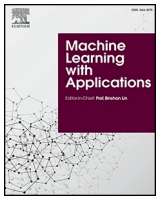
Machine Learning with Applications 15 (2024) 100521

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa



Spatiotemporal integration of GCN and E-LSTM networks for $PM_{2.5}$ forecasting

Ali Kamali Mohammadzadeh [a], Halima Salah [b], Roohollah Jahanmahin [a], Abd E Ali Hussain [c], Sara Masoud [a,*], Yaoxian Huang [b]

[a] Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48201 USA
[b] Department of Civil and Environmental Engineering, Wayne State University, Detroit, MI 48201 USA
[c] Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48201 USA

ARTICLE INFO

Keywords:
$PM_{2.5}$ AQI forecasting
Graph convolutional neural network
Exogenous long short-term memory
Air quality prediction
Urban air pollution

ABSTRACT

$PM_{2.5}$, inhalable particles, with a size of 2.5 micrometers or less, negatively impact the environment as well as our health. Monitoring $PM_{2.5}$ is critical to guard against extreme events by alerting people and initiating actions to alleviate $PM_{2.5}$'s impacts. Developing $PM_{2.5}$ forecasting frameworks empowers the authorities to predict extremely polluted events in advance and gives them time to implement necessary strategies in advance (e.g., Action! Days). Understanding the spatiotemporal behavior of $PM_{2.5}$ and meteorological factors is of significance for having accurate predictions. This study utilizes EPA sensor data to quantify the $PM_{2.5}$ air quality index (AQI) and meteorological factors such as temperature over 2015–2019 across Michigan, USA. Here, a spatiotemporal deep neural structure is proposed through integrating graph convolutional neural (GCN) and exogenous long short-term memory (E-LSTM) networks to incorporate spatial and temporal patterns within $PM_{2.5}$ AQI and meteorological factors for predicting $PM_{2.5}$ AQI. Results illustrate that not only does our proposed framework outperform the traditional approaches such as LSTM and E-LSTM, but also it is robust against the network structure of EPA stations. The study's findings demonstrate that the integration of GCN with E-LSTM significantly enhances the accuracy of $PM_{2.5}$ AQI predictions compared to traditional models. This advancement indicates a promising direction for environmental monitoring, offering improved forecasting tools that can aid in timely and effective decision-making for air quality management and public health protection.

1. Introduction

Atmospheric particulate matter (PM) consists of extremely fine solid particles and minuscule liquid droplets that are prevalent in the Earth's atmosphere (Xing et al., 2016). These particles exhibit a diverse range of sizes, leading to their classification based on their aerodynamic diameter. PM with diameters of 10, 2.5, and 1 micrometer (μm) are denoted as PM_{10}, $PM_{2.5}$, and PM_1, respectively. PM is composed of a mixture of components including sea salt, nitrates, organic carbon, sulfate, black carbon, ammonium, and dust (Crippa et al., 2019). The health implications of PM exposure are significant and contingent upon factors such as size, composition, source, and solubility. In a notable inverse relationship, smaller PM particles pose higher health risks. When inhaled, these diminutive particles can penetrate deeply into the respiratory system and even enter the bloodstream, giving rise to serious health concerns. For instance, prolonged exposure to $PM_{2.5}$ elevates the likelihood of developing cardiovascular disorders and lung cancer (Crippa et al., 2019). Notably, the United States Environmental Protection Agency (USEPA) underscores the adverse impact of $PM_{2.5}$ on the natural environment. Given its role as a primary contributor to haze pollution, $PM_{2.5}$ has the potential to inflict harm upon various ecosystems, contingent on its specific chemical composition (Sun et al., 2006).

The impact of $PM_{2.5}$ on human health remains a significant concern, particularly in urban environments. Across numerous urban areas in the United States, there has been a concerning rise in childhood asthma cases, with $PM_{2.5}$ identified as a contributing trigger (Masoud et al., 2021). Southeast Michigan, in particular, has maintained a vigilant, over 40-year-long monitoring program for PM levels (Michigan Department of Environment, 2021). While Michigan currently adheres to all PM regulatory requirements, there have been historical instances of non-attainment concerning $PM_{2.5}$ levels in Southeast Michigan (Michigan Department of Environment, 2021). For example, Detroit faces unique

* Corresponding author.
E-mail address: saramasoud@wayne.edu (S. Masoud).

https://doi.org/10.1016/j.mlwa.2023.100521
Received 19 October 2023; Received in revised form 5 December 2023; Accepted 14 December 2023
Available online 15 December 2023

air quality challenges as it is situated amidst several interstate roads related to the US-Canada international crossing, and is impacted by emissions from key industrial facilities such as the Marathon oil refinery, the US Steel plant, and DTE Energy's River Rouge Power plant. This has resulted in Detroit having the highest rate of preterm births (PTB) among major U.S. cities (Cleary et al., 2017). Given the disproportionate exposure of residents in Macomb, Wayne, and Oakland counties to ambient air pollutants, particularly $PM_{2.5}$, monitoring the Air Quality Index (AQI) for $PM_{2.5}$ becomes imperative (Schulz et al., 2018). There is compelling evidence highlighting that individuals who are Black or Hispanic, as well as those with lower to moderate incomes, bear a disproportionate burden of air pollution in Michigan (Schulz et al., 2018). These disparities underscore the urgency of addressing $PM_{2.5}$ pollution in a manner that promotes environmental justice and safeguards the health of vulnerable populations.

Despite ongoing efforts in $PM_{2.5}$ monitoring and prediction, there remains a gap in accurately forecasting its spatiotemporal dynamics, particularly in urban areas like Detroit with complex pollution sources. This paper endeavors to bridge this gap and fulfill the imperative need for precise forecasting of $PM_{2.5}$ concentrations through the development of a spatiotemporal forecasting model that leverages Graph Convolutional Networks (GCN) and Exogenous Long Short-Term Memory (E-LSTM) networks, designed to capture both the spatial distribution and temporal progression of $PM_{2.5}$ levels, taking into account the impact of local meteorological conditions. This work is distinct in its integration of these advanced neural network techniques, which have not been extensively applied in the realm of air quality forecasting. The driving force behind this research is the commitment to equip stakeholders, especially local residents, with timely insights into episodes of high air pollution. By proactively mitigating excessive air pollution exposure through predictive modeling, this study aspires to deliver substantial health benefits to the vulnerable population. Given the pivotal role that $PM_{2.5}$ plays in shaping the air quality and public health at both local (e.g., Detroit) and regional (e.g., Michigan) levels, it becomes imperative to scientifically construct a framework capable of accurately modeling and forecasting the spatiotemporal dynamics of $PM_{2.5}$ mass concentrations (Huang et al., 2021). On the other hand, meteorological factors such as wind speed, precipitation, relative humidity, and air temperature can significantly influence $PM_{2.5}$ concentrations throughout the year. Extensive research underscores the substantial temporal and spatial variations in $PM_{2.5}$ levels, attributable to the movement of air pollutants across diverse spatiotemporal scales (Huang et al., 2015). These variations manifest locally, originating from primary sources, and extend to broader scales through secondary reactions and transport mechanisms. Temporally, the fluctuations in air pollution are shaped by meteorological conditions and anthropogenic emissions stemming from human activities (Beckerman et al., 2013). Consequently, the development of a spatiotemporal approach fosters a deeper comprehension of $PM_{2.5}$ dynamics and facilitating accurate forecasting. The potential outcome of this research is a more accurate, real-time prediction model that can inform timely public health responses and policy decisions, particularly in areas facing disproportionate air pollution impacts. This represents a significant advancement in the field of environmental monitoring and public health protection.

This work makes contributions by applying GCN and E-LSTM networks to explore the spatiotemporal forecasting of $PM_{2.5}$ AQI, offering a novel solution to tackle the challenges inherent in $PM_{2.5}$ prediction. Furthermore, our approach incorporates meteorological data into the model to capture the influence of weather conditions on $PM_{2.5}$ concentrations, thereby enhancing the precision of our forecasts. Ultimately, we will rigorously assess and compare the performance of our proposed model against existing methods, providing compelling evidence of its efficacy in predicting $PM_{2.5}$ AQI.

The remainder of this paper is structured as follows. The literature review provides an overview of previous research pertaining to $PM_{2.5}$ forecasting and spatiotemporal modeling. The methodology section elucidates our approach to data acquisition and preprocessing, as well as the development of our forecasting model utilizing GCN and E-LSTM networks. The results and discussion section comprehensively presents our findings and conducts an in-depth evaluation of the model's performance. Finally, the conclusion succinctly summarizes the key contributions of this work, outlines its implications, and points towards future research directions.

2. Literature review

In recent years, various domains, including healthcare (Etu et al., 2022), agro-industry (Masoud et al., 2019), and manufacturing (Chowdhury et al., 2021), have widely embraced machine learning and deep learning methods. Researchers have harnessed these techniques to explore the connection between $PM_{2.5}$ levels and meteorological factors, as well as to predict $PM_{2.5}$ concentrations and their corresponding Air Quality Index (AQI) worldwide (Wang et al., 2021). To create a predictive model for $PM_{2.5}$ levels using meteorological variables as inputs, Wang et al. (2021) recommended employing the backpropagation artificial neural network (ANN). The authors compared the projected $PM_{2.5}$ values with actual measurements derived from $PM_{2.5}$ concentration Aerosol Optical Depth (AOD) data collected from 2016 to 2017 in the Fulling district of Chongqing. Their findings revealed a significant negative correlation between wind speed and $PM_{2.5}$ concentration, a meaningful positive association with relative humidity, and a segmented linear relationship with temperature, attributed to diffusion and conversion rates. Surprisingly, the authors reported no clear linear association with precipitation, despite its known purifying effect on $PM_{2.5}$.

For the prediction of daily $PM_{2.5}$ concentrations, Zhu et al. (2018) proposed an approach that combined various techniques, including Complementary Ensemble Empirical Mode Decomposition (CEEMD), Support Vector Regression (SVR), Gravitational Search Algorithm (PSOGSA), Particle Swarm Optimization, Generalized Regression Neural Network (GRNN), and Grey Correlation Analysis (GCA). They tested the viability of this model in three Chinese cities, Jinan, Harbin, and Chongqing, each with distinct climatic, terrain, and pollution-source characteristics. The results suggested that their proposed approach could be effectively employed for air quality prediction and notifications. In the realm of image-based air quality analysis, Chakma et al. (2017) focused on estimating $PM_{2.5}$ concentrations. Their approach involved utilizing a deep Convolutional Neural Network (CNN) to categorize natural images based on their $PM_{2.5}$ levels. They curated a dataset comprising 591 photos from Beijing, each accompanied by $PM_{2.5}$ values. The model they proposed proved valid for calculating image based $PM_{2.5}$ concentrations.

$PM_{2.5}$ prediction can be categorized into temporal, spatial, and spatiotemporal analyses, with various methods and approaches contributing to our understanding and ability to forecast $PM_{2.5}$ concentrations.

2.1. Temporal analysis

Temporal literature primarily relies on time-dependent information as predictor variables to construct time series forecasting models. For instance, the Multi-Directional Temporal Convolutional Artificial Neural Network (MTCAN) model formulated by Samal et al. (2021), adeptly conducted feature learning and sequential modeling on an extensive historical timeseries dataset for forecasting $PM_{2.5}$ pollution levels over extended periods (Samal et al., 2021). To evaluate their proposed model, the authors conducted performance tests using two authentic datasets: the Multi-Site Air-Quality Dataset from the UCI Machine Learning Repository and the data provided by the Central Pollution Control Board in India (Samal et al., 2021). Based on their findings, it was evident that their proposed model surpassed CNN (Samal et al., 2021). Masoud et al. (2021) employed PurpleAir sensors to measure $PM_{2.5}$ levels and various methodological variables, including humidity. Their investigation

centered on a temporal analysis of PM2.5 dispersion across distinct regions within the Greater Detroit Area, utilizing algorithms such as Dynamic Time Warping (DTW) and LSTM. The results highlighted that LSTM, when augmented with external variables (namely existing $PM_{2.5}$ levels, meteorological characteristics, and atmospheric conditions), achieved accurate predictions of $PM_{2.5}$ quantities (with an average Root Mean Square Error of 3.2 g/m3). However, he authors reported no robust correlation between $PM_{2.5}$ levels and the specified meteorological factors within the Greater Detroit Area were identified (Masoud et al., 2021).

2.2. Spatial analysis

Spatial literature predominantly concentrates on discerning spatial patterns for predicting $PM_{2.5}$ concentrations. Zhao et al. (2019) conducted a comprehensive assessment of the spatial distribution of $PM_{2.5}$ and its alterations between 2015 and 2016 in every prefectural city across China, utilizing data from over 1000 monitoring stations. Before contrasting the changes between the two years, the authors started by examining the spatial distribution of $PM_{2.5}$ values for both 2015 and 2016. Subsequently, the authors delved into the relationship between socioeconomic variables and shifts in $PM_{2.5}$ concentrations. Remarkably, despite existing significant levels of $PM_{2.5}$ pollution in many cities, their findings revealed that a majority of cities in eastern China experienced a reduction in $PM_{2.5}$ concentration. To further explore these associations, the authors conducted an multiple linear regression model, employing socioeconomic characteristics as explanatory variables and changes in $PM_{2.5}$ concentration as response variables (Zhao et al., 2019). Addressing the challenge of varying spatial correlations between predictor variables, such as weather conditions and Aerosol Optical Depth (AOD), with $PM_{2.5}$ concentrations, Zhan et al. (2017) introduced a machine learning approach known as Geographically Weighted Gradient Boosting Machine (GW-GBM). This method demonstrated superior performance compared to the conventional GBM model in predicting daily $PM_{2.5}$ concentrations, even when confronted with partially missing AOD data in China (Zhan et al., 2017).

2.3. Spatiotemporal analysis

Spatiotemporal analysis endeavors to integrate available temporal and spatial data to enhance $PM_{2.5}$ concentration predictions. For instance, Di et al. (2019) employed an ensemble model combining Random Forest, neural networks, and Gradient Boosting algorithms. Various predictors, including satellite data, land-use variables, elevation, meteorological data, reanalysis datasets, and a chemical transport model, were harnessed to compute daily $PM_{2.5}$ levels across the United States from 2000 to 2015 at a 1 km x 1 km resolution. Epidemiological applications were also made to gauge the health effects of $PM_{2.5}$ using this prediction dataset, which encompassed downscaling and uncertainty predictions. Their model performed satisfactorily up to 60 g/m3. Luo et al. (2023) assessed soil erosion in the Three Gorges Reservoir area, China. This region had experienced significant soil loss for an extended period, leading to the implementation of vegetation restoration projects since 1999. They incorporated different scenarios and regional vegetation restoration projects to examine potential changes in land use and climate in the future. By considering these factors, the study aimed to estimate long-term soil erosion changes in the TGR area, contributing to our understanding of how regional policies and socio-economic development influence soil erosion in the context of climate change.

Huang et al. (2015) utilized daily and hourly $PM_{2.5}$ data from the Beijing Environmental Protection Bureau to analyze spatiotemporal patterns from August 2013 to July 2014 in Beijing. The authors also assessed the correlation between $PM_{2.5}$ and meteorological variables, employing various methods such as time-series graphs, Ordinary Kriging interpolation, coefficient of divergence, and Spearman correlation coefficient. The study explored disparities in $PM_{2.5}$ concentrations related to temporal spatial factors using statistical tests and examined connections between weather-related elements and daily $PM_{2.5}$ using the Generalized Additive Mixed Model (GAMM). Their findings revealed significant associations between meteorological elements and $PM_{2.5}$ concentration. Ma et al. (2022) developed a detailed spatiotemporal approach for estimating $PM_{2.5}$, considering data from multiple sources, including $PM_{2.5}$ and weather data collected hourly between April 1, 2015, and August 31, 2019 and Landsat 8 Operational Land Imager multispectral images. Their random forest (RF) approach achieved high spatiotemporal resolution and excellent accuracy, with a cross-validated R^2 of 0.86.

Shogrkhodaei et al. (2021) used stochastic gradient descent, frequency ratio, and AdaBoost to prepare $PM_{2.5}$ risk mapping and spatiotemporal models in Tehran, Iran. The frequency ratio model demonstrated the highest modeling accuracy, especially in autumn and winter. Thongthammachart et al. (2021) developed a land use regression model combining the frequency ratio technique and the Community Multiscale Air Quality (CMAQ) modeling system to predict $PM_{2.5}$ levels in the Kansai area of Japan. Their model incorporating CMAQ variables outperformed the model without CMAQ data.

Bi et al. (2022) used multiscale geographically weighted regression to analyze the relationship between urban green space morphology (UGSM) and $PM_{2.5}$ at different scales in Wuhan. The authors reported that UGSMs significantly influenced $PM_{2.5}$, with the point-line-polygon UGSM having the most effective reduction in forecast errors.

Song et al. (2022) analyzed the spatial and temporal dispersion trends and potential risks associated with exposure to ground-level $PM_{2.5}$ data for China from 2001 to 2020. They noted substantial reductions in $PM_{2.5}$ concentrations in urban agglomerations between 2014 and 2020. Chen et al. (2021) introduced a convolutional recursive neural network to generate $PM_{2.5}$ prediction maps from real-time air quality data in Taiwan. Zhu et al. (2021) used Weibo data to extract PM2.5-related health information and employed GWR models to clarify the relationship between $PM_{2.5}$ and related parameters. Tan et al. (2022) addressed data gaps in $PM_{2.5}$ time series data using a two-step hybrid model called ST-SILM, which combined spatiotemporal modeling with single exponential smoothing and inverse distance weighting. Their approach improved estimation accuracy compared to the original dataset.

Zhang et al. (2021) developed a Spatiotemporal Causal CNN (ST-CausalConvNet) for short-term $PM_{2.5}$ forecasting. They used spatiotemporal correlation analysis to select relevant data from monitoring stations, achieving consistent trends between expected and observed $PM_{2.5}$ concentrations. Pak et al. (2020) proposed a spatiotemporal model, combining CNN and LSTM networks, for daily $PM_{2.5}$ forecasting in Beijing City. Their approach outperformed other models in terms of stability and prediction performance. Qi et al. (2019) used GCN and LSTM to analyze spatiotemporal changes in $PM_{2.5}$ concentrations in China based on data from 76 stations. Jiang et al. (2023) estimated CO_2 emissions using remote sensing night-time light data and observed a reduction in the combined "CO_2- $PM_{2.5}$" effect. The authors identified spatial variations in influencing factors and highlighted the synergy between pollution and carbon reduction through clean energy development.

Rahman and Kabir (2023) analyzed the spatial and temporal distribution of air quality indicators in the greater Dhaka region, forecasting weekly Air Quality Index (AQI) values and evaluating a particulate matter filtration unit's effectiveness. Liu et al. (2021) developed a multi-data-driven spatiotemporal prediction approach for $PM_{2.5}$, using a combination of methods, including graph convolutional networks, LSTM, and Q-learning. Tan et al. (2022) created an Ensemble Graph Attention Reinforcement Learning Recursive Network for $PM_{2.5}$ prediction, outperforming state-of-the-art models. Shi et al. (2023) proposed a Balanced Social LSTM for $PM_{2.5}$ concentration prediction, taking into account the influence of stations with higher concentrations on

nearby areas and the correlation between adjacent stations.

Ades and Pires (2019) assessed spatial and temporal patterns of $PM_{2.5}$ levels in Europe, using threshold models and ANN fine-tuned with Genetic Algorithms (GA). Gokul et al. (2023) predicted $PM_{2.5}$ levels in Hyderabad using various machine learning models, with LSTM outperforming traditional machine learning models. Cheng et al. (2023) evaluated Convolutional LSTM and Deep Convolutional Generative Adversarial Network models for spatiotemporal predictions of ozone concentrations in the Beijing-Tianjin-Hebei region from 2013 to 2018, finding machine learning-based models superior in spatiotemporal performance.

3. Methodology

Our proposed framework comprises three core phases: data acquisition, processing, and analysis, as depicted in Fig. 1. In the data acquisition phase, we gather $PM_{2.5}$ Air Quality Index (AQI) and meteorological data, including parameters like relative humidity and pressure. These data span the years 2015 to 2019, encompassing the State of Michigan. To ensure the integrity of our analysis, we intentionally exclude data from 2020 to 2021 due to potential anomalies stemming from the COVID-19 pandemic. Our data sources include information from the EPA website (United States Environmental Protection Agency, 2022) and an online open weather database (Weather Underground, 2022).

As illustrated in Fig. 1, the initial stage of the data processing phase involves imputation. Here, we address missing values through iterative imputing, which proves highly effective in multi-variant environments. In this method, each feature is modeled as a function of the remaining features, enabling us to reasonably estimate missing values based on the known feature values (Sadhu et al., 2020). To ensure the robustness of predictive models in handling outliers, we employ the winsorization method. Winsorization transforms data by capping extreme values (outliers) rather than discarding them, thus preserving valuable information (Wilcox, 2005). Subsequently, we proceed with feature engineering, which involves formatting features to align with our specific structural requirements. Following this, we conduct correlation analysis to reduce data dimensionality. This step focuses on retaining methodological features that exhibit significant correlations with $PM_{2.5}$ AQI. Consequently, the feature set is narrowed down to include mean relative humidity, pressure, temperature, wind speed, and wind direction.

In the latter part of our proposed framework, we concentrate on data analysis, employing spatiotemporal modeling techniques for $PM_{2.5}$ AQI and meteorological factors. This entails the integration of E-LSTM and GCN networks, facilitating comprehensive data analysis and modeling.

3.1. Exogenous long short-term memory

An exogenous long short-term memory network is a multivariate LSTM that incorporates exogenous variables as predictors. LSTMs are modifications of the original recurrent neural networks (RNN), designed to analyze time-series data, addressing the vanishing gradient problem (Zilly et al., 2017). Fig. 2 illustrates how an LSTM functions based on its three main gates (i.e., input, output, and forget).

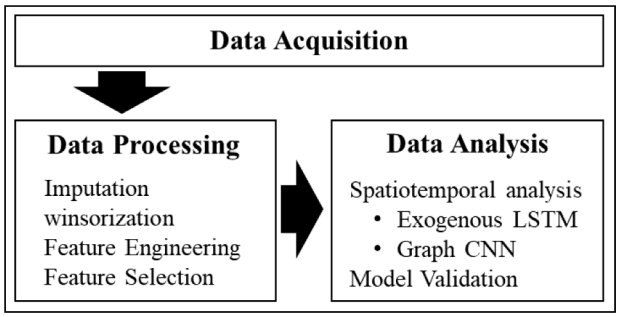


Fig. 1. Our proposed framework.

Given the structure of LSTM cells (three gates and memories), number of weights and biases is around four times of traditional RNNs (Ryu et al. 2022). The function of each Gate in LSTM can be described as follows (Ma et al., 2021). In the forget gate, it is decided which information from the preceding layer should be discarded and/or kept in the current state, as shown in (1).

$$f_t = \left(W_f.[h_{t-1}, x_t] + b_f\right) \tag{1}$$

The input gate updates the information upon entrance using the sigmoid function before deciding which information to store in memory cells, as illustrated by (2).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2}$$

The output gate determines the model's output as well as the proportion of the output of control unit state C_t to be distributed to the hidden layer elements of the model (*i.e.*, h_t). The sigmoid activation function produces the initial output, which is then reduced to a range of -1 to 1 by the tanh function, and then multiplied by the sigmoid output to produce the result as defined in (3) and (4).

$$O_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{3}$$

$$h_t = o_t.tanh(C_t) \tag{4}$$

The memory generates new candidate values utilizing the tanh function, then updates the memory state by combining the input gate's input information with the current state information. It determines what information are saved and what information are communicated to the next step so that it can forecast future data using prior data as demonstrated in (5).

$$\widetilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{5}$$

where W_f, W_i, W_c, and W_o are the weights of the forget gate, input gate, memory cell, and output gate, respectively. h_{t-1} represents the last value of the hidden unit, and x_t represents the input information at the current step. b_f, b_i, b_c, b_o are the biases of the forget gate, input gate, memory cell, and the output gate, respectively.

In the case of traditional LSTM models, the multivariate time-series dataset is represented as a three-dimensional tensor of shape (N, Q, 1), where N is the number of samples, Q is the maximum number of time stamps across variables, and 1 indicates that only a single variable is processed per time step.

However, in this study, we employ an Exogenous LSTM (E-LSTM) model that incorporates meteorological factors as exogenous variables. This extends the dimensionality of the tensor to (N, Q, M), where M represents the number of variables processed per time step, including the meteorological factors. By incorporating these additional variables,

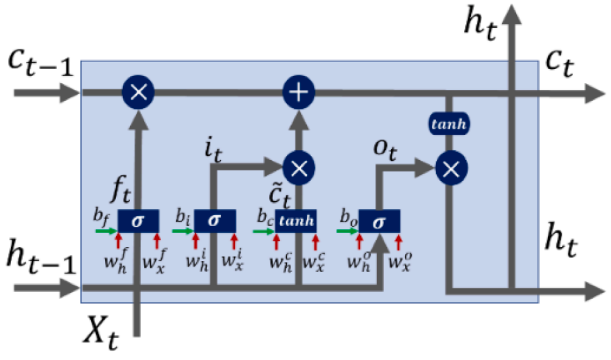


Fig. 2. A single LSTM cell including forget, input, and output gates (Masoud et al. 2019).

the E-LSTM model can capture the influence of meteorological conditions on $PM_{2.5}$ concentrations and enhance the accuracy of the forecasting model.

3.2. Graph convolutional network (GCN)

As a class of deep learning methods suitable for graphical data, GCNs have gained increasing popularity in recent years. GCNs enable deep learning models applicable to problems that can be described by graphs. Such problems can be found in several domains, including biology, natural language processing, medicine, telecommunication, and many more (Grattarola & Alippi, 2021). A GCN encodes the whole body structure of the problem as a graph *(G)* consisting of nodes *(V)* of size *n* and edges *(E)*. By takes advantage of graphs' relational structure, GCN makes better prediction (Haghighat &; Prince, 2021). GCNs, as graph-based multi-layer neural networks, operate on embedding vectors of nodes based upon the properties of their neighborhoods (Kipf & Welling, 2016). A GCN applies two types of operations: message passing and graph pooling. The former learns a non-linear transformation of the input graphs, and the latter reduces their size.

Here, $X \in R^{n\times m}$ reprecents a matrix of all n nodes and their respective features, with m being the feature vectors' dimension. Each row $x_v \in R^m$ denotes the feature vector for v. An adjacency matrix A can be defined with its degree matrix D, where the diagonal elements of A are set to one due to A's self-looping nature. For a one-layer network, $L^{(1)}$, the new k-dimensional node feature matrix is computed with (6).

$$L^{(1)} = \rho(A^{\sim} XW_0) \tag{6}$$

where ρ is the activation function, $W^0 \in R^{m\times k}$ is a weight matrix, and $A^{\sim}$ is the normalized symmetric adjacency matrix. As GCN can only gather information about immediate neighbors per each layer, several GCN layers are stacked to gather information about larger neighborhoods as defined in (7).

$$L^{(j+1)} = \rho\left(A^{\sim} L^{(j)} W_j\right) \tag{7}$$

where $L^{(0)} = X$ and j represent the layer number (Yao et al., 2019).

Our proposed model for this study, as demonstrated in Fig. 3, is an integration of E-LSTM and GCN. This integrated structure helps to better extract both temporal and spatial features of the data.

While the E-LSTM structure relies on time series data of $PM_{2.5}$ AQI and methodological factors, the GCN mainly relies on the adjacency matrix. The two networks process the data in parallel, as displayed in Fig. 3. Then, the outputs of the two networks are concatenated. A series of dense layers further process the data and provide us with the output.

4. Results and discussion

The proposed framework undergoes evaluation using data obtained from EPA monitoring sites spread across Michigan. Among the 13 active EPA sites within the state, seven presented notable gaps in their data records, as indicated by the red markers in Fig. 4. Consequently, our

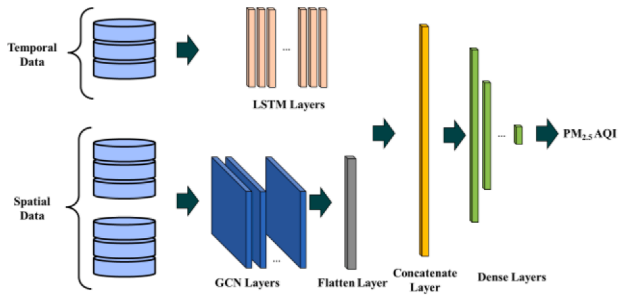


Fig. 3. The block diagram of our proposed E-LSTM-GCN.

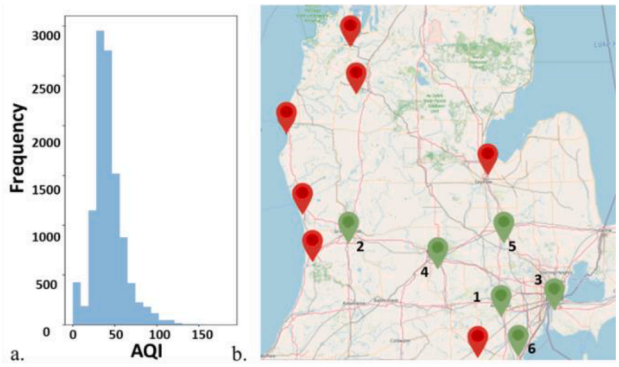


Fig. 4. Histogram of the average of $PM_{2.5}$ AQI (a); EPA active stations in Michigan (b).

analysis centers on the data sourced from the remaining six sites, marked in green on Fig. 4. These active monitoring stations are strategically located in Ann Arbor, Grand Rapids, Detroit, Lansing, Flint, and Monroe, each assigned a numerical code from 1 to 6, respectively. The dataset encompasses comprehensive information, including AQI measurements and various meteorological parameters such as humidity, pressure, and temperature. Data collection spans from January 1st, 2015, through December 31st, 2019, for all six of these operative monitoring stations.

Fig. 4 presents a histogram illustrating $PM_{2.5}$ AQI (a) and the geographical distribution of active EPA stations in Michigan (b). Our study focuses on data collected from stations marked in green, while those marked in red were excluded due to data incompleteness. The dataset utilized in this study exhibits gaps and outliers, necessitating two distinct data preprocessing techniques. Firstly, we employ iterative imputation to address missing values in the dataset. Secondly, we apply Winsorization with a threshold of 0.025 to handle outliers. This approach effectively mitigates the influence of outliers on our analysis while preserving the overall integrity of the data.

Fig. 5 provides a visual representation of the daily means for $PM_{2.5}$ AQI, temperature (°C), humidity (%), pressure (x10^3 pascal), wind direction (°), and wind speed (km/hr) across all six stations during the year 2019.

As shown in Fig. 5, while $PM_{2.5}$ AQI, humidity, pressure, wind direction, and wind speed exhibit stationary behavior, temperature stands out as the only factor displaying autocorrelation. This observation is further substantiated by the Dickey-Fuller test results. Consequently, we conducted an autocorrelation analysis to determine the lagged relationships and subsequently employed iterative differencing to remove the underlying trend in the temperature data. The next step involved

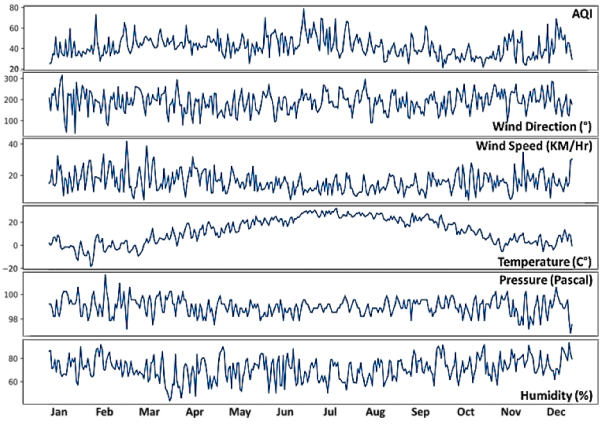


Fig. 5. Averaged time series of processed values of $PM_{2.5}$ AQI, humidity (%), pressure (x10^3 pascal), temperature (°c), wind direction (°), and wind speed (km/hr).

conducting correlation analysis, as illustrated in Fig. 6. Initially, our dataset included daily measurements of maximum, mean, and minimum values for pressure, humidity, wind speed, and wind direction, along with the daily mean temperature. Through the correlation analysis, we identified highly correlated features (i.e., those with correlation scores close to 1 or -1) and subsequently removed them from the dataset. This process helps reduce the feature space to the daily mean values for humidity (in percentage), pressure (in Pascal), temperature (in degrees Celsius), wind direction (in degrees), and wind speed (in kilometers per hour). By selecting these features, we ensured that the essential meteorological information was retained while minimizing the presence of redundant variables in our analysis.

Upon the completion of preprocessing steps, the dataset is partitioned into training (spanning the first 42 months: January 2015 to June 2018), validation (encompassing nine months: July 2018 to March 2019), and testing (covering nine months: April 2019 to December 2019) sets.

Before initiating model training, we determine the forecasting horizon and the training window by considering various combinations, ranging from one to five days of historical data to predict the $PM_{2.5}$ AQI for the next one to five days. After careful evaluation, the optimal combination identified was a training window of three days (including $PM_{2.5}$ AQI and meteorological factors) to forecast the $PM_{2.5}$ AQI for the following day. E-LSTM-GCN models are trained for all six locations, and hyperparameter tuning is performed using a grid search approach.

Fig. 7 illustrates the loss curves for both the training (depicted in blue) and test (depicted in orange) sets for the developed E-LSTM-GCN models. The observed decreases in loss for both training and test sets, as shown in Fig. 7, affirm the model's robust performance across all locations, resulting in the stabilization of the models' predictive capabilities.

To evaluate the performance of the proposed E-LSTM-GCN, the model is compared with traditional LSTM and E-LSTM using different metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), where at each location, three sets of models are used, and the error is calculated as the difference between the predicted and actual

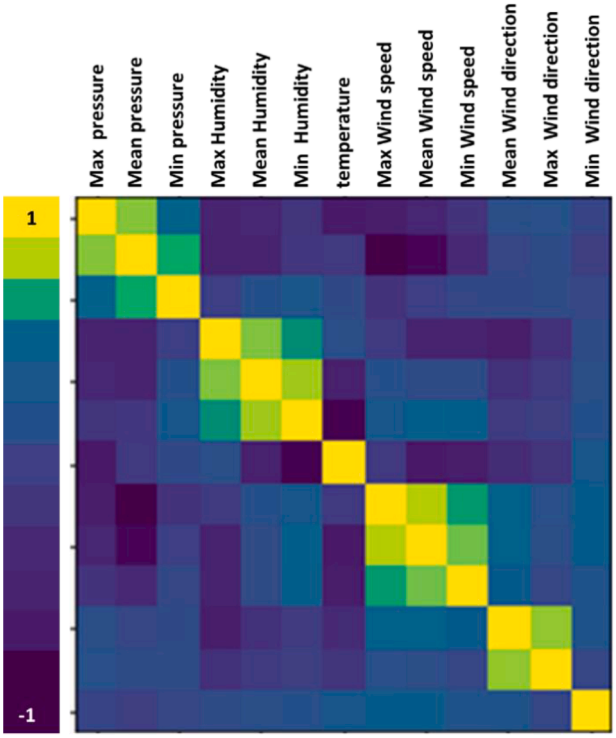


Fig. 6. Correlation analysis for feature selection.

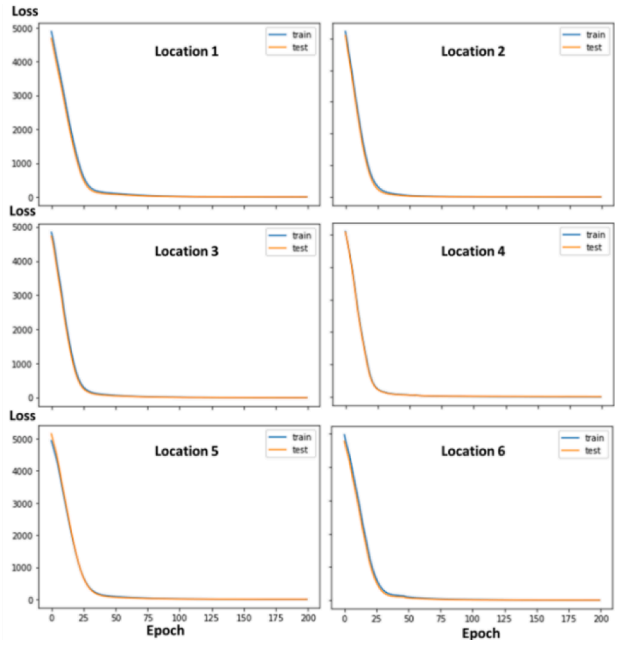


Fig. 7. The loss of the trained E-LSTM-GCN models over the train and test datasets.

values (i.e., LSTM, E-LSTM, and E-LSTM-GCN) are trained using the training dataset, and their MSE, MAE, MSE, and MAPE are calculated using 5-fold validation as demonstrated in Table 1. Table 1 compares the performance of all three models across all locations using the mean and standard deviation of MSE, RMSE, MAE, and MAPE.

Table 1 not only demonstrates the superior performance of our proposed E-LSTM-GCN model for reporting lower values of RMSE, MSE, MAE, and MAPE, but also it is of robustness compared to the other techniques by having the smallest standard deviation in each metric. To better understand the impact of the spatial structure of the adjacent stations, the performances of all models are evaluated using RMSE metric across all locations, as displayed in Fig. 8.

As displayed in Fig. 8, the proposed model outperforms the LSTM and E-LSTM models in each location, portraying its robustness against the structure of the adjacency graph. Fig. 8 demonstrates that although E_LSTM_GCN best performs when the targeted location is surrounded by other locations in all directions (e.g., Location 3, Detroit), it outperforms LSTM and E-LSTM even if the graph structure is not ideal (e.g., Location 6, Monroe).

5. Conclusion

Monitoring and predicting the spatiotemporal variability of $PM_{2.5}$ mass concentrations is of paramount importance, given its potential adverse effects on air quality and public health. The development of $PM_{2.5}$ forecasting frameworks empowers authorities to proactively safeguard citizens by providing lead time to implement preventive measures during periods of extreme pollution events. An in-depth comprehension of the spatiotemporal dynamics of $PM_{2.5}$ and its intricate relationship with meteorological factors is pivotal for enhancing

Table 1
Comparison of the performance (mean ± standard deviation) of the developed LSTM, E-LSTM, and E-LSTM_GCN using RMSE, MAE, MSE, and MAPE.

	RMSE	MSE	MAE	MAPE
LSTM	4.0±0.5	16.3±4.2	3.6±0.4	8.9±1.4
E-LSTM	3.9±0.4	15.6±3.1	3.5±0.4	8.8±1.7
E-LSTM-GCN	3.5±0.3	12.5±1.8	3.2±0.2	8.0±1.7

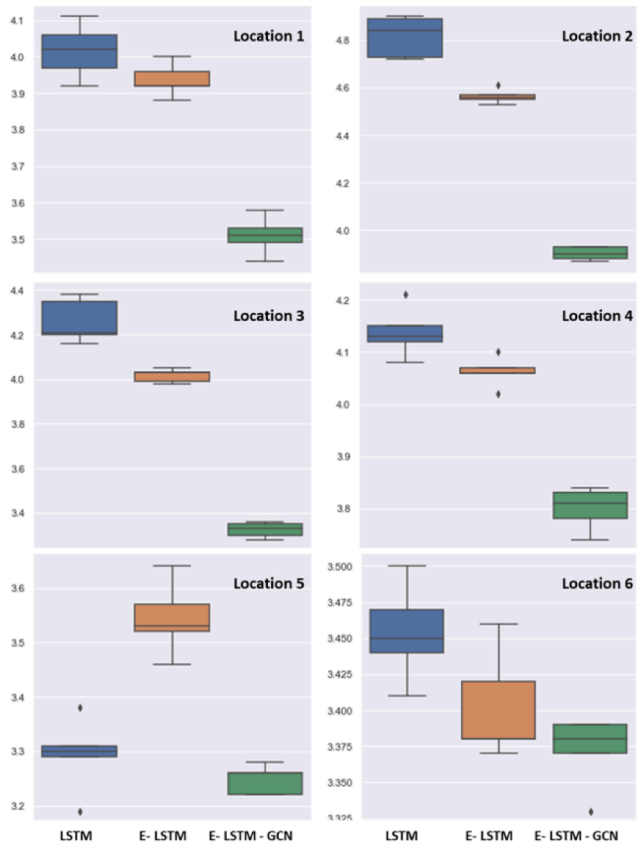


Fig. 8. Comparison of the performances of the developed LSTM, E-LSTM, and E-LSTM_GCN using RMSE at Locations 1 to 6.

forecasting systems. In this study, we harness EPA sensor data to quantitatively assess $PM_{2.5}$ Air Quality Index (AQI) alongside daily mean values of humidity, pressure, temperature, wind direction, and wind speed spanning the years 2015 to 2019 across Michigan, USA. We introduce a novel spatiotemporal deep neural architecture known as E-LSTM-GCN, designed to capture both spatial and temporal patterns within $PM_{2.5}$ AQI and meteorological factors for the prediction of $PM_{2.5}$ AQI. Our findings demonstrate the superiority of our proposed framework over conventional methods, such as LSTM and E-LSTM, as evidenced by an average Root Mean Square Error (RMSE) score of 3.52. Importantly, our framework exhibits resilience across diverse station networks. Specifically, our proposed model yields $PM_{2.5}$ AQI predictions for Ann Arbor, Grand Rapids, Detroit, Lansing, Flint, and Monroe with RMSE scores of 3.51, 3.32, 3.9, 3.25, 3.8, and 3.37, respectively. These results underscore the effectiveness and adaptability of our approach in $PM_{2.5}$ forecasting tasks.

Furthermore, the practical implications of our research are significant for urban planners and public health officials, offering a reliable tool for air quality management and policy formulation. The predictive capabilities of our model can be instrumental in formulating timely health advisories, urban planning decisions, and environmental policies. For future research, we propose further exploration into integrating additional environmental and socio-economic factors into the forecasting model to understand their impact on $PM_{2.5}$ levels. Another promising avenue is the application of our model to other regions with different climatic conditions and industrial landscapes, to test its adaptability and accuracy in varied environmental settings. Lastly, we see potential in exploring real-time data integration for dynamic and up-to-the-minute $PM_{2.5}$ forecasting, which could further enhance the utility of this model in immediate public health response scenarios.

CRediT authorship contribution statement

Ali Kamali Mohammadzadeh: Data curation, Formal analysis, Writing – original draft. **Halima Salah:** Data curation, Formal analysis, Writing – original draft. **Roohollah Jahanmahin:** Methodology, Writing – original draft. **Abd E Ali Hussain:** Data curation, Formal analysis. **Sara Masoud:** Writing – original draft, Project administration, Supervision. **Yaoxian Huang:** Project administration, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work has been partially funded by the Richard Barber Interdisciplinary Research Program for Undergraduates at Wayne State University.

References

Adães, J., & Pires, J. C. M. (2019). Analysis and modelling of PM2. 5 temporal and spatial behaviors in European cities. *Sustainability, 11*(21), 6019.

Beckerman, B. S., Jerrett, M., Serre, M., Martin, R. V., Lee, S. J., Van Donkelaar, A., … Burnett, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM2. 5 in the contiguous United States. *Environmental science & technology, 47*(13), 7233–7241.

Bi, S., Dai, F., Chen, M., & Xu, S. (2022). A new framework for analysis of the morphological spatial patterns of urban green space to reduce PM2. 5 pollution: A case study in Wuhan, China. *Sustainable Cities and Society, 82*, Article 103900.

Chakma, A., Vizena, B., Cao, T., Lin, J., & Zhang, J. (2017). Image-based air quality analysis using deep convolutional neural network. In *Proceedings of the IEEE international conference on image processing (ICIP)* (pp. 3949–3952). IEEE.

Chen, H.C., Trinanda Putra K., and Chun-WeiLin J.. "A novel prediction approach for exploring PM2. 5 spatiotemporal propagation based on convolutional recursive neural networks." arXiv preprint arXiv:2101.06213 (2021).

Cheng, M., Fang, F., Navon, I. M., Zheng, J., Zhu, J., & Pain, C. (2023). Assessing uncertainty and heterogeneity in machine learning-based spatiotemporal ozone prediction in Beijing-Tianjin- Hebei region in China. *Science of the Total Environment, 881*. https://doi.org/10.1016/j.scitotenv.2023.163146

Chowdhury, B. D. B., Masoud, S., Son, Y. J., Kubota, C., & Tronstad, R. (2021). A dynamic HMM-based real-time location tracking system utilizing UHF passive RFID. *IEEE Journal of Radio Frequency Identification, 6*, 41–53.

Cleary, E., Asher, M., Olawoyin, R., & Zhang, K. (2017). Assessment of indoor air quality exposures and impacts on respiratory outcomes in River Rouge and Dearborn, Michigan. *Chemosphere, 187*, 320–329.

Crippa, M., Janssens-Maenhout, G., Guizzardi, D., Van Dingenen, R., & Dentener, F. (2019). Contribution and uncertainty of sectorial and regional emissions to regional and global $PM_{2.5}$ health impacts. *Atmospheric Chemistry and Physics, 19*(7), 5165–5186.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Benjamin Sabath, M., et al. (2019). An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International, 130*, Article 104909.

Etu, E. E., Monplaisir, L., Masoud, S., Arslanturk, S., Emakhu, J., Tenebe, I., Joseph, B. M., Tom Hagerman, D. J., & Seth, K. (2022). A comparison of univariate and multivariate forecasting models predicting emergency department patient arrivals during the COVID-19 pandemic. *Healthcare, 10*(6), 1120. MDPI.

Gokul, P. R., Mathew, A., Bhosale, A., & Nair, A. T. (2023). Spatiotemporal air quality analysis and PM2.5 prediction over Hyderabad City, India using artificial intelligence techniques. *Ecological Informatics.* , Article 102067. https://doi.org/10.1016/j.ecoinf.2023.102067

Grattarola, D., & Alippi, C. (2021). Graph neural networks in TensorFlow and keras with spektral [application notes]. *IEEE Computational Intelligence Magazine, 16*(no. 1), 99–106.

Haghighat, P., Prince, A., & Jeong, H. (2021). Graph convolutional networks for exercise motion classification. In , *65. Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 685–689). SAGE Publications.

Huang, Y., Partha, D. B., Harper, K., & Heyes, C. (2021). Impacts of global solid biofuel stove emissions on ambient air quality and human health. *GeoHealth, 5*(3), Article e2020GH000362.

Huang, F., Li, X., Wang, C., Xu, Q., Wang, W., Luo, Y., et al. (2015). PM2. 5 spatiotemporal variations and the relationship with meteorological factors during 2013-2014 in Beijing, China. *PloS One, 10*(11), Article E0141642.

Jiang, F., Chen, B., Li, P., Jiang, J., Zhang, Q., Wang, J., et al. (2023). Spatio-temporal evolution and influencing factors of synergizing the reduction of pollution and carbon emissions - utilizing multi-source remote sensing data and GTWR model. *Environmental Research, 229*, Article 115775. https://doi.org/10.1016/j.envres.2023.115775

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint* arXiv:1609.02907.

Liu, X., Qin, M., He, Y., Mi, X., & Yu, C. (2021). A new multi-data-driven spatiotemporal PM2. 5 forecasting model based on an ensemble graph reinforcement learning convolutional network. *Atmospheric Pollution Research, 12*(10), Article 101197.

Luo, Z., Chen, X., Li, N., Li, J., Zhang, W., & Wang, T. (2023). Spatiotemporal foresting of soil erosion for SSP-RCP scenarios considering local vegetation restoration project: A case study in the three gorges reservoir (TGR) area, China. *Journal of Environmental Management, 337*. https://doi.org/10.1016/j.jenvman.2023.117717

Ma, P., Tao, F., Gao, L., Leng, S., Yang, K., & Zhou, T. (2022). Retrieval of Fine-Grained PM2. 5 Spatiotemporal Resolution Based on Multiple Machine Learning Models. *Remote Sensing, 14*(3), 599.

Ma, R., Zheng, X., Wang, P., Liu, H., & Zhang, C. (2021). The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. *Scientific Reports, 11*(1), 1–14.

Masoud, S., Chowdhury B., Son Y.J., Kubota C., and Tronstad R.. "A dynamic modelling framework for human hand gesture task recognition." arXiv preprint arXiv: 1911.03923 (2019).

Masoud, S., Mariscal, N., Huang, Y., & Zhu, M. (2021). A Sensor-based data driven framework to investigate PM 2.5 in the greater detroit area. *IEEE Sensors Journal, 21*(14), 16192–16200.

Michigan Department of Environment, Great lakes, and energy, "annual ambient air monitoring network review," 1 July 2021. [Online]. Available: Https://www.michigan.gov/documents/egle/egle-aqd-amu-2021_air_monitoring_network_review_689434_7.pdf. [Accessed 14 March 2022].

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., et al. (2020). Deep learning-based PM2. 5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of The Total Environment, 699*, Article 133561.

Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2. 5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment, 664*, 1–10.

Rahman, R. R., & Kabir, A. (2023). Spatiotemporal analysis and forecasting of air quality in the greater Dhaka region and assessment of a novel particulate matter filtration unit. *Environmental Monitoring and Assessment, 195*(7). https://doi.org/10.1007/s10661-023-11370-y

Ryu, S., Kim, H., Kim, S. G., Jin, K., Cho, J., & Park, J. (2022). Probabilistic deep learning model as a tool for supporting the fast simulation of a thermal–hydraulic code. *Expert Systems with Applications, 200*, Article 116966.

Sadhu, A., Soni, R., & Mishra, M. (2020). Pattern-based comparative analysis of techniques for missing value imputation. In *Proceedings of the IEEE 5th international conference on computing communication and automation (ICCCA)* (pp. 513–518). IEEE.

Samal, K. K. R., Babu, K. S., & Das, S. K. (2021). Multi-directional temporal convolutional artificial neural network for PM2. 5 forecasting with missing values: A deep learning approach. *Urban Climate, 36*, Article 100800.

Schulz, AJ., Mentz, G. B., Sampson, N., Ward, M., Timothy Dvonch, J., Majo, R. D., Barbara, A., Israel, A. G. R., & Donele, W. (2018). Independent and joint contributions of fine particulate matter exposure and population vulnerability to mortality in the Detroit metropolitan area. *International Journal of Environmental Research and Public Health, 15*(no. 6), 1209.

Shi, J., Liu, S., Qu, Y., Zhang, T., Dai, W., Zhang, P., … Cao, J. (2023). Variations of the urban PM2. 5 chemical components and corresponding light extinction for three heating seasons in the Guanzhong Plain, China. *Journal of Environmental Management, 327*, Article 116821.

Shogrkhodaei, S. Z., Razavi-Termeh, S. V., & Fathnia, A. (2021). Spatio-temporal modeling of pm2. 5 risk mapping using three machine learning algorithms. *Environmental Pollution, 289*, Article 117859.

Song, J., Saathoff, H., Gao, L., Gebhardt, R., Jiang, F., Vallon, M., … Leisner, T. (2022). Variations of PM2. 5 sources in the context of meteorology and seasonality at an urban street canyon in Southwest Germany. *Atmospheric Environment, 282*, Article 119147.

Sun, Y., Zhuang, G., Tang, A., Wang, Y., & An, Z. (2006). Chemical characteristics of PM2. 5 and PM10 in haze– fog episodes in Beijing. *Environmental science & technology, 40*(10), 3148–3155.

Tan, S., Wang, Y., Yuan, Q., Zheng, L., Li, T., Shen, H., & Zhang, L. (2022). Reconstructing global PM2. 5 monitoring dataset from OpenAQ using a two-step spatio-temporal model based on SES-IDW and LSTM. *Environmental Research Letters, 17*(3), Article 034014.

Thongthammachart, T., Araki, S., Shimadera, H., Eto, S., Matsuo, T., & Kondo, A. (2021). An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal PM2. 5 predictions in the Kansai region of Japan. *Atmospheric Environment, 262*, Article 118620.

"United States Environmental Protection Agency," [Online]. Available: Https://aqs.epa.gov/aqsweb/airdata/download_files.html. [Accessed 14 3 2022].

Wang, X., Yuan, J., & Wang, B. (2021). Prediction and analysis of PM2. 5 in Fuling District of Chongqing by artificial neural network. *Neural Computing and Applications, 33*(no. 2), 517–524.

"Weather Underground," [Online]. Available: Https://www.wunderground.com/history. [Accessed 14 3 2022].

Wilcox, R. "Trimming and winsorization." Encyclopedia of biostatistics 8 (2005).

Xing, Y. F., Xu, Y. H., Shi, M. H., & Lian, Y. X. (2016). The impact of PM2. 5 on the human respiratory system. *Journal of thoracic disease, 8*(1), E69.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. In *, 33. Proceedings of the AAAI conference on artificial intelligence* (pp. 7370–7377).

Zhan, Yu, Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Lizhong, Z., & Minghua, Z. (2017). Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric Environment, 155*, 129–139.

Zhang, L., Na, J., Zhu, J., Shi, Z., Zou, C., & Yang, L. (2021). Spatiotemporal causal convolutional network for forecasting hourly PM2. 5 concentrations in Beijing, China. *Computers & Geosciences, 155*, Article 104869.

Zhao, X., Zhou, W., Han, L., & Locke, D. (2019). Spatiotemporal variation in PM2. 5 concentrations and their relationship with socioeconomic factors in China's major cities. *Environment International, 133*, Article 105145.

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., Xuanlin, Q., et al. (2018). PM2. 5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. *Atmospheric Environment, 183*, 20–32.

Zhu, J., Deng, F., Zhao, J., & Zheng, H. (2021). Attention-based parallel networks (APNet) for PM2. 5 spatiotemporal prediction. *Science of The Total Environment, 769*, Article 145082.

Zilly, J. G., Srivastava, R. K., Koutnık, J., & Schmidhuber, J. (2017). Recurrent highway networks. *International conference on machine learning* (pp. 4189–4198). PMLR.