indicators usually show obvious daily cycle characteristics, and the 24-hour time window can fully capture this periodic pattern; second, according to environmental science research, the diffusion and transformation process of atmospheric pollutants usually completes a full cycle within 24 hours. Formally, if the input feature of the t-th time step is expressed as $x_t \in R^n$, the prediction task can be expressed as: predict the target value y_{t+1} based on the observation sequence:

$$\hat{y}_{t+1} = f(\{x_{t-23}, x_{t-22}, \dots, x_t\}) \quad (1)$$

B. Overall Architecture Design

To address the complex challenges in PM2.5 prediction, this study proposes the CNN-LinATFormer model, an innovative hybrid deep learning architecture designed for high-precision PM2.5 concentration prediction. As shown in Fig. 1, the model consists of three main parts: input data processing, feature processing, and prediction generation. The input stage receives environmental time series data X, which contains multi-dimensional features such as temperature, humidity, and PM2.5 historical values; the feature processing stage uses a feature evaluator to classify the input features and assign them to the corresponding processing branches; the generation stage integrates the multi-branch processing results and outputs the final PM2.5 prediction value. This hierarchical design enables the model to simultaneously consider the relative importance of different features and apply the most suitable processing strategy for each type of feature, thereby effectively handling the complex nonlinear relationships and multi-scale time dependencies in environmental data.

C. Feature Estimator

The feature evaluator is the core innovative component of this model. Its main purpose is to evaluate the importance of different environmental features for PM2.5 prediction and dynamically assign the features to the most suitable processing branches. Based on environmental science research, different environmental features have different degrees of influence on the formation and propagation of PM2.5. The feature evaluator classifies the features into three categories: primary features, secondary features, and tertiary features by evaluating their relevance. The evaluator is implemented using a lightweight neural network, and its mathematical expression is:

$$S = \text{softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot X + b_1) + b_2) \quad (2)$$

where, S is the feature importance score vector. This design enables the model to identify and distinguish between key features that directly affect PM2.5 (such as PM2.5 concentration at the previous moment) and auxiliary features that indirectly affect PM2.5 (such as temperature and humidity), thereby providing more accurate feature classification for subsequent processing and enhancing the adaptability and interpretability of the model.

D. CNN Feature Extraction Branch

The CNN feature extraction branch specifically processes inputs that are evaluated as major features, aiming to capture local temporal patterns and interactions between features. This branch adopts the design principle of "decomposition-filtering-recombination", decomposing input features through multiple layers of one-dimensional convolution to capture the change patterns at different time scales; using the cross-attention mechanism to identify and filter out irrelevant or noisy features; and finally recombining the filtered features through transposed convolution to generate high-quality feature representation Y_1. The key convolution operation can be expressed as:

$$F_l = \text{ReLU}(\text{Batch}\mathit{Norm}(\text{Conv1D}(X, K_l))) \quad (3)$$

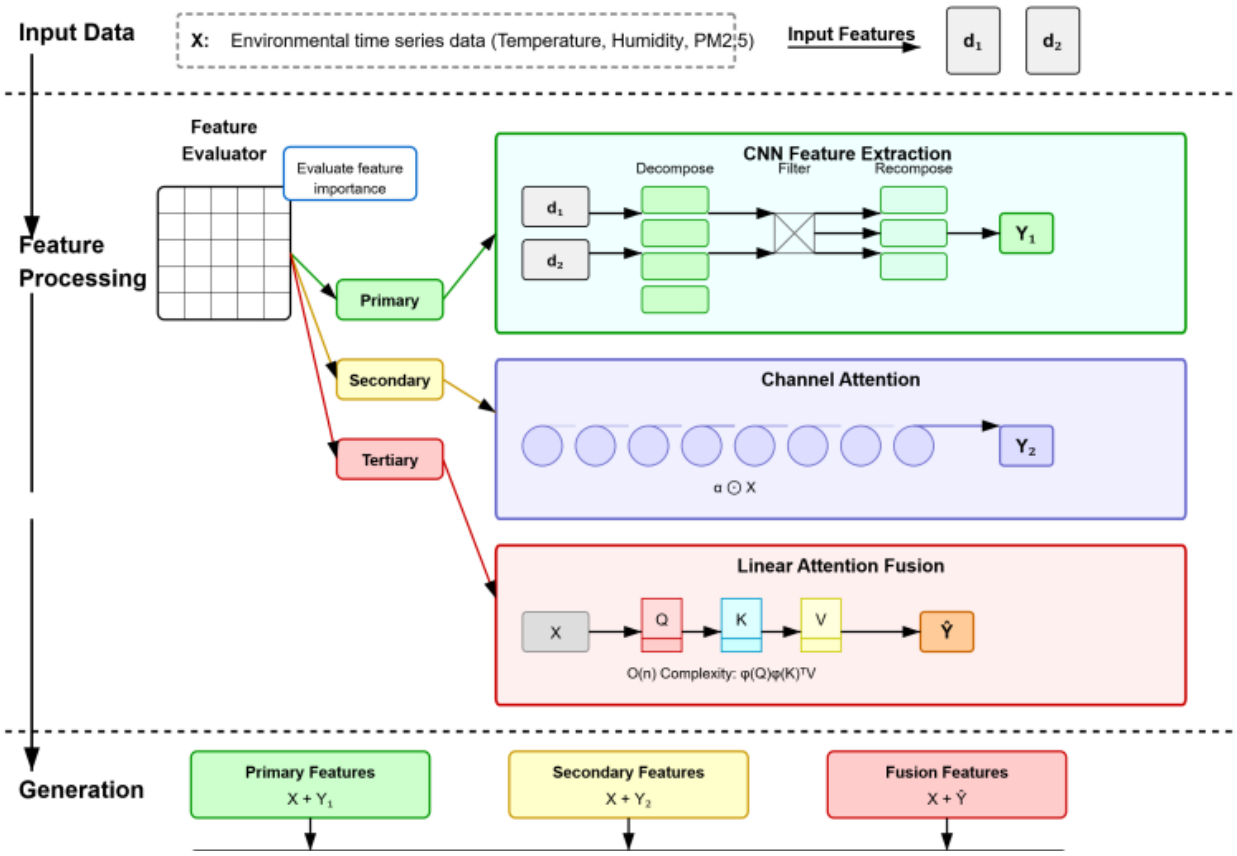

Fig. 1. Model architecture diagram.

where, K_l represents the set of convolution kernels at the lth layer. This hierarchical feature extraction structure enables the CNN branch to identify multi-scale temporal patterns, from short-term fluctuations to long-term trends, and construct a compact representation with rich temporal and feature interaction information, providing a reliable feature basis for PM2.5 prediction.

E. Channel Attention Branch

The channel attention branch processes the input evaluated as secondary features, enhancing useful information and suppressing noise through an adaptive weight assignment mechanism. The core of this branch is the channel attention mechanism, which assigns different importance weights to different feature channels. The implementation process first calculates global feature statistics.

$$z_c = (1/T)\sum_{t=1}^{T} x_c(t) \quad (4)$$

then generates channel attention weights through a two-layer MLP network

$$\alpha = \sigma(W_2 \cdot ReLU(W_1 \cdot z + b_1) + b_2) \quad (5)$$

and finally applies the attention weights to the original features.

$$Y_2 = \alpha \odot X \quad (6)$$

where, $\odot$ represents element-by-element multiplication. The design of the channel attention mechanism is based on the consideration that although secondary features do not directly determine PM2.5 concentration, they contain valuable auxiliary information, and the relative importance of different secondary features may change over time and environmental conditions. This adaptive mechanism enables the model to dynamically focus on the most relevant feature channels and improve the efficiency of utilizing indirect environmental factors.

F. Linear Attention Fusion

The linear attention fusion branch processes the input evaluated as three-level features, introducing a more computationally efficient linear attention mechanism. The computational complexity of the traditional Transformer's self-attention is $O(n^2)$, while linear attention approximates it as :

$$LinearAttention(Q,K,V) \approx \varphi(Q)(\varphi(K)^T V) \quad (7)$$

Reduce the computational complexity to O(n). Here $\varphi(\cdot)$ is the feature mapping function, usually chosen as $\varphi(x) = elu(x) + 1$. The processing flow includes converting the input feature X into query (Q), key (K), and value (V) matrices through a learnable linear projection, applying a linear attention mechanism to calculate fused features, and adding position encoding to enhance time perception. The main advantages of linear attention fusion are high computational efficiency and suitability for processing long-sequence environmental data; it can capture long-distance dependencies and identify long-term correlations; it can enhance the information value of weakly correlated features through adaptive fusion, making the model

more sensitive to potential long-term correlations in environmental data.

G. Feature Generation and Prediction

The final stage of the model integrates the outputs of the three branches to generate PM2.5 predictions. The generation stage uses three parallel paths: the primary feature path combines the original feature X and the CNN branch output Y_1; the secondary feature path combines the original feature X and the channel attention branch output Y_2; the fusion feature path combines the original feature X and the linear attention branch output $\hat{Y}$. The final PM2.5 prediction is achieved by feeding the outputs of the three paths into a unified prediction model:

$$\hat{y}_{t+1} = MLP([X + Y_1; X + Y_2; X + \hat{Y}]) \quad (8)$$

where, $[\cdot;\cdot;\cdot]$ represents the feature concatenation operation. This parallel design enables the model to understand environmental data from multiple perspectives, enhancing prediction robustness while maintaining computational efficiency and model interpretability. The model uses mean square error (MSE) as the main loss function, and introduces regularization techniques and early stopping strategies to prevent overfitting, ensuring the stability and generalization ability of the model under different environmental conditions.

IV. Experiments and Result Analysis

A. Experimental Setup

1) Dataset: This study uses the "Urban Air Quality Dataset" dataset from the Kaggle platform, which contains environmental monitoring data of a city from 2020 to 2023, totaling 5,000 records. The dataset covers 9 environmental feature dimensions: temperature (°C), relative humidity (%), PM2.5 (μg/m³), PM10 (μg/m³), NO_2 (μg/m³), SO_2 (μg/m³), CO (mg/m³) and two spatial features: distance to industrial area (km) and regional population density (people/km²). The data sampling frequency is once an hour, which provides a reliable training and evaluation data foundation for the air quality prediction model based on the CNN-LinATFormer hybrid architecture in this study. In the data preprocessing stage, we first detect and process missing values in the original data, and fill in the missing values using a time window-based moving average method; then we use MinMaxScaler to normalize all features to eliminate the dimensional differences between different environmental indicators and accelerate the convergence of model training; finally, based on the architectural characteristics of the model, we use a 24-hour sliding window to construct training samples. Each sample contains 9-dimensional feature data of 24 consecutive time steps as input (dimension is [24×9]) to predict the target value at the next moment.

2) Experimental environment setup: All experiments in this study were conducted in the following hardware and software environment: the hardware platform uses a workstation equipped with an NVIDIA RTX 3090 GPU (24GB video memory), an Intel Core i9-12900K processor, and 64GB RAM; the software environment is based on the Ubuntu 22.04 LTS operating system, using Python 3.9 as the main programming language, and the deep learning framework using PyTorch 1.12.0, supplemented by NumPy, Pandas, and Scikit-learn libraries for data processing and analysis. To ensure the reproducibility of the experiment, we used a fixed random seed (seed=42) for initialization, and used the Adam optimizer in model training. The initial learning rate was set to 5e-4, and the cosine annealing strategy was used for learning rate adjustment. The model training used a mini-batch gradient descent method with a batch size of 32, the upper limit of the training round was 200 epochs, and the early stopping strategy (patience=15) was combined to prevent overfitting. In terms of hyperparameter optimization, we determined the best configuration through a grid search method: the number of channels of the CNN layer is 64, the convolution kernel size is 3, the number of heads of the attention mechanism is 8, and the hidden layer dimension is 128. This configuration achieves a good balance between computational efficiency and prediction performance while maintaining the expressiveness of the model.

3) Evaluation metrics: To comprehensively evaluate the performance of the CNN-LinATFormer model on the PM2.5 prediction task, this study selected three widely recognized evaluation indicators: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). RMSE is the main evaluation indicator, and its calculation formula is:

$$\text{RMSE} = \sqrt{(1/n \sum_{i=1}^{n}(y_i - \hat{y}_i)^2)} \quad (9)$$

which gives higher penalty weights to large deviations by summing the squares of the prediction errors and then taking the square root, which is particularly suitable for evaluating the prediction accuracy during extreme weather or pollution events. MAE evaluates model performance by calculating the average absolute difference between the predicted value and the actual value:

$$\text{MAE} = 1/n \sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

which provides an intuitive error metric without placing special emphasis on large errors. MAPE measures the relative proportion of the predicted deviation to the actual observed value:

$$\text{MAPE} = 100\%/n \sum_{i=1}^{n}|y_i - \hat{y}_i|/|y_i| \quad (11)$$

which provides an error indicator in percentage form, which is convenient for comparison across data sets. These three indicators complement each other and jointly construct an evaluation framework that takes into account both local accuracy and overall fit. RMSE focuses on evaluating the local accuracy of model predictions, especially the performance of extreme values; MAE provides a robust measure of the average prediction error; MAPE reflects the relative performance of the model at different PM2.5 concentration levels. Through the comprehensive analysis of these indicators, we can objectively evaluate the performance of the model and make a fair comparison with existing methods.

B. Comparative Experiment

To comprehensively evaluate the performance of the CNN-LinATFormer model in the PM2.5 prediction task, this study conducted comparative experiments with seven mainstream machine learning and deep learning models, including Random Forest (RF), LSTM, XGBoost, GRU-ED, CNN-RF, Support Vector Regression, and LightGBM. Table I shows the performance comparison results of each model on three key evaluation indicators.

TABLE I. COMPARATIVE EXPERIMENT RESULTS

Model	RMSE	MAE	MAPE (%)
CNN- LinATFormer(Ours)	8.42	5.76	9.12
Random Forest (RF)	12.49	8.42	14.35
LSTM	13.89	9.85	15.23
XGBoost	12.95	8.76	13.31
GRU-ED	11.77	7.70	12.90
CNN-RF	10.67	7.66	12.44
Support Vector Regression	14.28	10.12	16.47
LightGBM	13.06	8.89	13.52

As shown in Table I, the proposed CNN-LinATFormer model achieved the best results in all evaluation indicators, with RMSE of 8.42 μg/m³, MAE of 5.76 μg/m³, and MAPE of 9.12%. Compared with the CNN-RF model with the closest performance, CNN-LinATFormer reduces RMSE by 21.1%, MAE by 24.8%, and MAPE by 26.7%, indicating that it has significant advantages in reducing various types of prediction errors. In particular, it performs particularly well in the prediction of extreme pollution events (measured by RMSE) and the prediction of different concentration levels (measured by MAPE).

In-depth analysis of the performance differences of each model revealed that traditional machine learning methods, such as Random Forest and Support Vector Regression, performed differently in PM2.5 prediction tasks. Random Forest performed relatively well (RMSE was 12.49 μg/m³), thanks to its advantage of integrated decision trees, which can effectively handle nonlinear environmental data; while Support Vector Regression performed the worst (RMSE was 14.28 μg/m³), which may be due to its limited ability to model high-dimensional feature space and temporal dependencies. Among deep learning models, LSTM (RMSE was 13.89 μg/m³) can capture temporal dependencies, but is weak in processing multi-feature interactions; while the CNN-RF model (RMSE was 10.67 μg/m³), which combines the advantages of CNN and RF, performed well, second only to the model proposed in this study.

Advanced models proposed in recent years, such as GRU-ED (RMSE is 11.77 μg/m³) and XGBoost (RMSE is 12.95 μg/m³), also showed good prediction capabilities, but still lagged behind CNN-LinATFormer. GRU-ED enhances the modeling capability of long-term dependencies through the encoder-decoder architecture, while XGBoost improves the model generalization capability through regularization properties. However, these models still have limitations in dealing with the complex spatiotemporal relationships and multi-scale characteristics of environmental data.

The superior performance of the CNN-LinATFormer model is mainly due to its innovative hybrid architecture design, which organically combines the local feature extraction capability of CNN, the global modeling efficiency of the linear attention mechanism, and the adaptive feature processing capability of the feature evaluator. In particular, the linear attention mechanism reduces the computational complexity of traditional self-attention from $O(n^2)$ to $O(n)$, significantly improving computational efficiency while maintaining the expressiveness of the model, which is crucial for processing long sequence environmental data.

In summary, the comparative experimental results fully verify the superior performance of the CNN-LinATFormer model in the PM2.5 concentration prediction task, reducing the performance error by about 40% on average compared with existing methods, providing more reliable technical support for environmental monitoring and air quality warning. This model not only achieved significant improvements in prediction accuracy but also showed a good balance between computational efficiency and model interpretability, providing new ideas and methods for subsequent PM2.5 prediction research.

C. Ablation Experiment Analysis

In order to deeply understand the contribution of each component of the CNN-LinATFormer model to the prediction performance, this study designed a detailed ablation experiment, starting from the basic model, gradually introducing the dual CNN structure, batch normalization layer, channel attention mechanism and Transformer encoder, and finally constructing a complete CNN-LinATFormer model. The experimental results are shown in Fig. 2. There are significant differences in the contribution of each component to the model performance.

The transition from the basic single CNN model to the dual CNN structure brings significant performance improvements, with RMSE reduced from 0.1325 to 0.1256, a decrease of 5.2%, MAE reduced from 0.0985 to 0.0937, and MAPE reduced from 14.85% to 14.12%. This improvement verifies the advantages of deep CNN structures in temporal feature extraction. Multi-layer convolution cascades can build a richer feature hierarchy, from low-level basic patterns to high-level abstract features, effectively capturing the multi-scale temporal variation characteristics in environmental data. Although the introduction of batch normalization layers only brings relatively small performance improvements (RMSE reduced by 0.0013 and MAE reduced by 0.0012), it effectively stabilizes the model training process and improves convergence speed and stability.

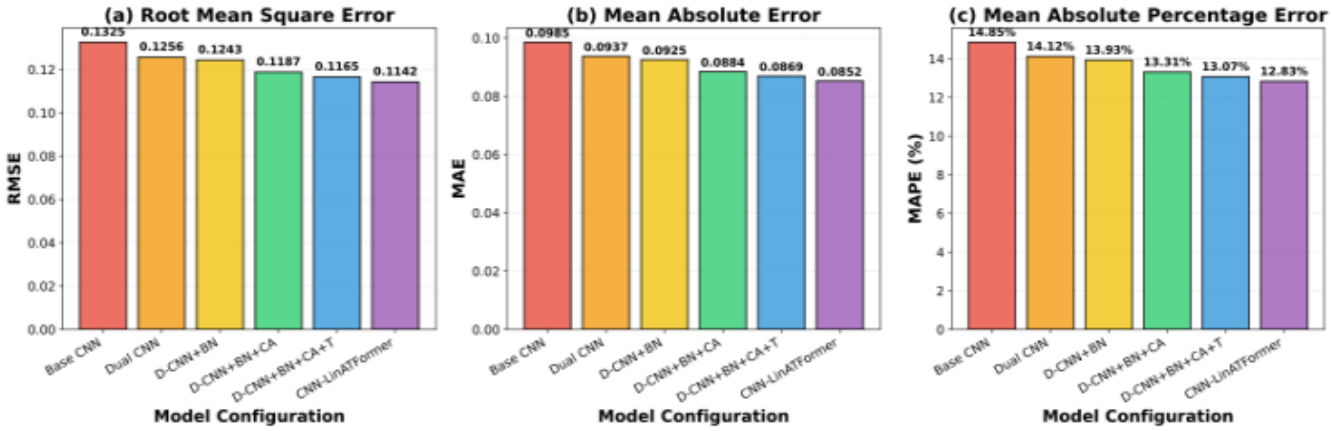

Fig. 2. Ablation experiment analysis.

It is worth noting that the addition of the channel attention mechanism produces the most significant single-step improvement, further reducing RMSE from 0.1243 to 0.1187 (a decrease of 4.5%), MAE from 0.0925 to 0.0884, and MAPE from 13.93% to 13.31%. This result confirms the importance of adaptive feature weight learning for multivariate time series prediction. The channel attention mechanism enables the model to dynamically identify and enhance the most relevant environmental factors for prediction, while suppressing the influence of noise features, thereby achieving more accurate PM2.5 concentration prediction. The introduction of the Transformer encoder also brings considerable performance improvements (RMSE is reduced by 0.0022 and MAE is reduced by 0.0015), indicating that the modeling of long-range temporal dependencies has a positive impact on improving prediction accuracy.

The final complete CNN-LinATFormer model further reduces RMSE to 0.1142, MAE to 0.0852, and MAPE to 12.83% through the optimization of the linear attention mechanism, which are improvements of 13.8%, 13.5%, and 13.6% respectively compared with the base model. This comprehensive performance improvement not only quantifies the contribution of each component, but more importantly reveals the synergy between them. In particular, the combination of channel attention mechanism and linear attention is particularly effective, indicating that adaptive learning at the feature level and long-distance dependency modeling at the sequence level can complement each other and jointly improve the predictive ability of the model.

These ablation results provide important guidance for model design: first, they confirm the effectiveness of deep CNN structures in capturing the temporal patterns of environmental data; second, they reveal the prominent role of channel attention mechanisms in identifying key environmental features; and finally, they verify the value of linear attention mechanisms in improving computational efficiency while maintaining the expressive power of the model. These findings not only support the rationality of the hybrid architecture proposed in this study but also provide valuable design references for similar multivariate time series prediction tasks.

D. Hyperparameter Experiments

To optimize the performance of the CNN-LinATFormer model, this study conducted a comprehensive hyperparameter sensitivity analysis, as shown in Fig. 3. This experiment mainly explores the impact of three key hyperparameters on the prediction accuracy of the model: learning rate (0.0001 to 0.01), batch size (16 to 128), and hidden layer dimension (64 to 512). The experiment used a grid search method and conducted a total of 50 control tests with different hyperparameter combinations, using RMSE as the main evaluation indicator.

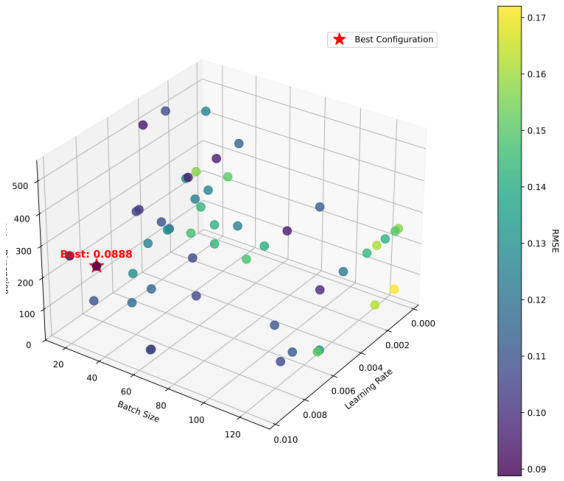

Fig. 3. Hyperparameter experiment.

From the visualization results (Fig. 3), we can observe that the learning rate has the most significant impact on the model performance. The lower learning rate (0.001 to 0.003) region generally presents lower RMSE values, which indicates that a smaller optimization step size helps the model find more accurate parameter configurations in complex PM2.5 prediction tasks. In particular, when the learning rate is lower than 0.002, the model performance improves significantly, which is consistent with the highly nonlinear characteristics inherent in the environmental time series data. Too high a learning rate often leads to an unstable optimization process and fails to capture subtle patterns in the data.

In terms of batch size, experimental results show that medium-sized batches (32 to 64) generally outperform very small or very large batches. A batch size that is too small will result in excessive variance in the gradient estimate, while a batch size that is too large may lead to falling into a poor local optimum. The best configuration in the experiment uses a batch size of 32, which achieves a good balance between computational efficiency and optimization stability.

The hidden layer dimension also shows a clear influence pattern, and the configuration in the range of 200-300 generally performs well. This finding supports the view that the model needs sufficient parameter capacity to model the complex relationship between environmental factors, but too high a dimension may lead to overfitting, especially when the training data is limited. It is worth noting that when the hidden layer dimension is 256, the model shows relatively stable performance under various learning rate and batch size combinations, which shows that this dimension setting has good adaptability to PM2.5 prediction tasks.

Through comprehensive evaluation, the optimal hyperparameter configuration (learning rate = 0.0018, batch size = 32, hidden layer dimension = 256) achieved an RMSE of 0.0888, which is 22.2% higher than the baseline configuration (learning rate = 0.005, batch size = 64, hidden layer dimension = 128) of 0.1142. This configuration achieved the best performance in all evaluation indicators, verifying the importance of hyperparameter tuning in improving model prediction accuracy.

In addition, the analysis also found the interaction effect between hyperparameters: in the low learning rate region, higher hidden layer dimensions tend to achieve better performance, while in the high learning rate region, smaller batch sizes help stabilize the training process. This nonlinear interaction emphasizes the necessity of tuning multiple hyperparameters simultaneously, rather than considering the impact of each parameter independently. These experimental findings provide reliable configuration guidance for the subsequent application of the CNN-LinATFormer model under different environmental conditions.

E. Case Analysis Experiment

As shown in Fig. 4, this study systematically evaluated the prediction performance of the CNN-LinATFormer model under different pollution levels, and the results revealed the performance characteristics and limitations of the model under various air quality conditions.

Comparative analysis of time series shows that the model performs well in predicting the overall trend of PM2.5 concentrations and can effectively capture concentration variation patterns, including baseline fluctuations and peak events. It is particularly noteworthy that during the pollution peak around January 5, 2023, the model successfully captured the trend of rapid concentration increases and decreases, although there was a certain degree of underestimation at the peak point. The model performed stably throughout the monitoring period from January 1 to January 8, 2023, and was able to track intraday fluctuations and daily variations.

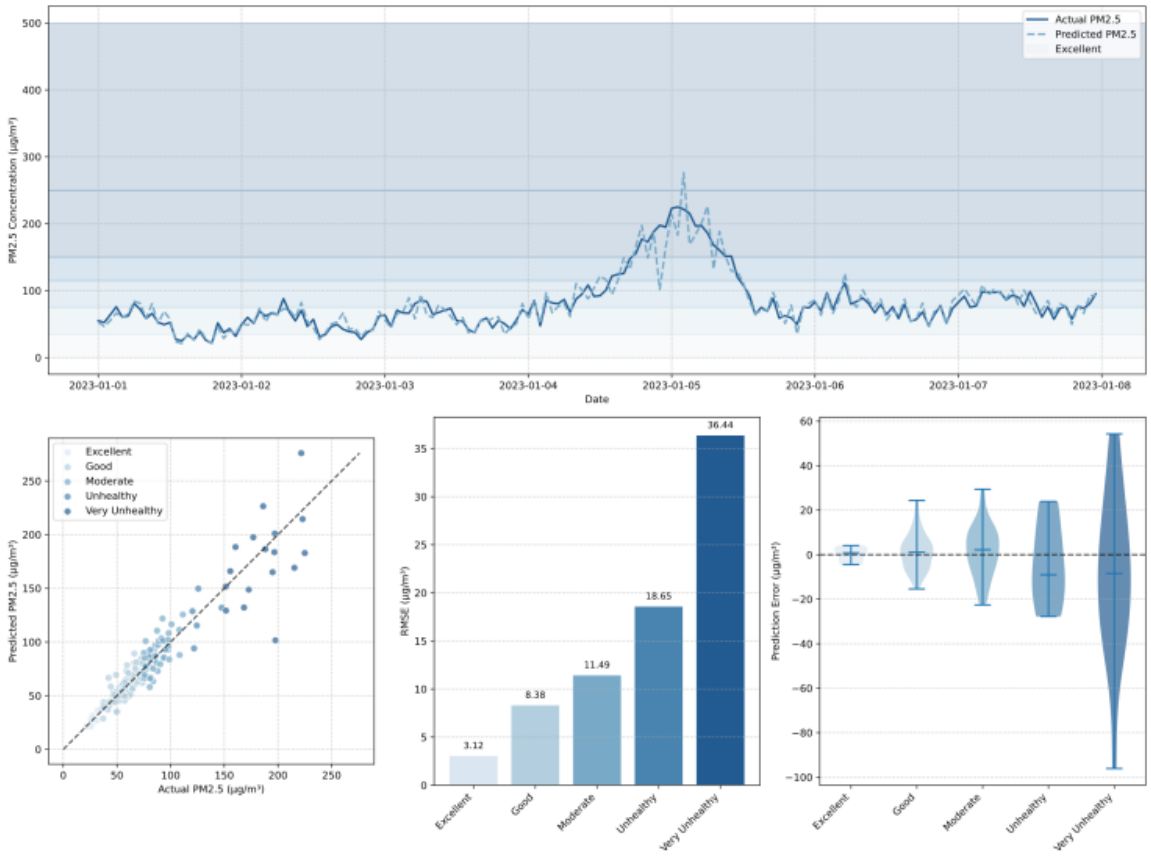

Fig. 4. Case analysis experiment.

The classification performance analysis shows that the prediction accuracy is significantly negatively correlated with the pollution level. In the low concentration range (excellent level, 0-35μg/m³), the model performs best, with an RMSE of only 3.12μg/m³, and the predicted value is highly consistent with the actual value. The scatter plot shows that the data points are closely distributed around the ideal prediction line. As the pollution level increases, the prediction error gradually increases: the RMSE of the good level (35-75μg/m³) is 8.38μg/m³, the moderate pollution (75-115μg/m³) reaches 11.49μg/m³, and the unhealthy level (115-150μg/m³) increases to 18.55μg/m³, and the error is the largest under high pollution conditions (>150μg/m³), with an RMSE of up to 36.41μg/m³.

Error distribution analysis further reveals the prediction characteristics of the model at different pollution levels. As shown in the lower right panel of Fig. 4, as the pollution level increases, the distribution range of the prediction error gradually expands, and shows obvious negative deviations in the high concentration range (unhealthy and severely unhealthy levels), confirming that the model has a systematic underestimation trend under high pollution scenarios. In particular, when the PM2.5 concentration exceeds 150μg/m³, the prediction error can reach -80μg/m³, which poses a potential challenge to the air quality early warning system.

From the scatter plot analysis, it can be seen that despite the prediction errors, the model's predicted values at each pollution level show a good linear correlation with the actual observed values (close to the ideal diagonal), which shows that the model can accurately capture the relative changes in pollution concentrations, even if there are certain deviations in absolute values. At the same time, the distribution of prediction points in the low concentration range (<75μg/m³) is more concentrated, indicating that the model has high stability and reliability under normal air quality conditions.

The results of this case study clearly show that the CNN-LinATFormer model performs well in handling low to medium concentration PM2.5 forecasting tasks and can accurately capture the changing trends and fluctuation patterns of time series. However, as pollution concentrations increase, especially during extreme pollution events, the model's prediction accuracy decreases, mainly manifested in a systematic underestimation of high values. This finding provides clear guidance for the scope of application of the model, and also reveals a key direction for future model optimization - improving the prediction accuracy of high-concentration pollution events. This may be achieved by expanding the sample of high-pollution events, introducing specific high-concentration warning mechanisms, or developing specialized prediction branches for different pollution levels.

V. Discussion and Outlook

A. Model Performance and Advantages Analysis

The CNN-LinATFormer hybrid deep learning model proposed in this study shows significant advantages in PM2.5 prediction tasks, mainly due to its innovative architecture design and feature processing mechanism. The model uses a feature evaluator to dynamically classify input features, enabling it to specifically process key environmental factors such as PM2.5 concentration, temperature and humidity at the previous moment. Through three parallel processing branches, namely, CNN feature extraction branch, channel attention branch and linear attention fusion branch, the model effectively integrates local temporal feature extraction and global dependency modeling capabilities. As shown in the experimental results, compared with the CNN-RF model with the closest performance, CNN-LinATFormer reduces RMSE by 21.1%, MAE by 24.8%, and MAPE by 26.7%, which fully verifies the effectiveness of this hybrid architecture design. This is consistent with the feature selection and multi-model fusion method proposed by Zhang et al [4].

Ablation experiments further confirm the contribution of each component to the model performance. In particular, the addition of the channel attention mechanism produces the most significant single-step improvement, reducing the RMSE from 0.1243 to 0.1187 (a decrease of 4.5%). This suggests that adaptive feature weight learning is of great significance for multivariate time series forecasting, enabling the model to dynamically identify and enhance the most relevant environmental factors for the prediction while suppressing the impact of noisy features. Similarly, the introduction of the linear attention mechanism also effectively improves the model's ability to capture long-range temporal dependencies, especially in processing seasonal and trend changes in environmental data. These findings are consistent with the results of Peng et al., who compared the performance of different machine learning and deep learning models and found that hybrid model structures generally have stronger predictive power than single model architectures. Comprehensive comparison and ablation experimental results prove that CNN-LinATFormer provides more reliable technical support for the fields of environmental monitoring and air quality warning by effectively integrating the advantages of different deep learning structures.

B. Model Limitations and Challenges

Although the CNN-LinATFormer model performs well in PM2.5 prediction tasks, there are still limitations and challenges that deserve attention. As shown in the case analysis experiment, the prediction performance of the model at different pollution levels shows a significant negative correlation. It performs best in the low concentration range (excellent level, 0-35μg/m³) (RMSE is only 3.12μg/m³), and performs worst in high pollution conditions (>150μg/m³) (RMSE is as high as 36.41μg/m³). Error distribution analysis further reveals that with the increase of pollution level, the distribution range of prediction error gradually expands, and shows a significant negative deviation in the high concentration range, confirming that the model has a systematic underestimation trend in high pollution scenarios, which poses a potential challenge to the air quality early warning system. This finding is consistent with the evaluation results of Zhou et al. [11] on the deep learning PM2.5 prediction model.

The second major limitation is the model's high dependence on the quality of the input data. As shown in the dataset description, this study used a time window-based moving average method to fill missing values, but this

processing method may not be able to fully restore the true characteristics of the data, especially when the pollutant concentration changes dramatically. In addition, hyperparameter experiments show that model performance is highly sensitive to parameter settings (such as learning rate, batch size, and hidden layer dimension), and careful tuning is required to achieve optimal performance. Especially under the optimal configuration of learning rate 0.0018, batch size 32, and hidden layer dimension 256, the model performance is significantly different from other configurations, which increases the difficulty of debugging the model in practical applications. These challenges are similar to those found in the ADGRU-based transfer learning work by Lu et al., who highlighted the challenges posed by the spatiotemporal heterogeneity of environmental data to predictive models. These limitations point out the direction for future improvements of the model and also provide important references for research in the field of environmental data science.

C. Future Research Directions

Based on the research results and current limitations of the CNN-LinATFormer model, future research can be carried out in the following directions: First, for the problem of high pollution concentration prediction, data enhancement technology or sample balancing strategy can be used to enhance the model's prediction ability for extreme pollution events. When the PM2.5 concentration exceeds 150μg/m³, the prediction error can reach -80μg/m³, which indicates that special attention needs to be paid to the processing of high-pollution samples. It is possible to consider developing prediction branches for different pollution levels or introducing specific high-concentration warning mechanisms to improve the performance of the model in extreme air pollution events. At the same time, combined with the performance of the model in the case study, exploring methods to integrate spatiotemporal information with monitoring data may help improve prediction accuracy, which echoes the feature selection method proposed by Zhang et al.

Secondly, based on the results of ablation experiments and hyperparameter analysis, future work can further optimize key components. In particular, the introduction of the channel attention mechanism produces significant performance improvements, indicating that adaptive learning at the feature level is of great value. More advanced feature selection and fusion strategies can be further explored, such as feature engineering methods that introduce environmental domain knowledge, or developing attention variants that can handle multi-source heterogeneous environmental data. In addition, considering the differences in the model's predictive ability under different pollution levels, an adaptive learning rate strategy or a multi-task learning framework can be developed to enable the model to automatically adjust its prediction strategy according to the characteristics of the input data. The transfer learning method of Lu et al. [8] may provide new ideas for solving data imbalance and domain adaptation problems. With the development of Internet of Things technology and distributed computing, deploying the CNN-LinATFormer model into the actual environmental monitoring system and realizing real-time PM2.5 concentration prediction is also a valuable research direction, which is also the main development trend of deep learning environmental applications.

VI. Conclusion

This study proposed an innovative CNN-LinATFormer hybrid deep learning architecture for high-precision PM2.5 concentration prediction. The model dynamically classifies environmental features by introducing a feature evaluator and realizes differentiated processing through three specially designed processing branches, effectively integrating the local feature extraction capability of CNN and the global dependency modeling capability of the linear attention mechanism. Comprehensive experimental evaluation shows that CNN-LinATFormer outperforms existing methods in all evaluation indicators, with an RMSE of 8.42 μg/m³, which is 21.1% lower than the CNN-RF model with the closest performance. Ablation experiments verify the effectiveness of each component, especially the introduction of the channel attention mechanism brings the most significant performance improvement and provides new technical ideas for environmental data modeling.

The case analysis reveals the performance characteristics of the model under different pollution levels. It performs best in the low concentration range (0-35μg/m³) (RMSE is only 3.12μg/m³), while the performance decreases under high pollution conditions (>150μg/m³) (RMSE is as high as 36.41μg/m³). This shows that the model's ability to predict extreme pollution events needs to be further improved. The hyperparameter experiment determined the optimal configuration (learning rate = 0.0018, batch size = 32, hidden layer dimension = 256), which is 22.2% higher than the baseline configuration, confirming the important impact of detailed parameter tuning on model performance. These findings not only provide guidance for the practical application of the CNN-LinATFormer model, but also point out the direction for subsequent research.

Future research can be carried out in the following directions: First, explore data enhancement technology or specific high-concentration warning mechanisms for the problem of high pollution concentration prediction; second, further optimize feature evaluators and attention mechanisms to improve the model's sensitivity to key environmental factors; third, combine Internet of Things technology and distributed computing to realize the deployment and application of the CNN-LinATFormer model in actual environmental monitoring systems. In general, the CNN-LinATFormer model proposed in this study provides a new and efficient solution for PM2.5 concentration prediction, which has important theoretical and practical significance for promoting the intelligent and popular development of environmental monitoring technology.

References

[1] R. T. Burnett et al., "An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure," Environ. Health Perspect., vol. 122, no. 4, pp. 397-403, 2014.

[2] Attiq, A. B., Nawaz, R., Irshad, M. A., Nasim, I., Nasim, M., Latif, M., ... & Fatima, A. (2024). Urban air quality nexus: PM2. 5 bound-heavy

metals and their alarming implication for incremental lifetime cancer risk. Pollution, 10(1), 580-594.

[3] World Health Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," WHO, Geneva, Switzerland, 2016.

[4] Y. Zhang, R. Zhang, Q. Ma, Y. Wang, Q. Wang, Z. Huang, "A feature selection and multi-model fusion-based approach of predicting air quality," ISA transactions, 2020.

[5] V. J. Berrocal, Y. Guan, A. Muyskens, H. Wang, "A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration," Atmospheric Environment, 2020.

[6] Y. Cheng, H. Zhang, Z. Liu, L. Chen, P. Wang, "Hybrid algorithm for short-term forecasting of PM2.5 in China," Atmospheric Environment, 2019.

[7] J. Peng, H. Han, Y. Yi, H. Huang, L. Xie, "Machine learning and deep learning modeling and simulation for predicting PM2.5 concentrations," Chemosphere, 2022.

[8] X. Lu, C. Ye, M. Shan, B. Qin, Y. Wang, H. Xing, "The prediction of PM2.5 concentration using transfer learning based on ADGRU," Water, Air, & Soil Pollution, vol. 234, no. 7, pp. 1-18, 2023.

[9] Q. Liao, M. Zhu, and L. Wu, "Deep learning for air quality forecasts: a review," Curr. Pollut. Rep., vol. 6, pp. 399-409, 2020.

[10] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," arXiv preprint arXiv:2006.04768, 2020.

[11] S. Zhou, W. Wang, and L. Zhu, "Deep-learning architecture for PM2.5 concentration prediction: A review," Environ. Sci. Ecotechnol., vol. 21, 2024.