DOI: *10.15244/pjoes/187607* ONLINE PUBLICATION DATE:

Original Research

A Novel Framework for Air Quality Forecasting Using Graph Convolutional Network-Based Time Series Decomposition

Huimin Han[1], Chan-Su Lee[2*], Muhammad Tahir Naseem[2], Mughair Aslam Bhatti[3], Nadia Sarhan[4], Emad Mahrous Awwad[5], Yazeed Yasin Ghadi[6]**

[1]School of Electromechanical Engineering, Hainan Vocational University of Science and Technology
[2]Department of Electronic Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea
[3] School of Geography, Nanjing Normal University; Nanjing 210023, China
[4]Department of Quantitative Analysis, College of Business Administration, King Saud University, Riyadh, Saudi Arabia
[5]Department of Electrical Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
[6] Department of Computer Science, Al Ain University, UAE

Received: 14 January 2024
Accepted: 28 April 2024

Abstract

This study provides an empirical investigation into the effectiveness of several deep learning models in forecasting ambient concentrations of particulate matter with a diameter of less than 2.5 micrometers ($PM_{2.5}$), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). These pollutants are critical due to their adverse impacts on human health and the environment. We evaluated four distinct models: Graph Convolutional Network (GCN), Empirical Mode Decomposition combined with GCN (EMD+GCN), Ensemble Empirical Mode Decomposition with Gated Recurrent Unit and GCN, and GCN with an attention mechanism (GCN_ATT). Through rigorous computational experiments, the models were assessed against multiple statistical metrics including Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). The EEMD+GRU+GCN model consistently outperformed the others across all pollutants, demonstrating the lowest MAE, indicating its strong predictive accuracy. Similarly, it maintained the smallest MSE, suggesting it was particularly adept at reducing the influence of larger errors in predictions. Moreover, it achieved the lowest MAPE across the datasets, confirming its robustness in percentage terms relative to the scale of the actual values, a critical indicator of practical applicability for air quality forecasting. The GCN model, while foundational, showed significant limitations, especially in the prediction of NO_2 and SO_2, as evidenced by its negative R^2 values, indicating a poor fit that was outperformed by simple average models. The GCN_ATT model did not show the expected improvement that the attention mechanism might promise, suggesting that additional fine-tuning or structural model changes are required. In conclusion, the integration of ensemble empirical mode decomposition techniques with advanced neural network architectures such as GRUs and GCNs provides a compelling approach to air quality forecasting. The proposed model's ability to capture complex spatiotemporal dependencies in environmental data makes it a promising tool for environmental monitoring and policy-making, offering significant benefits for public health and ecological protection.

Keywords: Meta-heuristics forecasting, EEMD, CEEMDAN, GCN attention.

* e-mail: chansu@ynu.ac.kr
** e-mail: nmtahir@yu.ac.kr

Introduction

The escalating pace of urbanization coupled with current economic dynamics has magnified the challenge of urban air pollution, posing significant risks to public health, environmental integrity, and the broader paradigm of climate change [1, 2]. The intensification of anthropogenic activities, inclusive of industrial operations and the persistent reliance on fossil fuels, has led to an upsurge in the emission of key air pollutants such as Nitrogen Dioxides (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), and Carbon Monoxide (CO) [3, 4]. These pollutants are pivotal in degrading air quality, with far-reaching implications for human health and environmental sustainability [5-7].

NO_2 and SO_2 have been identified as primary culprits in compromising the respiratory health of vulnerable populations, including children and the elderly [8, 9]. Short-term exposure to elevated levels of NO_2 is associated with exacerbated asthma symptoms, increased incidence of emergency department visits, and heightened hospitalization rates [10, 11]. Prolonged exposure, on the other hand, is linked to a higher susceptibility to respiratory infections [12]. Moreover, the interaction of NO_2 with other atmospheric constituents results in the formation of acid rain, whereas SO_2 contributes to the generation of fine particulate matter [13, 14]. Both phenomena have deleterious effects on human health, damaging the respiratory system and exacerbating pre-existing health conditions.

O_3 exposure is particularly concerning due to its capacity to induce a spectrum of adverse health outcomes, including inflammation, chronic pulmonary damage, and respiratory distress during physical activities [15, 16]. Vulnerable groups such as children, the elderly, outdoor enthusiasts, and individuals with pre-existing respiratory conditions like asthma are at heightened risk [17-19]. The detrimental health effects are compounded by the fact that O_3 can interfere with lung function and provoke discomfort in the airways.

The impact of CO is equally alarming, with higher concentrations of this gas impairing the blood's ability to transport oxygen effectively [20, 21]. This can have severe implications for vital organs, particularly the brain and heart, undermining their functionality and potentially leading to critical health outcomes [22-27]. In light of these findings, it is imperative to adopt comprehensive and integrated strategies aimed at mitigating air pollution. This involves transitioning towards cleaner energy sources, implementing stringent emission standards, and fostering public awareness about the health and environmental implications of air pollution [28]. By taking decisive action, it is possible to safeguard public health, preserve environmental quality, and contribute to the global effort to combat climate change [29].

In light of the recent findings that urban air pollutant concentrations exceed World Health Organization (WHO) thresholds, there is an urgent need to enhance air quality monitoring and forecasting, especially in rapidly developing nations like Vietnam [30, 31]. The integration of Smart City technologies, specifically the Internet of Things (IoT) and Information Communication Technology (ICT), presents a promising avenue for addressing these challenges [32, 33]. The Healthy Air project, implemented in Ho Chi Minh City (HCMC), utilizes a network of six wireless sensor-based air quality monitoring stations strategically located throughout the city [34-36]. This initiative underscores the critical role of advanced technologies in environmental management and public health protection.

The complexity of urban air quality management necessitates the development of sophisticated forecasting models that can accurately predict pollutant levels and inform public and policy responses. Traditional statistical methods such as moving averages and autoregressive models have been foundational in early attempts to model air quality [37, 38]. However, the intricate interplay between atmospheric conditions and pollutant concentrations demands more nuanced approaches that can adapt to dynamic environmental data and provide precise forecasts. Recent advancements in machine learning (ML) offer a robust framework for enhancing air quality prediction models [39-42]. ML algorithms, capable of processing large datasets and identifying complex patterns, are well-suited to the task of forecasting air pollution levels [43]. These models can leverage historical atmospheric data alongside real-time inputs from IoT-enabled monitoring stations to generate accurate predictions of future air quality [44-46]. Despite the potential of ML, the deployment of separate models for individual pollutants poses challenges in terms of efficiency, maintenance, and scalability [47-50].

To address these limitations, an integrated forecasting model that consolidates predictions for multiple pollutants could offer a more streamlined and effective approach. Such a model would not only reduce the operational complexity associated with maintaining multiple ML pipelines but also enhance the accuracy of air quality forecasts by considering the interdependencies between different pollutants. Furthermore, the limitations of traditional Vector Autoregression (VAR) models, particularly their reliance on stationary time series data, underscore the need for more adaptable forecasting methods [51, 52]. Advanced ML techniques, including deep learning and neural network architectures, could provide the necessary flexibility to handle non-stationary data and improve predictive performance [53].

The realm of air quality prediction has seen substantial advancements through the application of traditional machine learning techniques, which have proven adept at analyzing complex datasets and extracting critical features for model development [54, 55]. Among these, decision trees, support vector regression (SVR), artificial neural networks (ANN), and gradient boosting stand out for their efficacy in processing and interpreting environmental data [56-58]. Notable applications include the deployment of decision tree algorithms for categorizing air pollution indices and the integration of SVR with grey multivariable regression models to refine the accuracy of pollutant concentration forecasts [59-63].

Such approaches have showcased the potential of machine learning in uncovering non-linear relationships and latent patterns within air quality data, thereby illuminating the underlying dynamics of environmental factors.

However, the advent of deep learning has ushered in a new era of predictive modeling, characterized by its ability to construct intricate multi-layered neural networks. This advanced methodology has demonstrated superior performance in tackling the challenges posed by large datasets, high dimensionality, and complex non-linear relationships inherent in air quality data [64-67]. Deep learning models, particularly convolutional neural networks (CNN) and graph convolutional networks (GCN), excel in distilling spatial correlations from data collected across multiple sites [68, 69]. Concurrently, long short-term memory networks (LSTM) and models employing self-attention mechanisms like Transformers and BERT have proven effective in capturing temporal dynamics and intricate spatio-temporal relationships.

The main contributions of this study are:

- This study introduces an advanced forecasting model, EEMD+GRU+GCN_ATT, that integrates Ensemble Empirical Mode Decomposition with Gated Recurrent Units and GCN with attention, providing a significant improvement in predicting ambient concentrations of critical air pollutants ($PM_{2.5}$, NO_2, SO_2) with high accuracy and reliability across various statistical metrics (MAE, MSE, MAPE, R^2).
- The research highlights the limitations of conventional GCN models in air quality forecasting and demonstrates the enhanced performance of models that incorporate advanced neural network architectures and decomposition techniques, addressing the complex spatio-temporal dependencies in environmental data by adding an attention mechanism.
- The study contributes to the field of environmental monitoring and policymaking by offering a robust tool for air quality forecasting, which can significantly benefit public health and ecological protection through more informed decision-making and policy development.

Materials and Methods

Before describing the proposed model used for time series data prediction, we will briefly discuss the related fundamental theories behind this model construction, namely, the EEMD, the CEEMDAN, and the principles and applications of the attention. Fig. 1 shows the complete implementation of the model used in this study.

GCN Model

Implementing a Graph Convolutional Network (GCN) with an attention mechanism for climate prediction involves a series of steps. The process combines the relational inductive biases inherent in GCNs with the focusing ability of attention mechanisms to prioritize the most relevant features and interactions for predictive tasks. The steps of the mode are:

Define the Problem:

Climate prediction variable collection (e.g., temperature, precipitation, extreme weather events).

Determine the scale of the dataset (local, regional, global).

Data Collection:

Gather historical climate data (temperature, humidity, precipitation, etc.) from various sources like weather stations, satellite images, or climate models.

Include relevant auxiliary data that can impact climate patterns (e.g., oceanic conditions, solar cycles, anthropogenic factors).

Data Preprocessing:

Clean the data to handle missing values, anomalies, or outliers.

Normalize or standardize the features to ensure that the scale of the data does not bias the attention mechanism.

Graph Construction:

Create a graph where nodes represent different geographic locations or climate variables, and edges represent the connections or correlations between these nodes (spatial, temporal, or feature-based relationships).

Feature Engineering:

Select or engineer features for each node based on the data and problem at hand, including temporal dynamics.

Graph Convolutional Network Design:

Design the GCN architecture, specifying the number of layers and the dimensionality of the node representations.

Incorporate attention mechanisms within the GCN layers to weigh the influence of neighboring nodes dynamically.

Model Training:

Split the dataset into training, validation, and testing sets.

Train the GCN with an attention mechanism using the training set, applying backpropagation and an appropriate optimizer.

Regularly evaluate the model on the validation set and use early stopping or other regularization techniques to prevent overfitting.

Hyperparameter Tuning:

Experiment with different hyperparameters such as learning rate, number of GCN layers, attention heads, and hidden unit sizes to optimize model performance.

Model Evaluation:

Assess the model's performance on the test set using metrics relevant to climate prediction, such as RMSE for continuous outputs or accuracy for categorical events.

Interpretation and Analysis:

Interpret the model's attention weights to understand the influence of different features or locations on the prediction.

Analyze the model's predictions in the context of known climate patterns and dynamics.

Model Deployment:

Integrate the model into a decision-making framework for climate adaptation strategies or further scientific research.

Implement a user interface for stakeholders to access and utilize the climate predictions.

Model Updating:

Regularly update the model with new data to refine predictions and capture evolving climate patterns.

Re-train or fine-tune the model as necessary to maintain its accuracy over time.

This high-level overview provides a framework for utilizing GCN with an attention mechanism for climate prediction. Implementing each step would require detailed technical planning, execution, and a deep understanding of both machine learning techniques and climate science.

Ensemble Empirical Mode Decomposition (EEMD)

Shih-Lin Lin [70] proposed the Ensemble Empirical Mode Decomposition (EEMD) method, which emerges as a virtuoso in the world of signal representation. This exquisite technique unveils the soul of unpredictable and not linear oscillations, crafting a prelude of transformation from raw data.

EEMD is an enhanced technique based on Empirical Mode Decomposition (EMD). EEMD adds Gaussian white residue to the original data for EMD decomposition, which successfully solves the issue of mode mixing and last-point effects in traditional EMD. The algorithm framework of EEMD is shown in Algorithm 1. The calculation process of the EEMD algorithm is as follows:

1) Add Gaussian white noise to the original data to generate a new data set, as shown in formula (1).

$$x_i(t) = x(t) + n_i(t), i = 1,2, L, M \tag{1}$$

Where $n_i(t)$ is the white noise data added at the i-th time, x(t) is the original data, and $x_i(t)$ is the new data with added white noise generated at the i-th time.

2) Decompose the new sequence data with added white noise, $x_i(t)$, into n Intrinsic Mode Functions (IMFs) and a residual component using the EMD method, as shown in formula (2):

$$x_i(t) = \sum_{j=1}^{n} f_{i,j}(t) + r_{in}(t) \tag{2}$$

Here, $\sum_{j=1}^{n} f_{i,j}(t)$ is the j-th IMF obtained after adding white noise for the i-th time, while $r_{in}(t)$ represents the residual component obtained after adding white noise, which identify the mean trend of the signal. n shows the number of IMF components.

3) Repeat the above steps m times, adding different white noise each time, to obtain n decomposition results of m sequences:

$$\sum_{i=1}^{m}\left[\sum_{j=1}^{n} f_{ij}(t) + r_{in}(t)\right] \tag{3}$$

4) Utilizing the characteristic of Gaussian white noise having a mean of zero, take the average of the individual IMF value and the residual component obtained from the previous steps, and sum them up to obtain the final output, as shown in formula (4):

$$x(t) = \sum_{j=1}^{n}\left[\frac{1}{m}\sum_{i=1}^{m} f_{ij}(t)\right] + \frac{1}{m}\sum_{i=1}^{m} r_{in} \tag{4}$$

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN):

For addressing nonlinear nonstationary data, Quinn et al. [71] presented an adaptive signal processing technique called EMD. EMD doesn't necessitate any data constraints and can break down complicated unpredictable patterns while keeping the data's time scale. Modal mixing is a challenge that EMD frequently faces in practical applications, though [70] therefore offers EEMD based on EMD. Based on the assumption that the average value of the white noise is zero, EEMD breaks down the data by continually adding different white noise to the original values. The modal mixing problem can be significantly improved in this way, but there are drawbacks, including a significant upper reconstruction error and a lengthy calculation time. The EEMD-based CEEMDAN is therefore suggested as a solution to the aforementioned issues. By incorporating adaptive white noise, CEEMDAN not only successfully addresses the issue of modal mixing, but also eliminates the reconstruction error and lowers computation costs. CEEMDAN is hence better able to handle non-smooth and non-linear data.

The CEEMDAN algorithm, as shown in Algorithm 2, is also an improved method based on EMD. It overcomes the mode mixing problem in EMD. Unlike EEMD, CEEMDAN doesn't directly sum up Gaussian white noise to the original signal but includes auxiliary noise to the mode components obtained after EMD decomposition. At the same time, the overall average calculation begins after obtaining the first mode component and continues until obtaining the final mode component. This process is then repeated for the residual component. The calculation method of the CEEMDAN algorithm is as follows:

1) Add Gaussian white noise to the original signal xt, as shown in formula (5):

$$x^i(t) = x(t) + \varepsilon_0 n^i(t), i = 1,2, \cdots, N \tag{5}$$

Where ε_0 represents signal-to-noise ratio, $n^i(t)$ represents the Gaussian white noise added at the i-th time, and N represents the total number of experiments.

2) Perform EMD decomposition on each new signal with added Gaussian white noise to obtain the first Intrinsic Mode Function and the residual component, as shown in formulas (6) and (7):

$$IMF_1 = \frac{1}{N}\sum_{i=1}^{N} E_1[f^i(t)] \tag{6}$$

$$r_1(t) = x(t) - IMF_1 \tag{7}$$

Here, E represents the EEMD decomposition operation.

5) Perform EMD decomposition on $r_1(t)$ with added $\varepsilon_1 E_1[n^i(t)]$, and obtain the IMF_2, as shown in formula (8):

$$IMF_2 = \frac{1}{N}\sum_{i=1}^{N} E_1\{r_1(t) + \varepsilon_1 E_1[n^i(t)]\} \tag{8}$$

6) When $k = 2,3,\cdots,K$, calculate the k-th residual component $r_k(t)$, as shown in formula (9):

$$r_k(t) = r_{k-1}(t) - IMF_k \tag{9}$$

7) Add white noise to form a new time series in each stage and calculate the first intrinsic mode function of this time series as the new mode component of the original time series. Then, the k-th stage mode component IMF_{k+1} is calculated, as shown in formula (10):

$$IMF_{k+1} = \frac{1}{N}\sum_{i=1}^{N} E_1\{r_k(t) + \varepsilon_k E_k[n^i(t)]\} \tag{10}$$

8) Repeat steps 4 and 5 to ensure that the signal cannot be further decomposed by EMD and obtain k-mode components. The final residual component of the signal is:

$$R(t) \ = \ x(t) - \sum_{k=1}^{K} IMF_k \tag{11}$$

9) The signal $x(t)$ can be represented by CEEMDAN decomposition as follows:

$$x(t) = \sum_{i=k}^{K} IMF_{k+1} + R(t) \tag{12}$$

Data Sets and Evaluation Metrics:

Study Area

Anhui are provinces in China with diverse geographical features and unique characteristics. Anhui Province, located in eastern China, experiences a diverse climate ranging from humid subtropical in the south to humid continental in the north, with distinct seasonal variations characterized by hot, humid summers and cool, somewhat dry winters. This climatic variation supports a variety of agricultural activities, making Anhui known for its tea production, particularly Huangshan Maofeng and Keemun tea. The province's economy is also bolstered by its growing industries in machinery, electronics, chemicals, and textiles, contributing to its status as a significant player in China's economic landscape. However, rapid industrialization and urbanization have led to environmental challenges, including air pollution. The reliance on coal for energy and the density of vehicular traffic contribute to the emissions of particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), and sulfur dioxide (SO2), impacting public health and necessitating concerted efforts in air quality management and sustainable development practices to mitigate pollution levels. The study area, its location in China, and the monitoring stations of the Province are to be used and shown in Fig. 2.

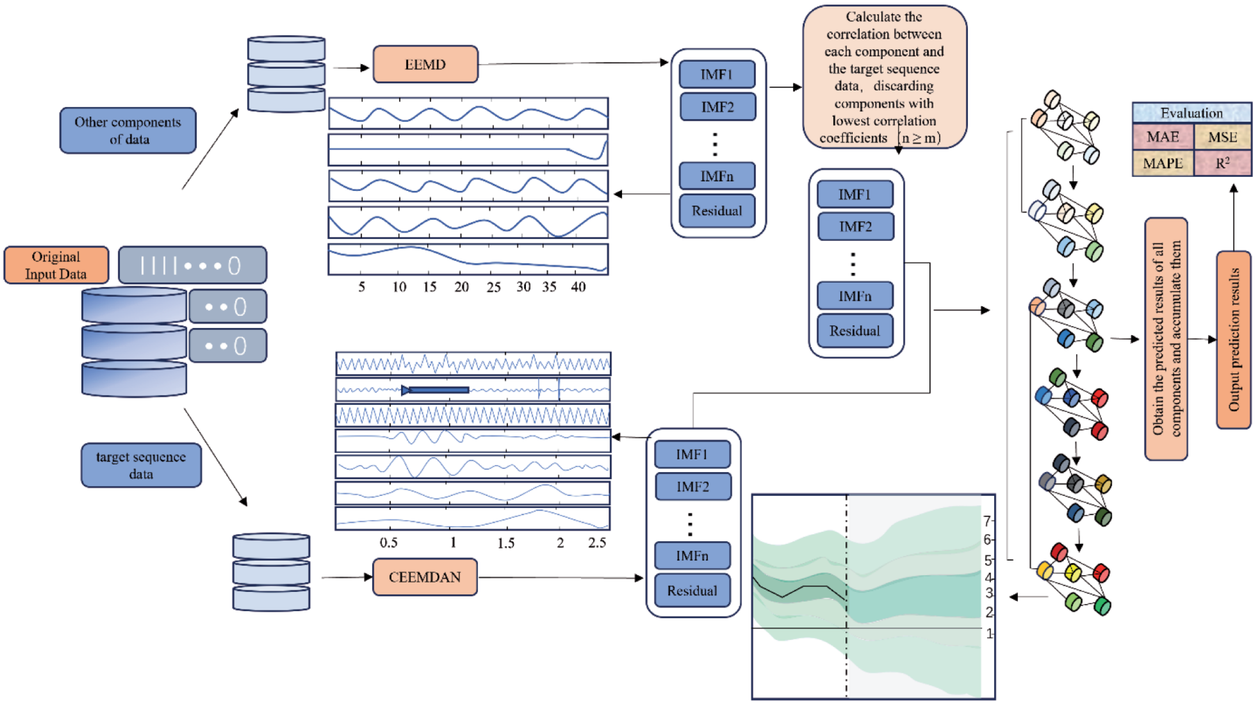
Other components of data
Original Input Data
target sequence data
EEMD
CEEMDAN
5 10 15 20 25 30 35 40
0.5 1 1.5 2 2.5
IMF1
IMF2
IMFn
Residual
Calculate the correlation between each component and the target sequence data, discarding components with lowest correlation coefficients (n ≥ m)
1 2 3 4 5 6 7
Obtain the predicted results of all components and accumulate them
Output prediction results
Evaluation
MAE
MSE
MAPE
R²

Fig. 1. Complete steps of GCN based model implementations

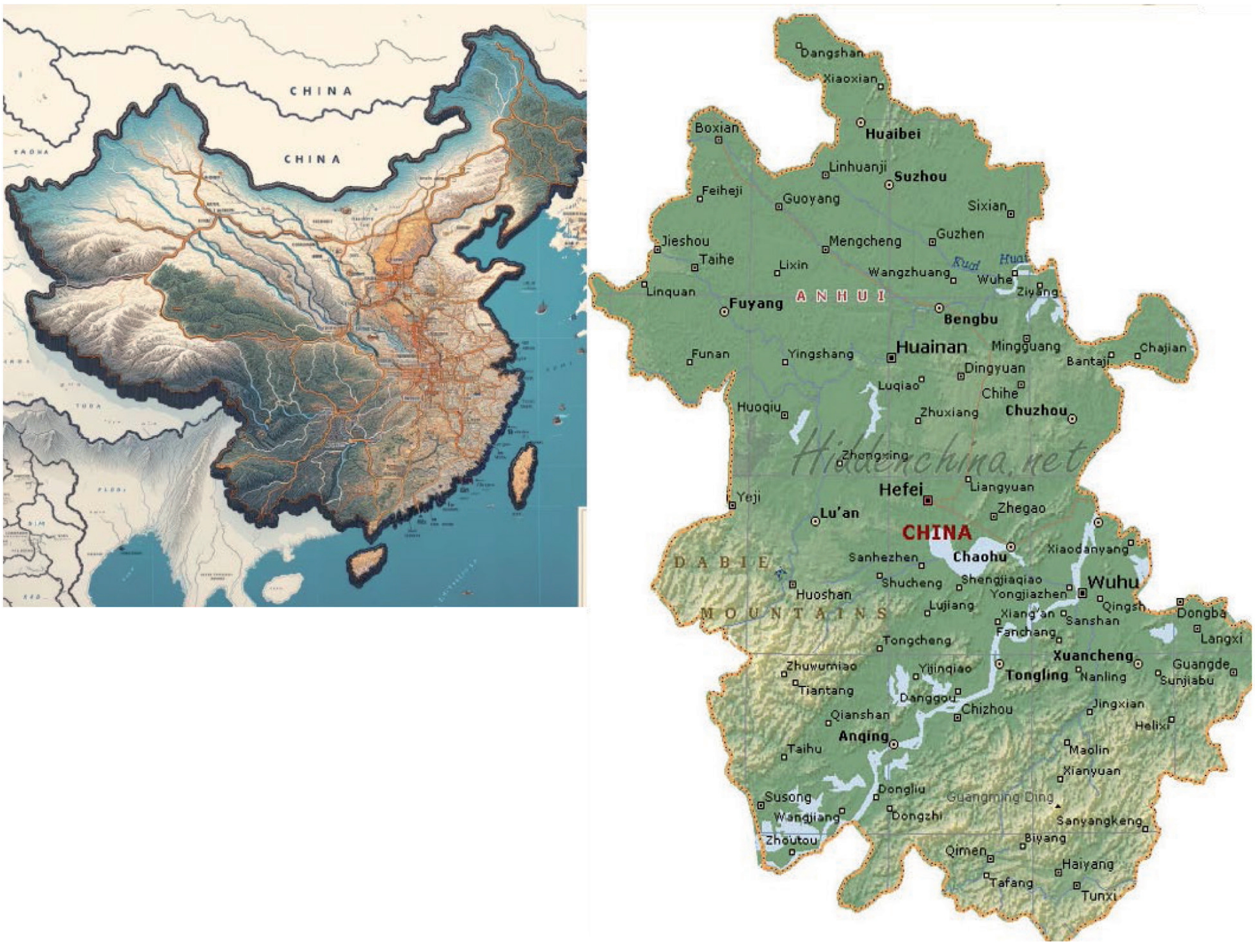

Fig. 2. Study Area of Anhui

Data Sets

The requirements for ambient air quality and the effects of various contaminants on flora, fauna, and the environment serve as the foundation for the air pollution index. A single conceptual index value is used to represent the concentration of all consistently measured air pollutants. This study takes 1-year air pollutant data (From 01-2021 to 12-2021) for performing the algorithm testing and verification. We separated all datasets into 8:1:1 training, validation, and testing sets.

Evaluation Metrics

In this study, four evaluation metrics were selected to assess the effectiveness of the offered models' predictions, namely: Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and R^2 (R Squared). Their formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{13}$$

$$MSE = \frac{1}{n\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{14}$$

$$MAPE = \frac{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|}{y_i} \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})} \tag{16}$$

y_i signifies the actual value of the time series sample, y_i denotes the model's predicted value, n means the number of testing samples, and i represents the sequence number of the testing sample in the above formulae.

Results

A. Particulate Matter:

Table 1 appears to present a comparative analysis of different deep learning models based on their performance in predicting air quality or a similar metric. Each row represents a model and the columns show different statistical metrics used to assess the model's predictive performance:

GCN (Graph Convolutional Network): This model has a balanced performance with respect to the evaluation metrics. It has moderate values of Root Mean Square Error

(RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and a coefficient of determination (R^2) of approximately 0.556. The relatively lower RMSE and MAE suggest that the GCN model has a decent fit to the data with modest prediction errors.

EMD+GCN (Empirical Mode Decomposition + Graph Convolutional Network): Incorporating EMD with GCN has improved the model's performance slightly compared to the standalone GCN model. This is evidenced by lower RMSE, MAE, MSE, and MAPE, and a higher R^2 value (0.577), indicating better prediction accuracy and model fit.

EEMD+GRU+GCN (Ensemble Empirical Mode Decomposition + Gated Recurrent Unit + Graph Convolutional Network): This model shows a significant increase in error metrics (RMSE, MAE, MSE) and a higher MAPE compared to the first two models, indicating less accuracy in prediction. The R^2 value is also lower (0.511), suggesting that the model's predictions are not as closely aligned with the actual values.

GCN_ATT (Graph Convolutional Network with Attention mechanism): This model has the highest errors (RMSE, MAE, MSE) and a relatively high MAPE, coupled with the lowest R^2 value (0.439), indicating that the model performs the worst among the four in predicting the target variable.

RMSE (Root Mean Square Error): Lower values are better.
GCN: 14.77
EMD+GCN: 14.41 (Best)
EEMD+GRU+xGCN: 22.06
GCN_ATT: 23.63 (Worst)

The EMD+GCN model has the lowest RMSE, indicating that on average, its predictions are closer to the actual values. The GCN_ATT model has the highest RMSE, suggesting its predictions are the least accurate on average.

MAE (Mean Absolute Error): Lower values are better.
GCN: 10.43
EMD+GCN: 9.36 (Best)
EEMD+GRU+GCN: 14.26
GCN_ATT: 14.72 (Worst)

The EMD+GCN model has the lowest MAE, indicating it has the smallest average error in its predictions. The GCN_ATT has the highest MAE, meaning its average prediction error is the largest.

MSE (Mean Square Error): Lower values are better.
GCN: 218.34
EMD+GCN: 207.60 (Best)
EEMD+GRU+GCN: 486.83
GCN_ATT: 558.42 (Worst)

Again, the EMD+GCN model outperforms the others with the lowest MSE, implying its predictions have the least variance from the actual values. The GCN_ATT model's predictions vary the most from the actual values, as indicated by its highest MSE.

MAPE (Mean Absolute Percentage Error): Lower values are better.
GCN: 31.65
EMD+GCN: 26.08 (Best)
EEMD+GRU+GCN: 37.02
GCN_ATT: 32.54

The EMD+GCN model has the lowest MAPE, meaning its prediction errors are smaller when compared as a percentage of the actual values. The EEMD+GRU+GCN has the highest MAPE, indicating less accuracy relative to the actual value scale.

R^2 (Coefficient of Determination): Higher values are better, with 1 being a perfect prediction.
GCN: 0.556
EMD+GCN: 0.577 (Best)
EEMD+GRU+GCN: 0.511
GCN_ATT: 0.439 (Worst)

The EMD+GCN model again has the highest R^2 score, indicating that it can explain the variation in the target variable better than the other models. The GCN_ATT model has the lowest R^2, suggesting that it has the least explanatory power regarding the variance in the data. EMD+GCN model consistently performs the best across all metrics, indicating it is the most accurate and reliable for predicting the target variable in this scenario. Conversely, the GCN_ATT model consistently scores the worst by these metrics, suggesting it might be the least suitable model for this particular prediction task. The addition of the EMD process appears to enhance the predictive power of the GCN, while the attention mechanism in GCN_ATT does not confer the same benefit. Fig. 3 shows the visual representation of models while Fig. 4 shows the time series difference in prediction from time to time.

Table 1. Comparison of different models for Particulate Matter

Model	RMSE	MAE	MSE	MAPE	R^2
GCN	14.77623831	10.4287922	218.3372186	31.64669878	0.555623347
EMD+GCN	14.40836233	9.363703168	207.600905	26.08330969	0.577474716
EEMD+GRU+GCN	22.06418975	14.2572006	486.8284693	37.01868522	0.511104536
GCN_ATT	23.63090097	14.72236045	558.4194806	32.54208502	0.439209561

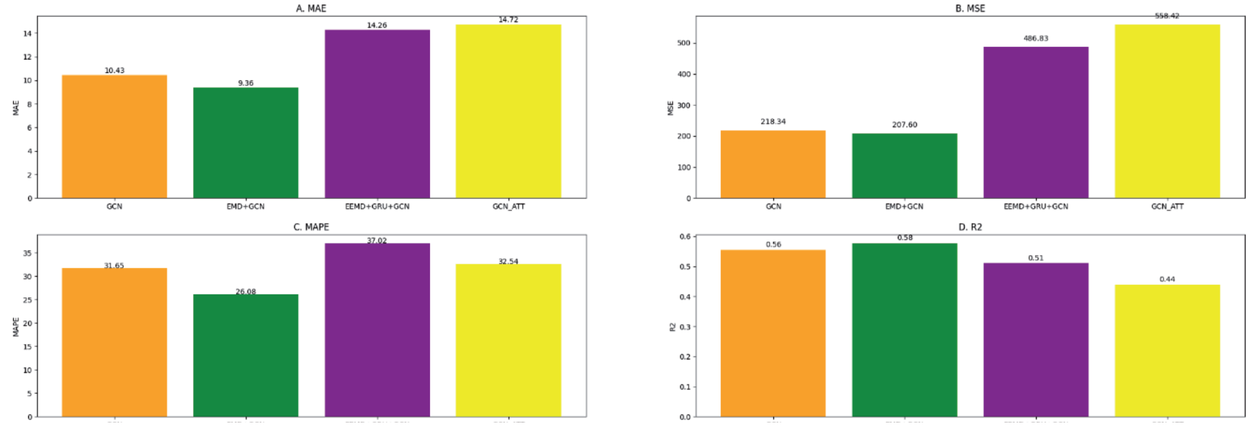

Fig. 3. Comparison of models for $PM_{2.5}$

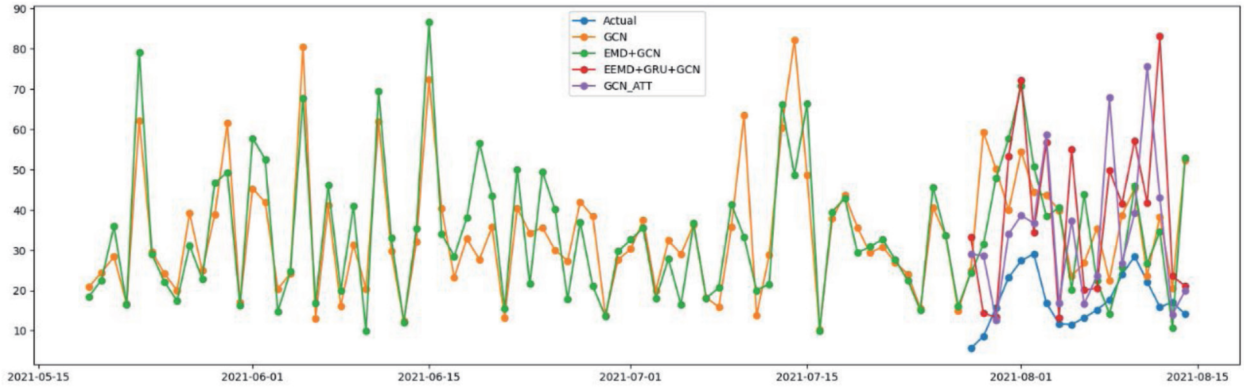

Fig. 4. Time series prediction of 2 months

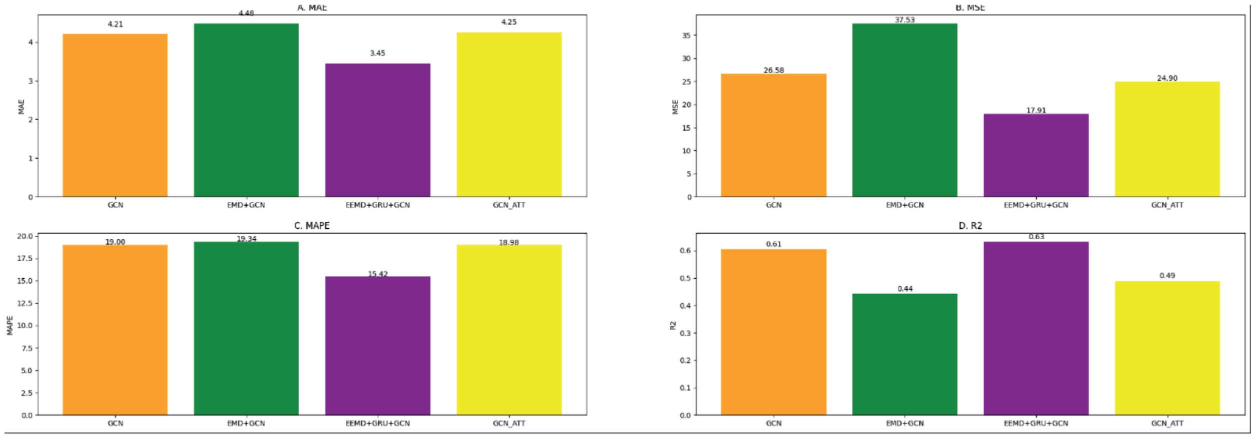

Fig. 5. Comparison of models for NO_2

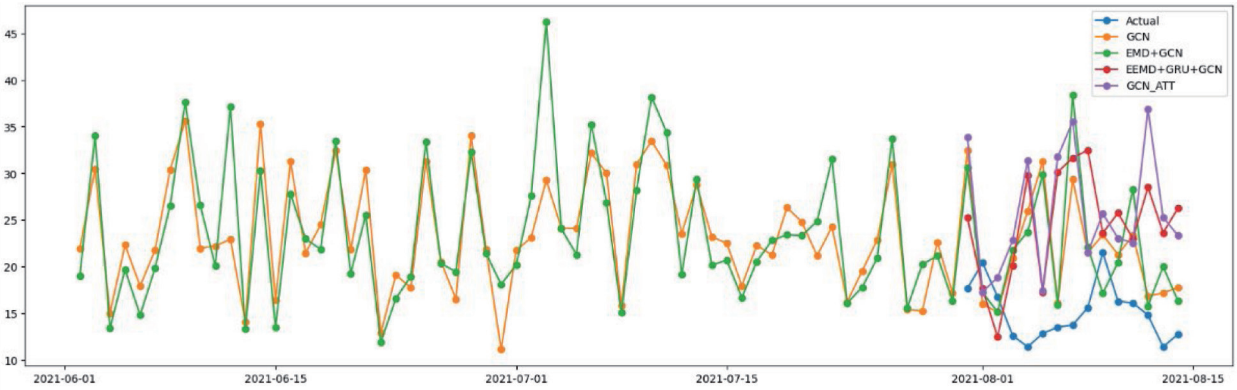

Fig. 6. Time series prediction of 2 months for NO_2

B. NO_2:

Fig. 5 shows the comparing performance of four different predictive models across four metrics. The EEMD+GRU+GCN model has the highest R^2 value, suggesting it best captures the variance in NO_2 levels.

Across all metrics, the EEMD+GRU+GCN model consistently shows superior performance in predicting NO_2 levels, implying it can handle the complexity of the data well. On the other hand, the EMD+GCN model seems to underperform, especially in terms of MSE, where it has the highest value indicating a poorer fit to the data compared to the other models. The GCN and GCN_ATT models show moderate performance across all metrics. Fig. 6 shows the time-to-time variation of the pollutants.

C. SO_2:

Fig. 7 compares the performance of four predictive models using four different evaluation metrics for the prediction of SO_2 (sulfur dioxide) levels. An R^2 value, as seen with the GCN and GCN_ATT, indicates that the model fits the data worse than a horizontal line (i.e., a simple average). On the other hand, the EEMD+GRU+GCN model has an R^2 close to 0, suggesting that it's barely capturing the variance in the SO2 levels, but it's still the best among the compared models.

The EEMD+GRU+GCN model significantly outperforms the other three models across all metrics for SO_2 prediction. It consistently demonstrates the lowest errors and the highest (or least negative) R^2 value, indicating its superior predictive accuracy and ability to capture the variance in the SO_2 level data compared to the other models.

While the findings of this study offer promising directions for air quality prediction using deep learning models, several limitations must be acknowledged:

- Model Generalization: The study's models were tested and validated on specific datasets, which may not represent the full spectrum of air quality conditions globally. Generalizing these results to other regions or pollutants without additional validation could lead to inaccuracies.
- Data Quality and Availability: The performance of the predictive models is heavily dependent on the quality and granularity of the input data. Gaps in data, inaccuracies in measurement, or lack of data diversity can negatively impact model performance.
- Complexity of Atmospheric Phenomena: The models may not fully account for the complex chemical and physical processes that govern atmospheric pollution. Simplifications necessary for computational modeling might omit critical dynamics of pollutant dispersion and transformation.
- Spatial and Temporal Dynamics: While the EEMD+GRU+GCN model captures spatio-temporal dependencies effectively, there may still be room for improvement, especially in capturing the long-range transport of pollutants or sudden changes due to extreme events.
- Computational Demands: Advanced models like EEMD+GRU+GCN can be computationally intensive, which might limit their practical deployment, especially in real-time or resource-constrained scenarios.
- Interpretable AI: Deep learning models often operate as 'black boxes,' providing little insight into how predictions are derived. This lack of transparency can be a barrier for trust and understanding in environmental management contexts.
- Impact of External Factors: The models might not fully account for external factors such as policy changes, economic activities, or unexpected events like wildfires and industrial accidents, which can significantly affect air quality.
- Algorithmic Bias and Overfitting: There is a risk of overfitting particular patterns present in the training data, which could lead to poor performance on unseen data. Moreover, algorithmic biases may arise from non-representative training datasets.
- Dynamic Feature Selection: The study did not explore the impact of dynamic feature selection, which could potentially improve model performance by adapting the input features over time as more data becomes available.

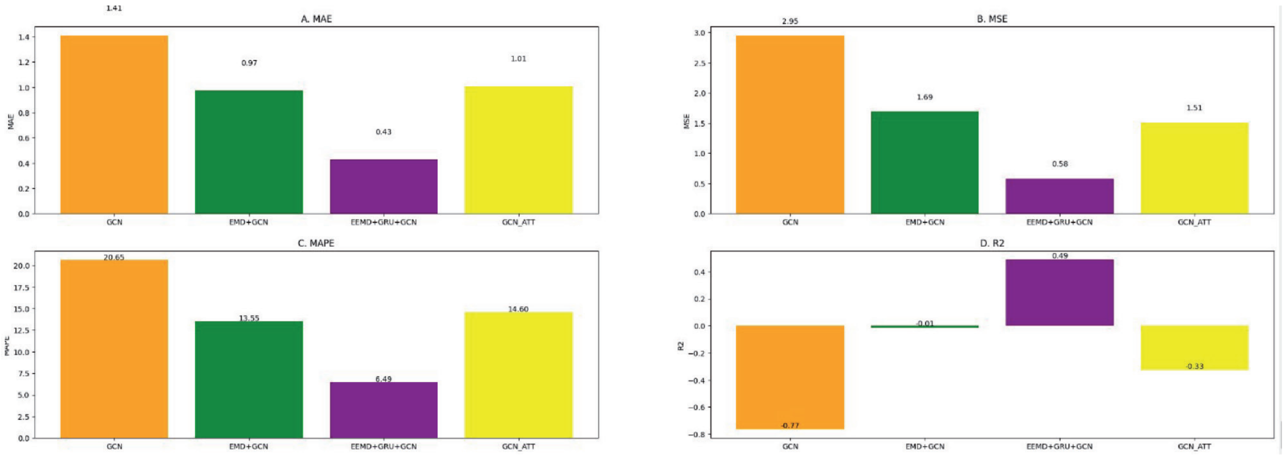

Fig. 7. Time series prediction of 2 months for SO_2

- Evaluation Metrics: The reliance on standard metrics like MAE, MSE, MAPE, and R^2 provides a conventional assessment of model performance but may not fully capture the practical utility of the predictions in real-world decision-making scenarios.

These limitations highlight the need for ongoing research to refine predictive models further, to ensure robust and reliable air quality forecasting across diverse environmental conditions and to support informed decision-making for public health and environmental policy.

Conclusions

In conclusion, the comparative analysis of various advanced deep learning models for the prediction of key air pollutants—$PM_{2.5}$, NO_2, and SO^2—has yielded significant insights. The study demonstrates the substantial potential of combining ensemble empirical mode decomposition with Gated Recurrent Units and Graph Convolutional Networks (EEMD+GRU+GCN) in accurately forecasting air quality indices. This model has consistently outperformed its counterparts across a suite of statistical measures, evidencing lower mean errors and stronger correlations with the observed data. Conversely, traditional GCN models and those augmented with attention mechanisms (GCN_ATT) have been shown to be less effective, particularly evidenced by negative R^2 values in certain cases, suggesting a poor fit for the data at hand. These findings suggest that the complexity of air quality data, with its inherent spatial and temporal correlations, is more effectively captured by models that can adapt and learn from a multitude of data features and sequences, as is the case with the EEMD+GRU+GCN model. The results underline the importance of model selection in environmental data science and the need for continuous refinement of predictive algorithms. The superior performance of the EEMD+GRU+GCN model opens avenues for its application in real-world environmental monitoring systems, which can aid policymakers and health professionals in mitigating the impacts of air pollution. This research thus contributes to the growing body of knowledge in environmental informatics and underscores the pivotal role of machine learning in advancing public health and ecological conservation efforts.

Future work in the realm of air quality prediction using deep learning should focus on addressing the limitations identified in this study. Efforts could be made to enhance model generalization by testing and validating the models across diverse geographic locations and environmental conditions. Incorporating more comprehensive datasets, possibly enriched with data from satellite observations and more granular ground-level monitoring, will improve the robustness and accuracy of the predictions. Furthermore, the integration of explainable AI techniques can provide transparency into the decision-making process of these models, increasing their trustworthiness and utility for policymakers.

Author Contributions

Funding

The authors present their appreciation to King Saud University for funding this research through the Researchers Supporting Program number (RSPD2024R1052), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement

Not applicable

Informed Consent Statement

Not applicable

Data Availability Statement

Data will be available on suitable request from corresponding author.

Acknowledgments

Thanks to reviewers and editors for providing valuable suggestions during the review process.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. BIKIS A. Urban Air Pollution and Greenness in Relation to Public Health. Journal of Environmental and Public Health, **2023**, 8516622, **2023**.
2. ZHAN C., XIE M., LU H., LIU B., WU Z., WANG T., ZHUANG B., LI M., LI S. Impacts of urbanization on air quality and the related health risks in a city with complex terrain. Atmospheric. Chemistry and Physics, **23** (1), 771, **2023**.
3. SEBASTIÃO B.F., HORTELÃO R.M., GRANADAS S.S., FARIA J.M., PINTO J.R., HENRIQUES H.R. Air quality self-management in asthmatic patients with COPD: An integrative review for developing nursing interventions to prevent exacerbations. International Journal of Nursing Sciences, **11** (1), 46, **2024**.
4. BARUA S., NATH S.D. The impact of COVID-19 on air pollution: Evidence from global data. Journal of Cleaner Production, **298**, 126755, **2021**.
5. MANISALIDIS I., STAVROPOULOU E., STAVROPOULOS A., BEZIRTZOGLOU E. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, **8**, **2020**.
6. ABUBAKAR I.R., MANIRUZZAMAN K.M., DANO U.L., ALSHIHRI F.S., ALSHAMMARI M.S., AHMED S.M.S., AL-GEHLANI W.A.G., ALRAWAF T.I. Environmental Sustainability Impacts of Solid Waste Management Practices in the Global South. International Journal of Environmental Research and Public Health, **19** (19), 12717, **2022**.

7. GUL H., DAS B. The Impacts of Air Pollution on Human Health and Well-Being: A Comprehensive Review. Journal of Environmental Impact and Management Policy, 1, **2023**.
8. ENKH-UNDRAA D., KANDA S., SHIMA M., SHIMONO T., MIYAKE M., YODA Y., NAGNII S., NISHIYAMA T. Coal burning-derived SO(2) and traffic-derived NO(2) are associated with persistent cough and current wheezing symptoms among schoolchildren in Ulaanbaatar, Mongolia. Environmental Health and Preventive Medicine, **24** (1), 66, **2019**.
9. BĂLĂ G.P., RÂJNOVEANU R.M., TUDORACHE E., MOTIȘAN R., OANCEA C. Air pollution exposure-the (in) visible risk factor for respiratory diseases. Environmental Science Pollution Research, **28** (16), 19615, **2021**.
10. ZHENG X.Y., ORELLANO P., LIN H.L., JIANG M., GUAN W.J. Short-term exposure to ozone, nitrogen dioxide, and sulphur dioxide and emergency department visits and hospital admissions due to asthma: A systematic review and meta-analysis. Environmental International, **150**, 106435, **2021**.
11. ZHAO Y., KONG D., FU J., ZHANG Y., CHEN Y., LIU Y., CHANG Z., LIU Y., LIU X., XU K., JIANG C., FAN Z. Increased Risk of Hospital Admission for Asthma in Children From Short-Term Exposure to Air Pollution: Case-Crossover Evidence From Northern China. Frontiers in Public Health, **9**, 798746, **2021**.
12. SURIT P., WONGTANASARASIN W., BOONNAG C., WITTAYACHAMNANKUL B. Association between air quality index and effects on emergency department visits for acute respiratory and cardiovascular diseases. PLoS One, **18** (11), e0294107, **2023**.
13. PRAKASH J., AGRAWAL S.B., AGRAWAL M. Global Trends of Acidity in Rainfall and Its Impact on Plants and Soil. Journal of Soil Science and Plant Nutrition, **23** (1), 398, **2023**.
14. GRENNFELT P., ENGLERYD A., FORSIUS M., HOV Ø., RODHE H., COWLING E. Acid rain and air pollution: 50 years of progress in environmental science and policy. Ambio, **49** (4), 849, **2020**.
15. LEE Y.G., LEE P.H., CHOI S.M., AN M.H., JANG A.S. Effects of Air Pollutants on Airway Diseases. International Journal of Environmental Research and Public Health, **18** (18), **2021**.
16. WANG C., WOLTERS P.J., CALFEE C.S., LIU S., BALMES J.R., ZHAO Z., KOYAMA T., WARE L.B. Long-term ozone exposure is positively associated with telomere length in critically ill patients. Environmental International, **141**, 105780, **2020**.
17. WANG Z., LIU J., WANG B., ZHANG B., DENG N. Health benefits from risk information of air pollution in China. Scientific Reports, **13** (1), 15432, **2023**.
18. GRIGORIEVA E., LUKYANETS A. Combined Effect of Hot Weather and Outdoor Air Pollution on Respiratory Health: Literature Review. Atmosphere, **12** (6), 790, **2021**.
19. KEARL Z., VOGEL J. Urban extreme heat, climate change, and saving lives: Lessons from Washington state. Urban Climate, **47**, 101392, **2023**.
20. TIAN X., GUAN T., GUO Y., ZHANG G., KONG J. Selective Susceptibility of Oligodendrocytes to Carbon Monoxide Poisoning: Implication for Delayed Neurologic Sequelae (DNS). Frontiers in Psychiatry, **11**, 815, **2020**.
21. ROCA-BARCELÓ A., CRABBE H., GHOSH R., FRENI-STERRANTINO A., FLETCHER T., LEONARDI G., HOGE C., HANSELL A.L., PIEL F.B. Temporal trends and demographic risk factors for hospital admissions due to carbon monoxide poisoning in England. Preventive Medicine, **136**, 106104, **2020**.
22. WEAVER L.K. Carbon monoxide poisoning. Undersea Hyperbaric Medical society, **47** (1), 151, **2020**.
23. NING K., ZHOU Y.Y., ZHANG N., SUN X.J., LIU W.W., HAN C.H. Neurocognitive sequelae after carbon monoxide poisoning and hyperbaric oxygen therapy. Medical Gas Research, **10** (1), 30, **2020**.
24. ALHARTHY N., ALANAZI A., ALMOQAYTIB A., ALHARBI B., ALSHAIBANI R., ALBUNIYAN J., ALSHIBANI A. Demographics and clinical characteristics of carbon monoxide poisoning for patients attending in the emergency department at a tertiary hospital in Riyadh, Saudi Arabia. International Journal of Emergency Medicine, **17** (1), 25, **2024**.
25. YU J., LEE J., CHO Y., OH J., KANG H., LIM T.H., KO B.S. Correlation between Carboxyhemoglobin Levels Measured by Blood Gas Analysis and by Multiwave Pulse Oximetry. Journal of Personalized Medicine, **14** (2), 168, **2024**.
26. GAO X., WEI W., YANG G.D. Clinical factors for delayed neuropsychiatric sequelae from acute carbon monoxide poisoning: a retrospective study. Front Med (Lausanne), **11**, 1333197, **2024**.
27. BHATTI U.A., HASHMI M.Z., SUN Y., MASUD M., NIZAMANI M.M. Editorial: Artificial intelligence applications in reduction of carbon emissions: Step towards sustainable environment. Frontiers in Environmental Science, **11**, **2023**.
28. AWEWOMOM J., DZEBLE F., TAKYI Y.D., ASHIE W.B., ETTEY E.N.Y.O., AFUA P.E., SACKEY L.N.A., OPOKU F., AKOTO O. Addressing global environmental pollution using environmental control techniques: a focus on environmental policy and preventive environmental management. Discover Environment, **2** (1), 8, **2024**.
29. JONIDI JAFARI A., CHARKHLOO E., PASALARI H. Urban air pollution control policies and strategies: a systematic review. Journal of Environmental Health Sciences Engineering, **19** (2), 1911, **2021**.
30. RAKHOLIA R., LE Q., VU K., HO B.Q., CARBAJO R.S. AI-based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. Urban Climate, **46**, 101315, **2022**.
31. SHEN Z., ZHANG Z., CUI L., XIA Z., ZHANG Y. Coordinated change of PM2.5 and multiple landscapes based on spatial coupling model: Comparison of inland and waterfront cities. Environmental Impact Assessment Review, **102**, 107194, **2023**.
32. SYED A.S., SIERRA-SOSA D., KUMAR A., ELMAGHRABY A. IoT in Smart Cities: A Survey of Technologies, Practices and Challenges. Smart Cities, **4** (2), 429, **2021**.
33. ALAHI M.E.E., SUKKUEA A., TINA F.W., NAG A., KURDTHONGMEE W., SUWANNARAT K., MUKHOPADHYAY S.C. Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends. Sensors, **23** (11), 5206, **2023**.
34. RAKHOLIA R., LE Q., VU K.H.N., HO B.Q., CARBAJO R.S. Outdoor air quality data for spatiotemporal analysis and air quality modelling in Ho Chi Minh City, Vietnam: A part of HealthyAir Project. Data Brief, **46**, 108774, **2023**.
35. CHRISTAKIS I., TSAKIRIDIS O., KANDRIS D., STAVRAKAS I. Air Pollution Monitoring via Wireless Sensor Networks: The Investigation and Correction of the Aging Behavior of Electrochemical Gaseous Pollutant Sensors. Electronics, **12** (8), 1842, **2023**.
36. BHATTI U., MENGXING H., NEIRA MOLINA H., MARJAN S., BARYALAI M., HAO T., WU G., BAZAI S. MFFCG – Multi feature fusion for hyperspectral image classification using graph attention network. Expert Systems with Applications, **229**, 120496, **2023**.
37. LI Y., GUO J.-E., SUN S., LI J., WANG S., ZHANG C. Air quality forecasting with artificial intelligence techniques: A scientometric and content analysis. Environmental Modelling & Software, **149**, 105329, **2022**.
38. BHATTI U., TANG H., WU G., MARJAN S., HUSSAIN A. Deep Learning with Graph Convolutional Networks:

An Overview and Latest Applications in Computational Intelligence. International Journal of Intelligent Systems, **2023**, **2023**.

39. MÉNDEZ M., MERAYO M.G., NÚÑEZ M. Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 1, **2023**.
40. RAVINDIRAN G., HAYDER G., KANAGARATHINAM K., ALAGUMALAI A., SONNE C. Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. Chemosphere, **338**, 139518, **2023**.
41. MAO W., WANG W., JIAO L., ZHAO S., LIU A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. Sustainable Cities and Society, 102567, **2020**.
42. ESSAMLALI I., NHAILA H., EL KHAILI M. Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review. Sustainability, **16** (3), 976, **2024**.
43. IMAM M., ADAM S., DEV S., NESA N. Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment. Intelligent Systems with Applications, **22**, 200333, **2024**.
44. SHETTY C., SHEDOLE S., SOWMYA B.J, NANDALIKE R., SHIVASHANKAR S., DAYANADA P., ROHITH S., VISHWANATH Y., RANJAN R., GOUD V. A Machine Learning Approach for Environmental Assessment on Air Quality and Mitigation Strategy. Journal of Engineering. **2024**, **2024**.
45. MOURSI A.S., EL-FISHAWY N., DJAHEL S., SHOUMAN M.A. An IoT enabled system for enhanced air quality monitoring and prediction on the edge. Complex & Intelligent Systems, **7** (6), 2923, **2021**.
46. CHEN L., HAN B., WANG X., ZHAO J., YANG W., YANG Z. Machine Learning Methods in Weather and Climate Applications: A Survey. Applied Sciences, **13** (21), 12019, **2023**.
47. ZIMELEWICZ E., KALINOWSKI M., MÉNDEZ FERNÁNDEZ D., GIRAY G., ALVES A., LAVESSON N., AZEVEDO K., VILLAMIZAR H., ESCOVEDO T., LOPES H., BIFFL S., MUSIL J., FELDERER M., WAGNER S., BALDASSARRE T., GORSCHEK T. ML-Enabled Systems Model Deployment and Monitoring: Status Quo and Problems. arXiv.org, **2024**.
48. ALDOSERI A., AL-KHALIFA K.N., HAMOUDA A.M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Applied Sciences, **13** (12), 7082, **2023**.
49. BHATTI U.A., MASUD M., BAZAI S.U., TANG H. Editorial: Investigating AI-based smart precision agriculture techniques. Frontiers in Plant Science, **14**, **2023**.
50. AHMED Z., MOHAMED K., ZEESHAN S., DONG X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford), **2020**, **2020**.
51. NAJDAWI F., VILLARREAL R. Utilizing the Vector Autoregression Model (VAR) for Short-Term Solar Irradiance Forecasting. Energy and Power Engineering, **15**, 353, **2023**.
52. IHNE-SCHUBERT S.M., KIRCHER M., WERNER R.A., LAPA C., EINSELE H., GEIER A., SCHUBERT T. Vector autoregression: Useful in rare diseases?-Predicting organ response patterns in a rare case of secondary AA amyloidosis. PLoS One, **18** (8), e0289921, **2023**.
53. ZHANG L., WANG R., LI Z., LI J., GE Y., WA S., HUANG S., LV C. Time-Series Neural Network: A High-Accuracy Time-Series Forecasting Method Based on Kernel Filter and Time Attention. Information, **14** (9), 500, **2023**.
54. KUMAR K., PANDE B.P. Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, **20** (5), 5333, **2023**.
55. GUPTA S., MOHTA Y., HEDA K., ARMAAN R., VALARMATHI B., GANESHAN A. Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health, **2023,** 1, **2023**.
56. BILBAN M., GÖKMEN M. Support Vector Machine, Gradient Boosting and Artificial Neural Network Techniques in Internal Combustion Engine Tests: A Review. Renewable Energy Sources Energy Policy and Energy Management. **2** (1), **2021**.
57. HASSAN M., BESHR E. Predicting soil cone index and assessing suitability for wind and solar farm development in using machine learning techniques. Scientific Reports, **14** (1), 2924, **2024**.
58. BENTI N.E., CHAKA M.D., SEMIE A.G. Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects. Sustainability, **15** (9), 7087, **2023**.
59. YAZDANI A., ZAHMATKESHAN M., RAVANGARD R., SHARIFIAN R., SHIRDELI M. Supervised Machine Learning Approach to COVID-19 Detection Based on Clinical Data. Medical_Journal of the Islamic Republic of Iran, **36,** 110, **2022**.
60. GONG J., DING L., LU Y., QIONG Z., YUN L., BEIDI D. Scientometric and multidimensional contents analysis of PM(2.5) concentration prediction. Heliyon, **9** (3), e14526, **2023**.
61. MA J., MA X., YANG C., XIE L., ZHANG W., LI X. An Air Pollutant Forecast Correction Model Based on Ensemble Learning Algorithm. Electronics, **12** (6), 1463, **2023**.
62. SHAZIAYANI W.N., UL-SAUFIE A.Z., MUTALIB S., MOHAMAD NOOR N., ZAINORDIN N.S. Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach. Atmosphere, **13** (4), 538, **2022**.
63. XIA W., JIANG Y., CHEN X., ZHAO R. Application of machine learning algorithms in municipal solid waste management: A mini review. Waste Management & Research, **40** (6), 609, **2022**.
64. FAN Z., YAN Z., WEN S. Deep Learning and Artificial Intelligence in Sustainability: A Review of SDGs, Renewable Energy, and Environmental Health. Sustainability, **15** (18), 13493, **2023**.
65. TAYE M.M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers, **12** (5), 91, **2023**.
66. BHATTI U.-A., BAZAI S.-U., HUSSAIN S., FAKHAR S., KU C.-S., MARJAN S., YEE P.-L., JING L. Deep Learning-Based Trees Disease Recognition and Classification Using Hyperspectral Data. Computers, Materials \& Continua. FRONTIERS in PLANT SCIENCE, **77** (1), 681, **2023**.
67. SHRESTHA Y., KRISHNA V., KROGH G. Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. Journal of Business Research, **123**, 588, **2021**.
68. ZHUANG W., LI Z., WANG Y., XI Q., XIA M. GCN–Informer: A Novel Framework for Mid-Term Photovoltaic Power Forecasting. Applied Sciences, **14** (5), 2181, **2024**.
69. KOPALIDIS T., SOLACHIDIS V., VRETOS N., DARAS P. Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information, **15** (3), 135, **2024**.
70. LIN S.-L. Application of empirical mode decomposition to improve deep learning for US GDP data forecasting. Heliyon, **8** (1), e08748, **2022**.
71. QUINN A.J., LOPES-DOS-SANTOS V., DUPRET D., NOBRE A.C., WOOLRICH M.W. EMD: Empirical Mode Decomposition and Hilbert-Huang Spectral Analyses in Python. Journal Open Source Software, **6** (59), **2021**.