Air Quality Prediction Based on VMD-CNN-BiLSTM-Attention

Huang Xinxin[1], Mohd Suffian Sulaiman[2]*, Marshima Mohd Rosli[3]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia[1,2,3]
Cardiovascular Advancement and Research Excellence Institute (CARE Institute),
Universiti Teknologi MARA, Sungai Buloh Campus[3]

***Abstract*—With the advancement of industrialization, air pollution has emerged as a critical global health and environmental concern. This study presents an air quality prediction model based on variational mode decomposition, a convolutional neural network, bidirectional long short-term memory, and an attention mechanism. The variational mode decomposition method is employed to decompose the Air Quality Index sequence, capturing different local characteristics of the original data. A hybrid model is constructed by integrating the convolutional neural network for feature extraction, the bidirectional long short-term memory for temporal pattern recognition, and the attention mechanism for focusing on significant data features. The model is optimized using the Grey Wolf Optimizer for hyperparameter tuning, thereby enhancing prediction accuracy. The proposed model is evaluated using air quality data from Changsha, China, covering the years 2015 to 2023. The results demonstrate that our model outperforms several other models in terms of mean absolute error, mean squared error, root mean squared error, and R-squared. This study provides a robust approach to air quality prediction, offering valuable insights for residents and policymakers.**

***Keywords*—Air quality prediction; variational mode decomposition; convolutional neural network; bidirectional long short-term memory; hyperparameter optimization; air quality index**

I. Introduction

With the development of global industrialization, material living standards have continued to improve; however, air pollution has become increasingly severe [1], [2]. Since the mid-19th century, when London experienced severe smog pollution, people have gradually become aware of the dangers posed by air pollution. In the 20th century, smog events in cities such as Los Angeles and London further heightened public concern and prompted more extensive investigation into air pollution issues.

Current research on air pollution primarily focuses on outdoor atmospheric pollution, which includes major components such as ozone, suspended particulate matter, and nitrogen oxides. These pollutants originate from various complex sources, including industrial emissions, vehicle exhaust, and coal-fired power generation.

The Air Quality Index (AQI) is a crucial metric used to assess air quality. It consolidates the concentrations of several routinely monitored pollutants—including PM2.5, PM10, nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), into a single, conceptual index value. This index categorizes air pollution levels and air quality conditions, allowing the extent of pollution to be reflected in a scientific and intuitive manner.

According to data from the World Health Organization (WHO), air pollution accounted for 7.6% of total global deaths in 2015 [3], [4], [5]. Modern epidemiological studies have shown that air pollution can severely damage the human immune system and may cause irreversible long-term harm to future generations. In addition, air pollution results in significant social and economic losses. Experts estimate that, based on the value of life, 3.25% of deaths in China in 2017 were attributable to air pollution, resulting in an economic loss equivalent to approximately 1.53% of the country's GDP that year [6], [7].

Most countries now have air quality monitoring systems in place; however, these systems often experience delays in providing air quality data and cannot offer reliable early warnings for governments or the public. Additionally, due to the non-constant and non-linear nature of air pollution concentrations, predicting air quality remains a significant challenge.

In recent years, many researchers have proposed various air quality prediction models. Existing methods can generally be divided into two categories: classic dispersion models [8], [9] and data-driven models [10], [11], [12]. Classic dispersion models are based on physical principles, taking into account factors such as wind speed, wind direction, atmospheric stability, and emission source characteristics to simulate and estimate the dispersion and concentration distribution of pollutants in the atmosphere.

In contrast, data-driven models represent a modern approach to air quality prediction. These models rely primarily on large volumes of historical data and statistical learning techniques to forecast future air quality. Unlike traditional physics-based models, data-driven models emphasize learning and extracting patterns from past data. Common types include statistical methods, machine learning algorithms, and deep learning models.

Applying statistics to air quality prediction research mainly includes autoregressive (AR) models, autoregressive integrated moving average (ARIMA) models, grey models, and multiple kinear regression(MLR) models. In 2006, Pagowski et al. used a dynamic linear regression model to predict ozone concentrations within 24 hours, achieving reduced bias under conditions with fewer monitor measurements [13]. Pai et al. used the Grey prediction model GM(1,1) to predict ozone

*Corresponding authors.

concentrations, and the results showed that GM(1,1) had a mean absolute percentage error (MAPE) of 19% based on a small sample size, indicating average prediction accuracy. The advantage of linear regression models is their fast modeling speed and good interpretability, but they handle non-linear issues in air quality datasets poorly and cannot fit non-linear data well [14]. Donnelly et al. combined a non-parametric kernel regression description of NO2 concentrations varying with wind speed and direction to obtain seasonal and daily variation factors, achieving good prediction results on the basis of the multiple linear regression model. Niu et al. established an autoregressive moving average (ARMA) model for air quality prediction in Chengdu [15], and in the same year, Zhang et al. used wavelet decomposition methods to improve the ARMA prediction model, decomposing and reconstructing air pollutant concentrations. The results proved that wavelet multi-scale decomposition could significantly improve the prediction accuracy of the ARMA model [16].

Machine learning models may learn patterns from historical data and use these patterns to predict future air quality. It also optimize prediction performance by adjusting model parameters and structures. Liang et al. used an 11 years dataset collected by Taiwan's Environmental Protection Administration (EPA) and then employed machine learning methods including adaptive boosting (AdaBoost), artificial neural networks(ANN), random forest (RF), stacking ensemble, and support vector machines(SVM) for individual predictions. The experimental results indicated that the stacking ensemble performed the best in terms of prediction accuracy [17]. Castelli et al. used support vector regression(SVR) to predict pollutant and particulate matter levels for the prediction of the AQI, and in their experimental scheme, SVR with a radial basis function (RBF) achieved the best prediction results [18]. Liu et al. constructed regression models using SVR and random forest regression (RFR) to predict the AQI in Beijing and the nitrogen oxides(NOx) concentration in an Italian city. The experimental results showed that the SVR-based model performed better in AQI prediction, while the RFR-based model performed better in NOx concentration prediction [19].

Deep learning models, in contrast to machine learning models that typically require less data for training, rely on large amounts of data and complex model structures for training. However, deep learning models can better capture the complex nonlinear relationships in the learning data and can integrate multiple data sources to improve the accuracy and comprehensiveness of predictions [20], [21]. Baniasadi et al. proposed a novel binary chimp optimization algorithm(BChOA) and combined it with long short-term memory(LSTM) networks.The BChOA algorithm fine-tunes the optimization parameters of the LSTM model to achieve more accurate and reliable air pollution forecasting. Experimental results demonstrated that the BChOA-LSTM model achieved the highest accuracy of 96.41% [22]. Bun Theang Ong et al. proposed deep recurrent neural networks(DRNN), which use an autoencoder enhancement pre-training method specifically designed for time series prediction to achieve PM2.5 concentration prediction. The experimental results showed that the DRNN model could handle air quality prediction problems well [23]. Wang et al. noticed that existing prediction models did not accurately capture the regularities between haze concentration and real-world influencing factors and proposed a two-layer model prediction algorithm based on long short-term memory neural networks and threshold regression units (LSTM and GRU). It is an improvement and enhancement of the existing forecast method of LSTM [24]. Wu et al., a novel air quality forecasting model based on the improved sparrow search algorithm (ISSA) and LSTM networks has been proposed,integrating the mRMR-RF feature selection method to enhance the model's predictive accuracy. By combining mRMR and RF,the model effectively selects variables that impact the AQI. The ISSA algorithm is utilized for the hyperparameter optimization of the LSTM. Experimental results indicate that this approach significantly improves the accuracy and efficiency of the forecasts [25].

A. Our Contribution

In this study, the VMD method is used to decompose the input AQI sequence. The resulting IMF components represent different local characteristics of the original data. Predicting each decomposed IMF component in parallel reduces the time required to identify signal features, thereby improving both the training efficiency of the algorithm and the accuracy of the predictions.

For the construction of the prediction model, we develop a hybrid model for air quality forecasting. Recognizing that LSTM networks do not fully resolve the vanishing gradient problem, we introduce CNN into the model to more effectively extract features from the air quality data. However, we also acknowledge that this integration increases the model's dependence on large volumes of training data and raises the risk of convergence to local optima. To address these challenges, we explore improvements to the internal structure of the neural network and consider incorporating additional, more effective neural architectures for feature extraction.

This study combines the CNN's capability to extract multi-scale temporal features with the strengths of BiLSTM, which processes sequences through both forward and backward layers of improved LSTM units. After feature extraction by the CNN, the BiLSTM computes the temporal dependencies, and the Attention mechanism is employed to calculate the attention weights of data features at different time steps with respect to the predicted value. This reveals the correlation between the time series data and the prediction target. By integrating these specialized neural networks, we construct an enhanced CNN-BiLSTM-Attention model capable of forecasting air quality for future periods with high accuracy.

Furthermore, to avoid arbitrary hyperparameter selection in the CNN-BiLSTM-Attention model, we apply the Grey Wolf Optimizer (GWO) to systematically search for optimal hyperparameters. This contribution demonstrates how modern machine learning techniques can be effectively combined and how model performance can be significantly improved by optimizing the training process through algorithmic tuning.

Experimental results confirm that, compared to other models with similar structures used for air quality prediction, the proposed model achieves higher accuracy.

II. Model Introduction

A. Variational Mode Decomposition (VMD)

Variational Mode Decomposition (VMD) is an adaptive signal processing method proposed by Konstantin Dragomiretskiy

and Zoubin Ghahramani in 2014. This method has the advantage of being able to determine the number of modal decompositions. Its adaptability is reflected in the ability to determine the number of modal decompositions for a given sequence based on actual conditions. In the subsequent search and solution process, it can adaptively match the best central frequency and finite bandwidth for each mode, and can effectively separate the Intrinsic mode functions (IMF) and divide the signal into frequency domains, thereby obtaining the effective decomposition components of the given signal and ultimately obtaining the optimal solution to the variational problem. It can avoid aliasing by controlling the bandwidth and can effectively address the mode mixing defect present in the Empirical Mode Decomposition (EMD) method [26], [27]. The basic principle includes the following steps:

The original signal is assumed to be decomposed into K components, ensuring that the decomposed sequences are modal components with finite bandwidth centered at specific frequencies. The variational problem can be described as finding K modal functions $u_k(t)(k = 1, 2, ..., K)$, such that the sum of the estimated bandwidths of each mode is minimized, subject to the constraint that the sum of all modes equals the original signal. The modal functions are defined by the following Eq. (1):

$$u_k(t) = A_k(t)cos(\phi_k(t)) \tag{1}$$

Use the Hilbert transform to calculate the analytic signal corresponding to each modal function $u_k(t)$, and obtain the one-sided spectrum, as shown in Eq. (2):

$$(\delta(t) + \frac{j}{k}) * u_k(t) \tag{2}$$

where, $\delta(t)$ is the Dirac function, based on the estimated central frequency e^{-jw_kt} of the mixed analytic signals of each mode, modulate the frequency of each mode to the corresponding base frequency band, see Eq. (3):

$$[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{-jw_kt} \tag{3}$$

Calculates the square norm L^2 of the gradient of the aforementioned signals, estimate the sum of the bandwidths of the modal signals, and the constrained variational problem is represented as Eq. (4):

$$\begin{cases} \min\limits_{u_k,\omega_k} \{\sum_k \|\partial_t[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{-jw_kt}\|_2^2\} \\ \text{s.t.} \sum_{k=1}^K u_k = f \end{cases} \tag{4}$$

In Eq. (4), $u_k = u_1, u_2, ..., u_k$ represents the modal functions with finite bandwidth, $\omega_k = \omega_1, \omega_2, ..., \omega_k$ denotes the central frequencies of each mode, * is the convolution operator, ∂_t signifies taking partial derivatives, and e^{-jw_kt} indicates shifting the spectrum to the base frequency band.

Solving the variational problem by introducing the Lagrange penalty operator and the penalty coefficient, transforming the constrained variational problem described in Eq. (4) into an unconstrained variational problem, resulting in the augmented Lagrange expression, as shown in Eq. (5).

$$\begin{aligned} L(u_k, \omega_k, \lambda) = \alpha \sum_k \|\partial_t[(\delta(t) + \frac{j}{k}) * u_k(t)]e^{j w_k t}\|_2^2 + \\ \|f(t) - \sum_k u_k(t)\|_2^2 + (\lambda(t), f(t) - \sum_k u_k(t)) \end{aligned} \tag{5}$$

To solve the unconstrained problem, it is necessary to find the saddle point of Eq. (5). The alternating direction method of multipliers (ADMM) is used to iteratively update $\hat{u}_k^{n+1}$, ω_k^{n+1}, and $\hat{\lambda}_k^{n+1}$. The iterative update process is shown in Eq. (6).

$$\begin{cases} \hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \hat{\lambda}(\omega)/2}{1 + 2\alpha(\omega - \omega_k)^2} \\ \omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}(\omega)^2 d\omega|}{\int_0^\infty |\hat{u}(\omega)^2 d\omega|} \\ \hat{\lambda}_k^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau(\hat{f}(\omega) - \sum_k^{n+1} \hat{u}_k(\omega)) \end{cases} \tag{6}$$

In this context, $\hat{f}(\omega)$ and $\hat{u}_i(\omega)$ represent the Fourier transforms of the original signal and the modal components, respectively, n is the number of iterations, and τ is the fidelity coefficient.

The frequency center and bandwidth of each IMF component are continuously updated during the separation from the original signal, until the iterative stopping conditions, as shown in Eq. (7) are met.

$$\sum_{i=1}^K (\frac{\|\hat{u}_i^{n+1}(\omega) - \hat{u}_i^n(\omega)\|_2^2}{\|\hat{u}_i^n(\omega)\|_2^2}) < \epsilon \tag{7}$$

ϵ is the discrimination precision; if the convergence condition is met, then the iteration is stopped and K IMFs are obtained.

From the above reasoning, we may deduce that the VMD algorithm transforms the modal decomposition problem into a search for the optimal solution of a variational equation. It then obtains the corresponding modal components through continuous iterative updates. Compared to the EMD algorithm, the VMD algorithm can effectively suppress the endpoint effect and modal component aliasing. Additionally, the VMD algorithm reduces the non-stationarity of time series with high complexity and strong non-linearity, resulting in decomposed subsequences that contain multiple different frequency scales and are relatively stationary. Therefore, it is more suitable for non-stationary time series such as air quality.

B. Convolutional Neural Networks (CNN)

Convolutional neural networks(CNNs) are a special type of deep feedforward neural network that have been widely used in image processing, machine vision, and other fields. The most basic structure of a CNN is shown in Fig. 1, which includes convolutional layers, pooling layers, fully connected layers, and an output layer. The CNN analyzes the features of

the input image or one-dimensional data through convolution operations. The pooling layers are used to merge and extract features obtained from each convolutional layer, and finally, the fully connected layers perform dimensionality reduction and output results. This network has characteristics such as local connections and weight sharing in its structure [28], [29].

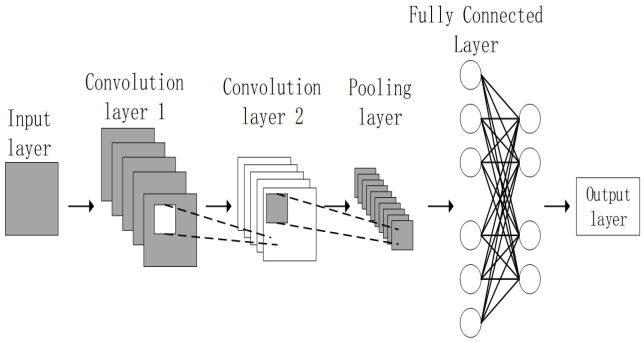

Fig. 1. CNN Network.

The fundamental principle of CNN is to utilize convolutional operations to extract features from the input, where the convolutional kernels are automatically learned by the network, thus endowing CNNs with excellent feature extraction capabilities. Each neuron in the convolutional layer only computes a small region of the input and shares weights, which can significantly reduce the number of parameters that need to be learned. The mathematical definition of the convolution operation is: the input matrix X and the convolutional kernel W, the convolution operation is represented as:$Y = X * W$. In the convolution operation, the convolutional kernel slides over the input matrix and performs convolution operations at each position to generate the output matrix Y. This process is a sliding window operation of the convolutional kernel over the input matrix, where the values in each sliding window are multiplied by the values in the convolutional kernel and summed to obtain the corresponding value in the output matrix. Pooling layers, on the other hand, perform dimensionality reduction on the feature maps output by the convolutional layers, typically using max pooling or average pooling. Max pooling selects the largest value in each pooling window as the output, while average pooling calculates the average value of the values in the pooling window as the output. Pooling operations may reduce the size of the feature maps and also enhance the robustness and generalization ability of the model [30], [31].

This study aims to improve the prediction accuracy of long-term meteorological data by utilizing a one-dimensional CNN neural network to extract meteorological influencing factor data $x = (x_1, x_2, ..., x_n)$. It employs convolutional layers and pooling layers to obtain effective representations, then flattens the acquired data and introduces it into a fully connected layer. The model analyzes the extracted feature data and realizes the output of feature results $c = (c_1, c_2, ..., c_n)$.

C. Bidirectional Long Short-Term Memory Networks-BiLSTM

Recurrent Neural Networks(RNNs) often encounter the vanishing or exploding gradient problem when dealing with relationships between nodes that are far apart, while LSTMs can better retain information provided by nodes that are distant from each other, enhancing performance on longer sequences of temporal data. Each LSTM unit is composed of three gating mechanisms: the forget gate, the input gate, and the output gate. The functional relationships between these gating units are as shown in Eq. 8 [32], [33]:

$$\begin{aligned} f_t &= \delta(W_f.[h_{t-1}, x_t] + b_f) \\ i_t &= \delta(W_i.[h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_c.[h_{t-1}, x_t] + b_c) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \delta(W_o.[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{8}$$

Specifically, x_t represents the input time-series data; f_t, i_t, o_t represent the outputs of the forget gate, input gate, and output gate, respectively; W_f, W_i, W_o are the weight matrices for the three gates, and b_f, b_i, b_o are the corresponding bias units; δ is the sigmoid function; $tanh$ is the hyperbolic tangent function; * denotes the inner product operation; $\tilde{C}_t$ represents the candidate vector created by passing through the tanh layer; W_c, b_c, are the weight matrix and bias unit, respectively, of the candidate layer; C_t represents the cell state; and h_t represents the hidden state.

During the data training process, LSTM can only use information from the forward sequence as the network's prediction result and cannot perceive backward data during model training. The advent of the bidirectional LSTM(BiLSTM) completely solves the problem of the model being unable to utilize future data. The term "bidirectional" refers to the existence of two LSTM networks within BiLSTM : one LSTM processes the forward sequence values, and the other processes the backward sequence values. These two networks operate independently, and the final output is achieved by vector concatenation to produce the final predicted features. Extensive research has shown that BiLSTM performs significantly better than LSTM in time series forecasting. Fig. 2 illustrates the structure of the BiLSTM network.

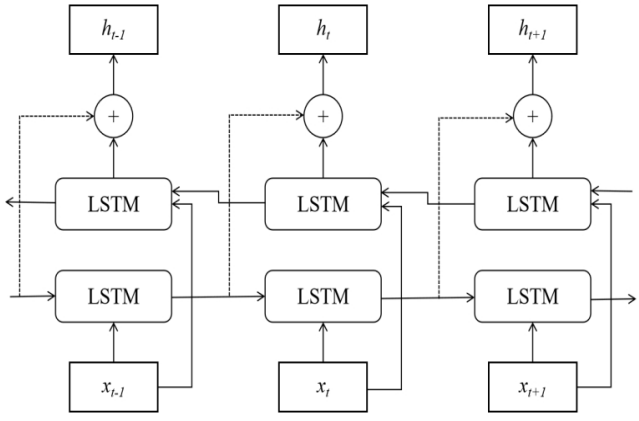

Fig. 2. BiLSTM network structure.

D. Attention

The attention mechanism is commonly used in tasks such as image classification and semantic understanding, where

it has achieved good results. Attention was first proposed in the context of image classification, primarily to enhance image understanding by omitting secondary information and emphasizing important information. Many computer vision algorithms still use Attention and its improvements to enhance their performance. However, since Bahdanau et al. applied Attention to neural translation and achieved good results, an increasing number of researchers have started to use it in natural language processing. For instance, Google's neural translation machine uses self-attention mechanisms exclusively to complete the entire model. Simply put, the attention mechanism is related to attention. Taking humans as an example, when receiving different messages simultaneously, different levels of response are given to different messages, and this difference is a manifestation of attention. Translating this to machine learning, different data require different processing methods, which means we pay different levels of attention to different aspects of the data based on its characteristics [34], [35].

In the process of meteorological data prediction training and learning, especially with multiple input variables, although effective correlation information is retained, it can lead to a complex network topology, resulting in slower learning speeds and difficulties in algorithm convergence. Therefore, this study incorporates an attention mechanism into the model, which adaptively and dynamically weights the factors affecting temperature, precipitation, and wind speed predictions. By reducing the weights of factors with weaker correlations to actual observed values, it allocates higher weights to the main influencing factors, achieving high-precision meteorological information forecasting. As shown in Fig. 3, the structure of the BiLSTM network after incorporating the attention mechanism is depicted.

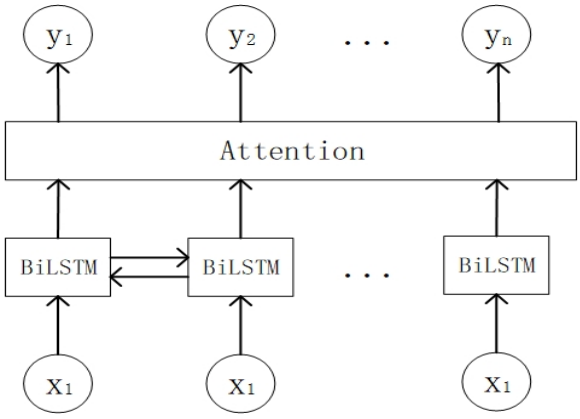

Fig. 3. BiLSTM-Attention network structure.

In the implementation process of introducing an attention mechanism into the BiLSTM model, a series of historical data $x = (x_1, x_2, ..., x_n)$ that affect meteorological predictions (temperature) are treated as the stored content of Key and Value during the addressing operation. Here, Key represents the data address, and Value represents the attention value. The attention calculation formula is as shown in Eq. (9):

$$A_i = \sum_{i=1}^{n} a_i v_i \tag{9}$$

The formula for calculating is given in Eq. (10):

$$a_i = softmax(Sim_i)\frac{eSim_i}{\sum_{k=1}^{n} e^{Sim_k}} \tag{10}$$

Specifically, v_i represents the value of i the data in the attention mechanism, and a_i represents the weight coefficient of v_i. We obtain the final attention by weighted averaging. Furthermore, Sim_i represents the cosine similarity of the i data, and the calculation formula is as shown in Eq. (11):

$$Sim(X, Key_i) = \frac{X.Key_i}{\|X\|.\|Key_i\|} \tag{11}$$

III. Materials and Methods

A. Case Measurement

To verify the effectiveness of the VMD-CNN-BiLSTM-Attention model proposed in this study for air quality forecasting, a total of 9 years of air quality data from January 2015 to December 2023 in Changsha, China, were selected as the training set to predict the air quality for a future period. The sequence model established based on air quality is shown in the Fig. 4.

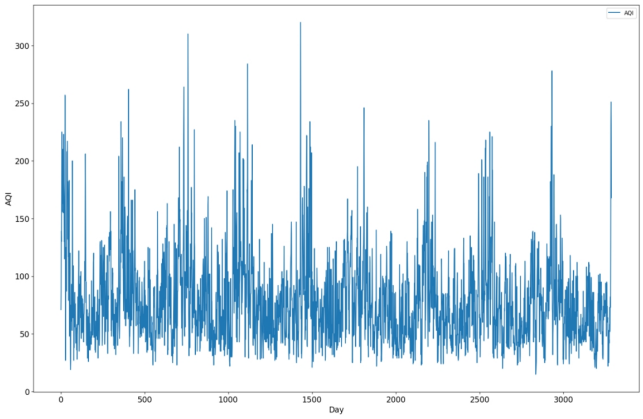

Fig. 4. The structure of the AQI data.

B. Data Processing

Due to the presence of seven different data indicators in the air quality forecasting model, namely AQI, PM2.5, PM10, SO2, NO2, O3, and CO, each indicator may have different characteristics, dimensions, orders of magnitude, and availability, making it impossible to directly analyze the characteristics and patterns of the research subjects. When there is a significant difference in the levels of various indicators, if we directly use the original values of the indicators for analysis, the role of indicators with higher values will be amplified in the comprehensive analysis, relatively weakening the role of indicators with lower values. Therefore, before using the data for prediction, we perform standardization processing on the data to ensure that different feature variables have the same scale. This allows the target variable to be controlled by multiple feature variables of the same size, and when using gradient descent to learn parameters, the impact of different

features on the parameters is consistent. This study uses min-max normalization for data standardization. The calculation formula is as shown in Eq. (12).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{12}$$

In which, X_{max} represents the maximum value in the sequence data sample; X_{min} represents the minimum value in the sequence data sample. VMD processing:this study employs the VMD algorithm to denoise and decompose air quality data, thereby enhancing the accuracy of the data. The sampling frequency is set to 2000, and after fine-tuning, the optimal penalty coefficient α is determined to be 1200, with the number of decomposition layers K set to 10. From the Fig. 5, it is evident that the decomposition of the air quality sequence using VMD yields 9 components with strong regularity. IMF1 is the dominant component, with a smooth curve that characterizes the overall trend of air quality changes. The remaining components, though of different frequencies, also exhibit regularity and can reflect the local characteristics of air quality data to a certain extent.

C. Model Frame

A combined forecasting model based on CNN-BiLSTM-Attention is used, with the model structure shown in Fig. 6. The model mainly consists of an input layer, a one-dimensional CNN layer, a dropout layer, a BiLSTM layer, an attention layer, a flatten layer, and a fully connected layer.

During the backpropagation process of the CNN-BiLSTM-Attention combined model, an appropriate algorithm is needed for parameter learning to guide the parameters of the objective function in the correct direction for updating the appropriate size, so that the updated parameters continuously approach the global optimum of the objective function value. Gradient descent principle is commonly used, following the negative gradient of the objective function to locate the minimum value of the function. This study uses the Adaptive Moment Estimation algorithm(Adam algorithm), with the purpose of accelerating the optimization process. At the same time, the Adam algorithm adjusts the throughput of the search process automatically for each variable encountered gradient(partial derivative) at each step length.

For the prediction of air quality data, the proposed model is implemented through the following processes:

- For the preprocessed time series $x = (x_1, x_2, ..., x_n)$, where, $x_i \in R^{sxf}, i \in (0, n)$ and s represents the length of each data's time window, and f is the dimension of the data feature vector. First, a one-dimensional CNN is used for feature extraction, and data padding is employed to ensure that the input time series dimensions remain unchanged.
- Further, for the extracted data $x = (x_1, x_2, ..., x_n)$, where, $x_i \in R^{sxd}, i \in (0, n)$, d denotes the size of the convolutional kernel in the convolutional neural network, adding a Dropout layer allows the model to adaptively block some hidden layer neurons without affecting the output dimension size, thereby enhancing the model's generalization ability.
- Secondly, the BiLSTM, which includes two LSTM units, receives forward and backward information, specifically calculated using Eq. (8). BiLSTM processes the output from the Dropout layer, achieving the exploration of periodicity and nonlinear relationships between time series.
- Then, the output of BiLSTM is taken as the input for the attention mechanism, which adaptively assigns weight coefficients to each input variable feature, further enhancing the model's perception of key information, with related calculation formulas shown in Eq. (9) to Eq. (11).
- Finally, after processing by the attention mechanism, the feature information is connected to the fully connected layer for prediction processing through the Flatten layer, and the final prediction results are outputted.

D. Experimental Design and Simulation

All programs in this study are completed using Python, with the training of the neural network part carried out using the PyTorch software library. PyTorch is an open-source neural network framework developed by the Torch7 team at Facebook AI Research. Its underlying implementation is based on Torch, but it is entirely implemented and utilized in Python. This framework is primarily used for scientific research and application development in the field of artificial intelligence. Torch is a classic tensor library for operating on multi-dimensional matrix data and is widely used in machine learning and other mathematically intensive applications. This study utilizes the GPU version of PyTorch, leveraging the excellent performance of GPU data processing to reduce model learning time. The entire experiment was conducted in a Windows environment, with the specific experimental environment as shown in Table I.

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Name	Versions
CPU	Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz (8CPUs)
Memory	DDR4 32GBytes
Operating system	Windows 10 Pro
GPU	NVIDIA Ge Force RTX 3090
Python	Python 3. 9. 13
PyTorch-GPU	PyTorch 1. 13. 0

The sample data is divided in a 9:1 ratio, where the 90% of the data is used as the training set and the remaining 10% as the test set.

In the experimental design of this study, we used batch processing. The batch processing refers to the batch size, which is the number of training samples used in each iteration. A larger batch size can better utilize the parallel computing capabilities of GPUs/CPUs but requires more memory. A smaller batch size occupies less memory but may lead to unstable convergence. Therefore, the batch size needs to be chosen based on actual memory conditions and model complexity. In this research, the batch size is set to 256.

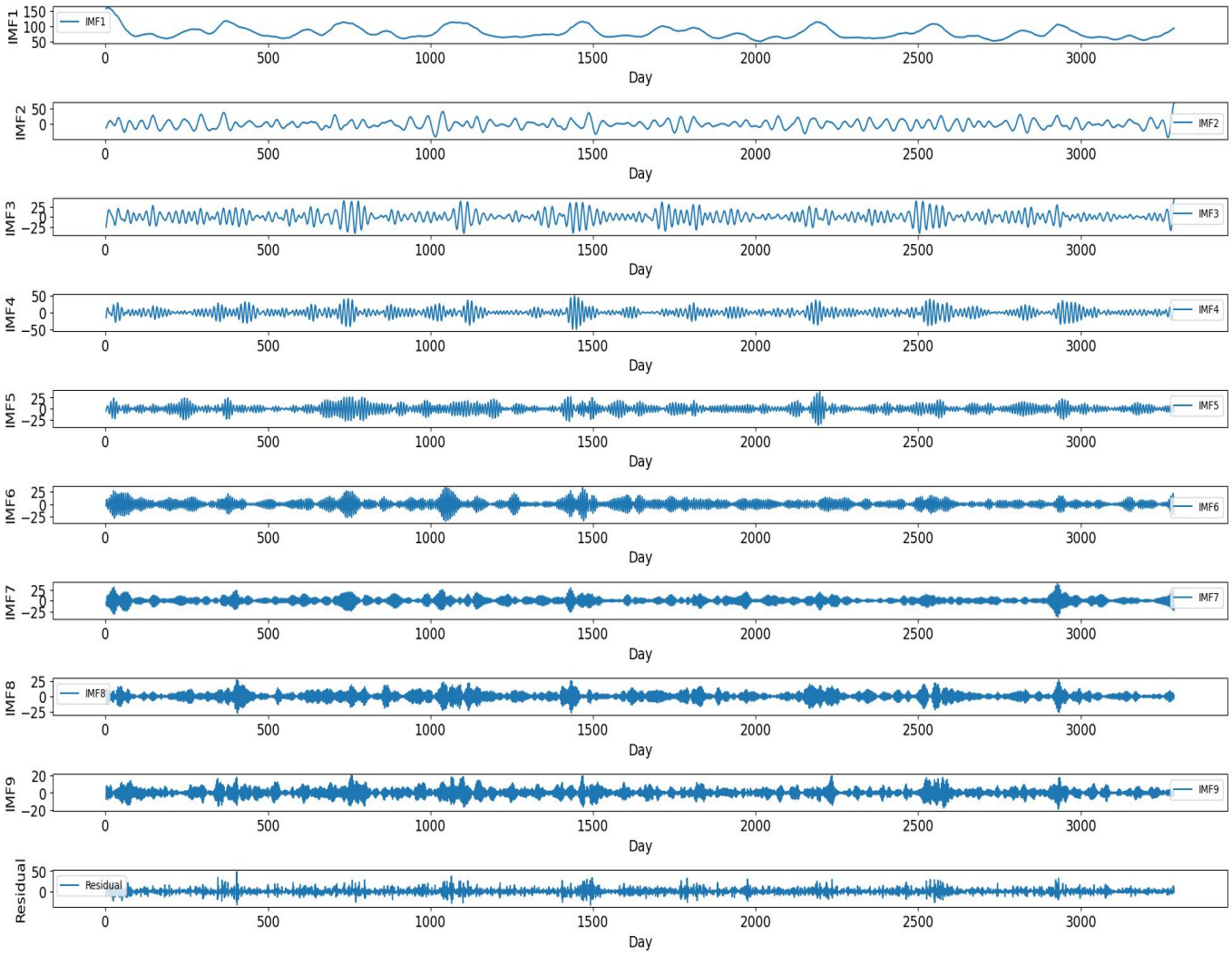

Fig. 5. The IMFs obtained after VMD.

The learning rate determines the update amplitude of weights in each training batch. A larger learning rate can speed up convergence but may miss the optimal solution. A smaller learning rate can increase accuracy but converges more slowly. In this research, the learning rate is set to 0.001.

The number of iterations is the number of times the entire training dataset is traversed. More epochs mean more thorough training but may also lead to overfitting. The number of iterations is set to 100 in this design.

Dropout is a regularization technique to prevent overfitting. It works by randomly "dropping" some neurons in the hidden layers during training, reducing the interdependence among neurons. This makes the network less sensitive to minor changes in input data and improves the model's generalization ability. A higher dropout rate can better prevent overfitting but may also lead to underfitting. In this research, the dropout rate is set to 0.1.

Time steps are mainly used for processing sequential data, such as natural language or time series data. It represents the number of steps when LSTM or RNN is unfolded. Longer time steps help capture long-term dependencies but also increase computational complexity and memory usage. Therefore, the time step should be chosen based on the nature of the task and hardware resources. In this experiment, the time step is set to 15, meaning that the model's training value at each moment is related to its previous 15 values.

The result measurement in these research, we use Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R Squared (R2) as metrics for evaluating the performance of the models. Lastly, the effectiveness of the hybrid model in air quality prediction is verified by comparing the VMD-CNN-BiLSTM-Attention hybrid model with RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM and CNN-LSTM-Attention models.

E. Result and Discussion

Table II and Fig. 7 are the result of air quality prediction performance of each model. The models are RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention and CNN-BiLSTM-Attention. The results shows that, for

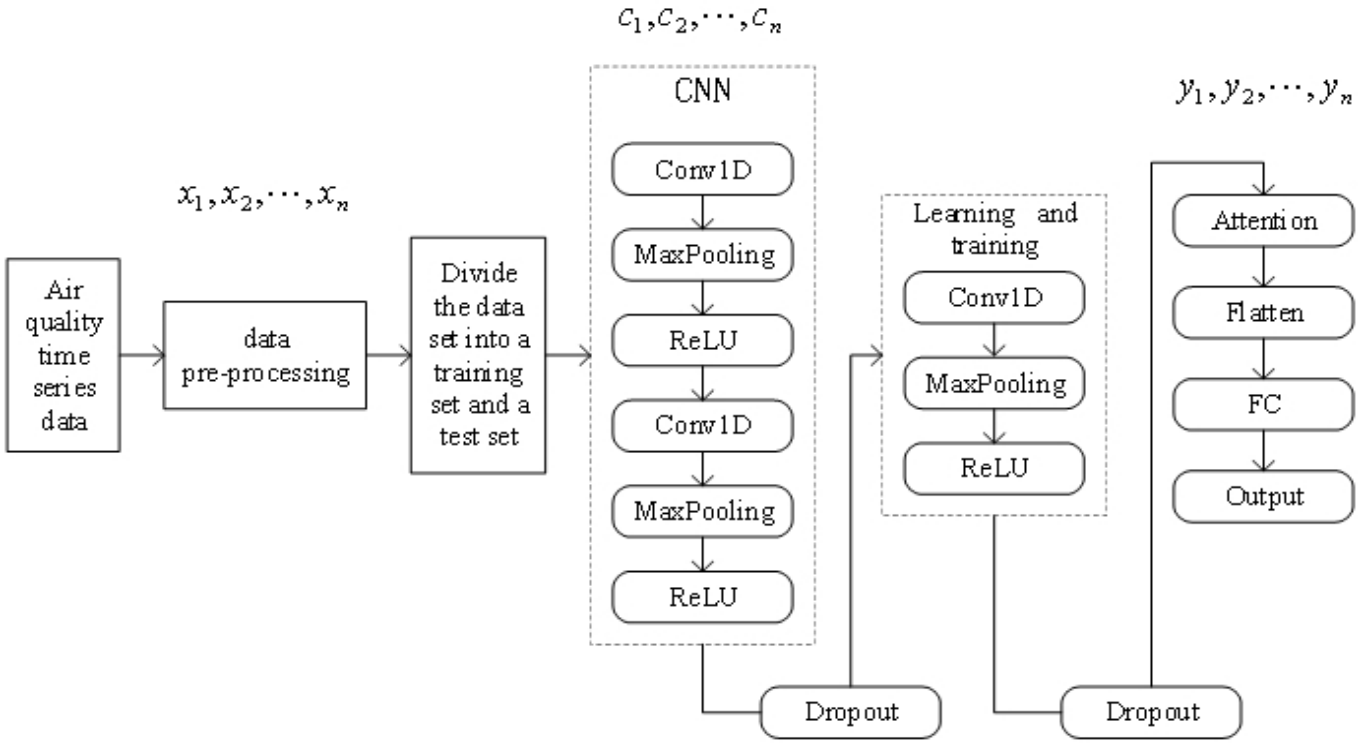

Fig. 6. Combined forecasting model based on CNN-BiLSTM-Attention.

individual models, BiLSTM and LSTM have better predictive performance compared to RNN, with BiLSTM outperforming LSTM. This indicates that RNN does not effectively utilize historical data of the air quality dataset, resulting in the poorest predictive performance. Incorporating bidirectional LSTM units can enhance the model's predictive performance, as BiLSTM offers higher efficiency and performance in data feature extraction compared to LSTM.

TABLE II. EVALUATION INDEX TABLE OF AIR QUALITY PREDICTION PERFORMANCE OF EACH MODEL

Model	MAE	MSE	RMSE	R2
RNN [36]	12.83	301.77	17.37	0.4226
LSTM [37]	12.81	276.47	16.63	0.2641
BiLSTM [38]	11.82	244.69	15.64	0.6288
CNN-LSTM [39]	8.28	133.71	11.56	0.7503
CNN-BiLSTM	7.34	100.71	10.04	0.8128
CNN-LSTM-Attention [40]	4.08	30.74	5.54	0.9588
CNN-BiLSTM-Attention	3.16	22.53	4.75	0.9685

However, when RNN, BiLSTM and LSTM models are combined with the neural network CNN, the predictive accuracy of the models significantly improved. CNN-LSTM, compared to LSTM, saw a 35% and 51% reduction in MAE and MSE, respectively, with R2 increasing to 75%. CNN-BiLSTM, compared to BiLSTM, experienced a 38% and 59% decrease in MAE and MSE, respectively, with R2 reaching 81%. This suggests that the addition of convolutional neural network CNN extracts effective features from the data, thereby enhancing the accuracy of predictions.

Upon integrating the Attention model, the predictive accuracy of the combined models further improved. CNN-LSTM-Attention, compared to CNN-LSTM, saw a 51% and 77% reduction in MAE and MSE, respectively, with R2 increasing by 28%. CNN-BiLSTM-Attention, compared to CNN-BiLSTM, experienced a 57% and 78% decrease in MAE and MSE, respectively, with R2 increasing by 19%. The attention mechanism can mimic the human brain's operational mechanism, focusing more on more important information when faced with varying external input data, thus avoiding the interference of irrelevant information. It can assign higher weights to key data features in air quality data, enabling the model to achieve higher predictive accuracy.

In summary, the ranking of predictive performance of different models is as follows:CNN-BiLSTM-Attention > CNN-LSTM-Attention > CNN-BiLSTM > CNN-LSTM > BiLSTM > LSTM > RNN. The predictive model CNN-BILSTM-Attention used in this study can effectively enhance the accuracy of predictions and obtain more precise outcomes. Compared to the initial RNN, CNN-BILSTM-Attention has reduced the MAPE.

IV. CONCLUSION

In recent years, as living standards have improved due to technological progress, the exploitation and use of fossil fuels by humans have also increased rapidly. This surge in fossil fuel consumption has led to serious air quality issues, posing a significant threat to both the environment and human health. In response to these challenges, this study focuses on predictive research related to air quality, aiming to achieve accurate forecasts through the development of an effective air

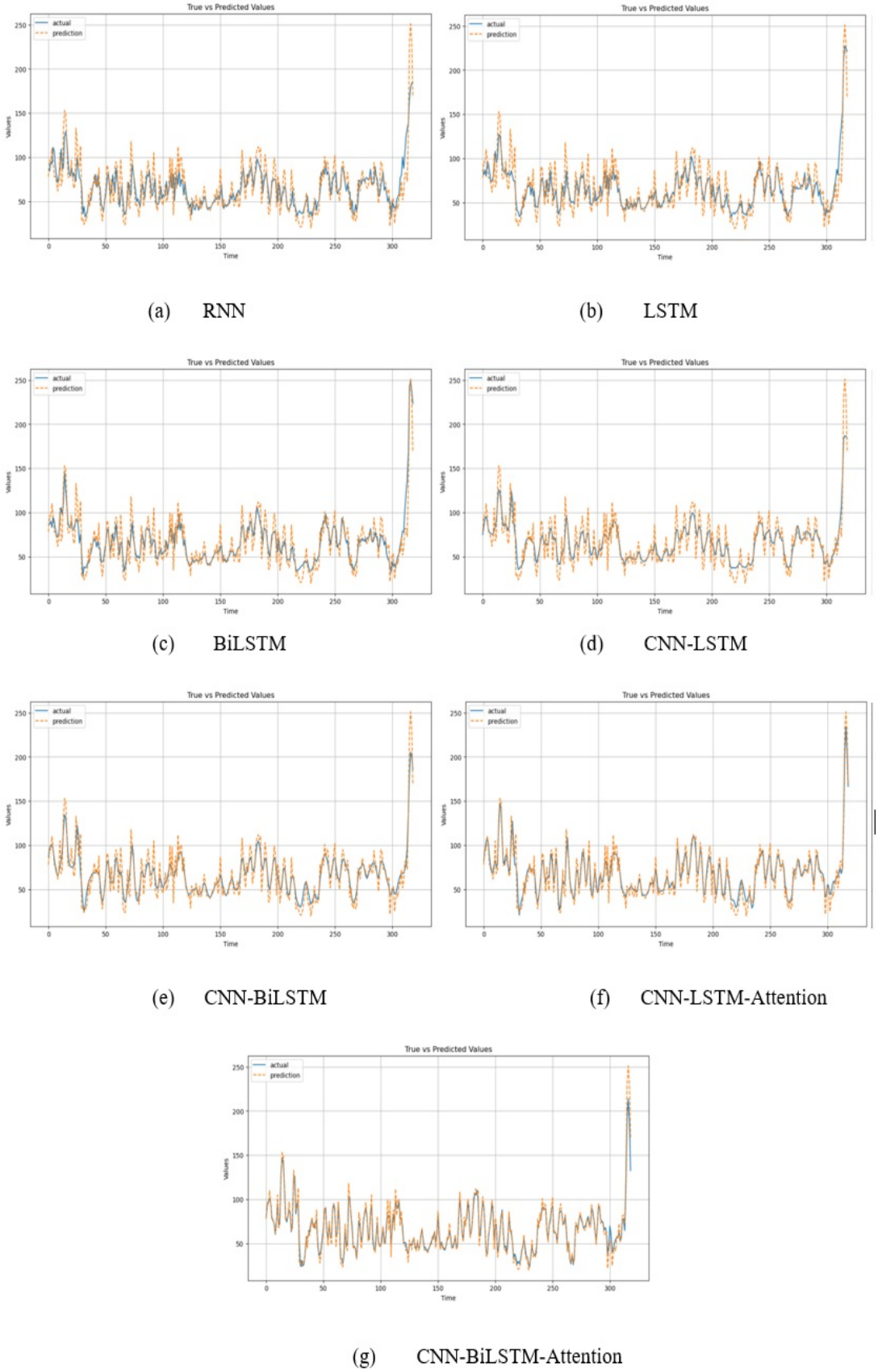

Fig. 7. Comparison of air quality prediction results from various models.

quality prediction model. Such a model can assist residents in planning their activities and provide scientific guidance to governments in formulating environmental protection policies.

Our research reveals that air quality time series data typically exhibit spatiotemporal diversity, meaning they display varying patterns across different times and locations. In addition, due to the nonlinear nature of meteorological and physical processes, as well as the influence of external factors, air quality data often show seasonal trends, daily cyclical variations, long-term patterns, and random fluctuations. These complexities make air quality prediction particularly challenging.

To address these issues, this study proposes a combined VMD-CNN-BiLSTM-Attention model designed to capture the spatiotemporal diversity and seasonality inherent in air quality time series. The model begins by applying the VMD algorithm to decompose and denoise the air quality data. Following decomposition, the dataset is standardized through normalization, improving data accuracy and facilitating more effective neural network predictions.

Within the proposed architecture, the CNN is responsible for extracting spatial features from the air quality data, while the BiLSTM network captures the temporal dependencies. The Attention mechanism then assigns varying weights to different elements of the data sequence, emphasizing the most influential features. This integrated model structure enhances the ability to capture complex spatiotemporal patterns, thereby improving prediction accuracy and reliability.

During model training, we evaluated the performance of seven different neural network models: RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention, and CNN-BiLSTM-Attention. Model performance was assessed using four evaluation metrics, there are MAE, MSE, RMSE, and R2. The results demonstrate that the CNN-BiLSTM-Attention model outperforms the others across all indicators, achieving the highest accuracy in air quality prediction.

ACKNOWLEDGMENT

This work was supported by the Journal Support Fund, UiTM.

REFERENCES

[1] Y. Ye, Q.Tao, "E The dynamic relationship among economic development, air pollution, and health production in China: the DNSBM efficiency model," *Frontiers in Environmental Science.*, vol, 11, pp. 1205712, 2023.

[2] Y.Tan, and X. Mao, "Assessment of the policy effectiveness of Central Inspections of Environmental Protection on improving air quality in China," *Journal of Cleaner Production.*, vol.288,no.12, pp. 125100,2020.

[3] O.Hahad, J.Lelieveld, F.Birklein, K. Lieb, A. Daiber and T.Münzel, "Ambient air pollution increases the risk of cerebrovascular and neuropsychiatric disorders through induction of inflammation and oxidative stress," *International journal of molecular sciences.*, vol. 21, no. 12, pp. 4306,2020.

[4] C.Xu, Z.Zhang, G.Ling, G. Wang, and M.Wang, " Air pollutant spatiotemporal evolution characteristics and effects on human health in North China," *Chemosphere.*, vol. 294, pp. 133814, 2022.

[5] D.Zhan, M. P.Kwan, W.Zhang, X.Yu, B.Meng, and Q. Liu, " The driving factors of air quality index in China,"*Journal of Cleaner Production.*, vol. 197, pp.1342–1351,2018.

[6] Y.Liu, J.Xu, D.Chen, P.Sun, and X.Ma,"The association between air pollution and preterm birth and low birth weight in Guangdong, China," *BMC public health.*, vol.19, pp.1–10,2019.

[7] X.G.Zeng, F.F.Ruan, and Y.Y.Peng, "Spatial distribution of PM2. 5 pollution health effects in China based on spatial grid scale ," *China Environ. Sci.*, vol.39, pp.2624-2632,2019.

[8] Q.Liao, M.Zhu, L.Wu, X. Pan, X.Tang, and Z.Wang, "Deep learning for air quality forecasts: a review,"*Current Pollution Reports.*,vol. 6, no.4,pp.399-409,2020.

[9] A.Aggarwal, D.Toshniwal, "A hybrid deep learning framework for urban air quality forecasting,"*Journal of Cleaner Production.*,vol. 329, pp.129660,2021.

[10] J.Wang, J.Li, X.Wang, J.Wang, and M. Huang, "Air quality prediction using CT-LSTM,"*Neural Computing and Applications.*, vol.33, pp.4779–4792, 2021.

[11] J. Ferreira , D.Lopes, S. Rafael, H. Relvas, , S. M.Almeida,and A. I. Miranda, "Modelling air quality levels of regulated metals: Limitations and challenges,"*Environmental Science and Pollution Research.*, vol.27, pp.33916–33928, 2020.

[12] A. Baklanov, Y. Zhang, "Advances in air quality modeling and forecasting,"*Global Transitions.*, vol.2, pp.261-270,2020.

[13] M.Méndez, M.G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey,"*Artificial Intelligence Review.*, vol.56, no. 9, pp.10031-10066,2023.

[14] H.Liu,G. Yan, Z. Duan, and C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," *Applied Soft Computing.*, vol. 102, PP.106957,2021.

[15] J.Zhang, and S. Li,"Air quality index forecast in Bei**g based on CNN-LSTM multi-model," *Chemosphere.*, vol.308, pp.136180,2022.

[16] S.Abirami, and P. Chitra, " Regional air quality forecasting using spatiotemporal deep learning," *Journal of cleaner production.*, vol.283, pp.125341,2021.

[17] Y.C.Liang , Y.Maimury,A.H.L. Chen, and J.R.C.Juarez," Machine learning-based prediction of air quality," *applied sciences.*, vol.10,no.24,pp.9151,2020.

[18] M.Castelli,F.M. Clemente, A. Popovič, S.Silva, and L. Vanneschi, " A machine learning approach to predict air quality in California,"*Complexity.*, vol.1,pp, 8049504,2020.

[19] J.M.Bertrand, F. Meleux, A.Ung, G. Descombes, and A. Colette, " Improving the European air quality forecast of the Copernicus Atmosphere Monitoring Service using machine learning techniques,"*Atmospheric Chemistry and Physics.*, vol.23,no.9, pp.5317-5333,2023.

[20] B.Zhang, Y.Rong, and R.H. Yong," Deep learning for air pollutant concentration prediction: A review,"*Atmospheric Environment.*, vol. 290, pp.119347,2022.

[21] C.L.Wu, R.F.Song, X.H.Zhu, Z.R.Peng,Q.Y. Fu, and J. Pan," A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network,"*Environmental pollution.*, vol.320, pp.121075,2023.

[22] S.Baniasadi, R.Salehi, S.Soltani,D. Martín, P. Pourmand, and E. Ghafourian, " Optimizing long short-term memory network for air pollution prediction using a novel binary chimp optimization algorithm,"*Electronics.*, vol.12, no.18, pp.3985,2023.

[23] N.A.Zaini, L.W. Ean,A.N. Ahmed, and M.A. Malek, "A systematic literature review of deep learning neural network for time series air quality forecasting," *Environmental Science and Pollution Research.*, pp.1–33,2020.

[24] S.Abirami, and P.Chitra, "Regional air quality forecasting using spatiotemporal deep learning ,"*Journal of cleaner production.*, vol.283, pp.125341,2021.

[25] H.Wu, T. Yang, H.Li, and Z. Zhou, "Air quality prediction model based on mRMR–RF feature selection and ISSA–LSTM,"*Scientific Reports.*, vol. 13,no.1, pp.12825,2023.

[26] Y.Zeng, J.Chen, N.Jin ,X.P. Jin,Y. Du, "Air quality forecasting with hybrid LSTM and extended stationary wavelet transform,"*Building and Environment.*, vol.213, pp.108822,2022.

[27] W.Huiyong,Y. Tongtong, and L. Hongkun, "Air quality prediction model based on mRMR-RF feature selection and ISSA-LSTM,"*Scientific reports.*, vol.13, no.1,pp. 12825-12835,2023.

[28] L. Alzubaidi, J. Zhang, A.J. Humaidi, .A.Al-Dujaili, Y.Duan, O.Al-Shamma, J.Santamaría,M.A.Fadhel,A.Muthana and F.Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,"*Journal of big Data.* ,vol. 8, pp.1–74,2021.

[29] A. Brahms, "Representation error for real numbers in binary computer arithmetic," IEEE Computer Group Repository, Paper R-67-85.

[30] A.Sayeed, Y.Choi, E.Eslami, Y.Lops, A.Roy, and J. Jung, "Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance," *Neural Networks,.*, vol.121, pp.396–408,2020.

[31] C.Shang, J.Gao, H.Liu, and F.Liu, "Short-term load forecasting based on PSO-KFCM daily load curve clustering and CNN-LSTM model," *IEEE Access.*, vol. 9, pp.50344-50357,2021.

[32] N.Jin,Y.Zeng, K.Yan, and Z.Ji, "Multivariate air quality forecasting with nested long short term memory neural network," *IEEE Transactions on Industrial Informatics.*, vol. 17,no.12, pp.8514–8522,2021.

[33] M.Lee, L. Lin, C.Y.Chen, Y. Tsao, T.H. Yao, M.H.Fei, S.H. Fang, "Forecasting air quality in Taiwan by using machine learning," *Scientific reports.*, vol.10,no.1,pp. 4153,2020.

[34] Gilpin W. "Deep reconstruction of strange attractors from time series,"*Advances in neural information processing systems*, vol. 29, pp.204–216,2020.

[35] A.Aggarwal, and D. Toshniwal, "A hybrid deep learning framework for urban air quality forecasting," *Journal of Cleaner Production*, vol.329, pp.129660,2021.

[36] J.M. Han, Y. Q. Ang, A. Malkawi, and H.W. Samuelson, "Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements," *Building and Environment*, vol. 192, pp.107601, 2021.

[37] D.N. Fente, and D.K. Singh, "Weather forecasting using artificial neural network," *In 2018 second international conference on inventive communication and computational technologies (ICICCT)*, pp. 1757-1761, 2018.

[38] M. Schuster, and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[39] M. Fan, O. Imran, A. Singh, and S.A. Ajila, "Using cnn-lstm model for weather forecasting," *In 2022 IEEE International Conference on Big Data (Big Data)*, pp. 4120-4125, 2000.

[40] J. Shen, W. Wu, and Q. Xu, "Accurate prediction of temperature indicators in eastern china using a multi-scale cnn-lstm-attention model," *arXiv preprint arXiv:2412.07997*, 2024.