SUSTAINABLE MACHINE INTELLIGENCE JOURNAL

# Integrating Machine Intelligence to Estimate PM2.5 Concentration for Sustainable Urban Air Quality Management

**Ahmed Metwaly[1,*], Ahmed Sleem[2], and Ibrahim Elhenawy[3]**

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt; a.met-waly23@fci.zu.edu.eg;
[2] Ministry of communication and information technology, Egypt; Asleem@mcit.gov.eg
[3] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt; ielhenawy@zu.edu.eg;
* Correspondence: a.metwaly23@fci.zu.edu.eg.

**Abstract: Air quality degradation, particularly the proliferation of fine particulate matter (PM2.5), poses a critical threat to environmental sustainability and public health. This paper introduces a comprehensive machine learning (ML) framework designed to predict PM2.5 concentrations, addressing the complexities inherent in heterogeneous urban environments. Drawing from a review of existing literature encompassing diverse ML methodologies applied to PM2.5 prediction, this study proposes an innovative approach amalgamating various data sources, including meteorological, geographical, and anthropogenic factors. Leveraging ensemble learning techniques and novel algorithmic models, our framework aims to surpass limitations encountered in current predictive models, enabling accurate and localized PM2.5 predictions. The significance of this research lies in its potential to offer a robust tool for environmental policymakers and urban planners, facilitating informed decisions towards mitigating PM2.5 pollution and fostering sustainable environments. Through evaluation of multiple ML algorithms, this paper contributes a novel predictive model crucial for enhancing air quality management.**

| Event | Date |
|---|---|
| Received | 01-06-2023 |
| Revised | 25-08-2023 |
| Accepted | 09-09-2023 |
| Published | 21-09-2023 |

**Keywords: PM2.5, Air quality, Machine intelligence, Urban environment, Sustainable development, Environmental monitoring, Artificial intelligence, Smart cities, Sensor networks, Predictive modeling, IoT (Internet of Things), GIS (Geographic Information System), Machine learning (ML).**

## 1. Introduction

The decline in air quality, especially the hike in PM2.5 levels, is endangering the future of urban areas on every continent. Urbanization and industrialization that have happened rapidly over the past years demands an effective approach to air management [1]. PM2.5 are tiny particles, 2.5 microns or smaller in diameter which are known to cause severe health effects and environmental impact [2]. Traditional air quality monitoring approaches give important information but may not always be dynamic enough for such urbanized settings necessitating novel methods like integrating machines thinking into pollution control as a result of our new understanding of cities [3-4]. Emerging AI technologies offer a number of prospects for addressing the issues associated with monitoring and predicting PM2.5 concentrations [5]. Machine learning algorithms have achieved considerable success in extracting meaningful patterns out of huge datasets. This potential has been demonstrated by AI-enabled solutions that can leverage various types of data including meteorological data, satellite imagery and ground-based sensor networks leading to improved accuracy and efficiency of predictions for PM2.5 concentrations [6]. Despite these advancements, there remains a need to explore the integration of AI technologies into urban air quality management comprehensively.

### 1.1. Research Gaps

Although the field has experienced significant progress due to artificial intelligence (AI) solutions, noticeable research gaps still exist. Therefore, the scalability and adaptability of existing models on different urban configurations should be critically analyzed [7]. Other challenges include data quality, sensor location and model interpretation that can affect the accuracy of PM2.5 concentration estimates. Additionally, societal-economic factors as well as environmental elements that determine air quality have to be better accommodated in AI models. In conclusion identifying these research gaps will result in more sustainable urban air quality management through robust and practical solutions [8].

### 1.2. Motivation & Contribution

This study is motivated by the urgent requirement to fill those gaps and advance machine intelligence integration in assessing PM2.5 concentrations for urban air quality management. Accurate and timely predictions can bring about wide-ranging environmental and societal benefits apart from health considerations. Enhancements in air quality not only support public health but also promote sustainable urbanism as well as resilience communities. The aim of this study is to enhance understanding of how AI can be applied in environmental science policy, leading to improved mitigation strategies against PM2.5 pollution in cities [9-11].

## 2. Literature Review

This part of the paper focuses on how machine intelligence has been linked with PM2.5 estimation and considers both the success and failure in those methodologies that have helped in laying foundations for the current study. Lv et al. [11] delved into the enhancement of numerical simulation predictions of PM2.5 and its chemical components through the application of machine learning algorithms. The study explored the effectiveness of these algorithms in refining the accuracy of simulations, shedding light on their potential for improving the understanding of PM2.5 dynamics. In addition, Qiao et al. [12] proposed a hybrid model that integrates wavelet transform and an improved deep learning algorithm for forecasting PM2.5. By incorporating wavelet transform, the model aimed to capture temporal patterns in PM2.5 data effectively. The study contributes to the exploration of novel hybrid approaches in improving the precision of PM2.5 predictions. More, Zamani Joharestani et al. [13] focused on PM2.5 prediction, employing random forest, XGBoost, and deep learning techniques. Notably, the study leveraged multisource remote sensing data, showcasing the importance of integrating diverse data streams to enhance prediction accuracy and robustness. Besides,Pak et al. [14] addressed PM2.5 prediction with a deep learning-based approach that considered spatiotemporal correlations. The study specifically focused on the case study of Beijing, China, illustrating the importance of accounting for spatial and temporal dynamics for accurate predictions in urban environments. Zhan et al. [15] developed a spatially explicit machine learning algorithm for spatiotemporal prediction of continuous daily PM2.5 concentrations across China. The study contributed insights into the geographic variations and temporal dynamics of PM2.5, emphasizing the need for location-specific models. Kumar et al. [16] presented a machine learning-based model tailored for estimating PM2.5 concentration levels in Delhi's atmosphere. The study likely considered the unique environmental and geographical characteristics of Delhi to develop a model suitable for the specific challenges faced by this urban area. Moreover, Choi and Kim [17] applied Principal Component Analysis (PCA) to deep learning forecasting models for predicting PM2.5. The incorporation of PCA aimed to enhance the interpretability of deep learning models, providing insights into the underlying factors influencing PM2.5 concentrations. Enebish et al. [18] focused on predicting ambient PM2.5 concentrations in Ulaanbaatar, Mongolia, utilizing various machine learning approaches. Given the specific environmental conditions of Ulaanbaatar, the study likely addressed region-specific challenges and contributed to the understanding of PM2.5 dynamics in the context of Mongolia. Further, Ejohwomu et al. [19] contributed to the literature by modeling and forecasting temporal PM2.5

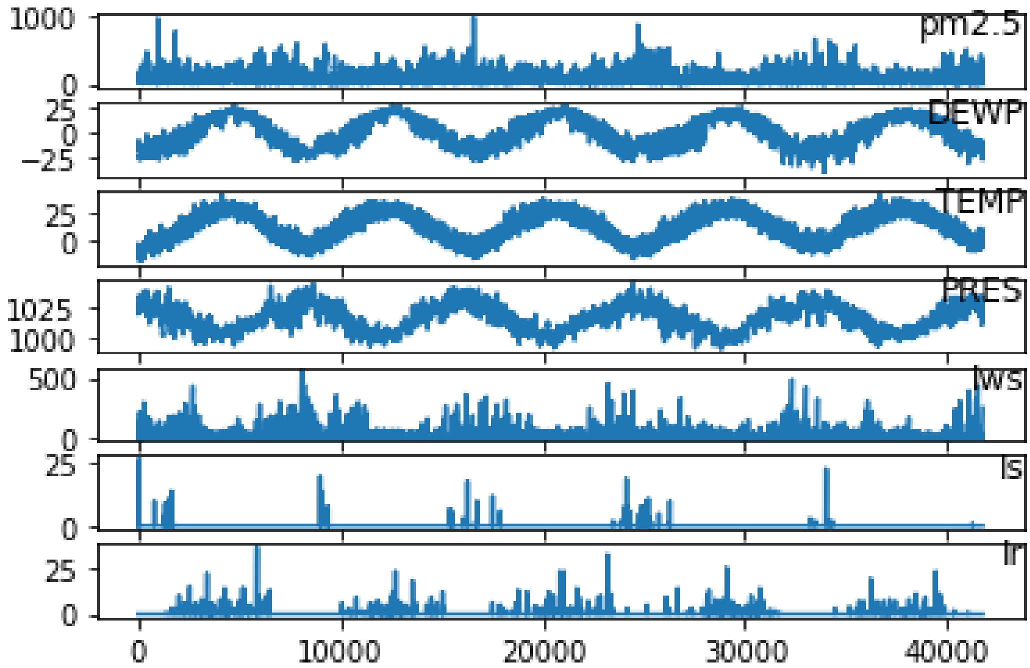


*Figure 1: Temporal Visualization of Key Atmospheric Factors Influencing PM2.5 Concentrations.*

concentrations using ensemble machine learning methods. Ensemble methods combine multiple models to improve predictive performance, and this study likely explored the advantages of such an approach for PM2.5 forecasting. Shen et al. [20] adopted the Prophet forecasting model, a machine learning approach, to predict the concentrations of various air pollutants, including PM2.5, in Seoul, South Korea. The study likely considered the unique characteristics of Seoul's air quality and demonstrated the applicability of the Prophet model to diverse pollutants in an urban setting.

## 3. Material and Method

In this section, the research methodology and the materials employed play a crucial role in ensuring the credibility and replicability of the study.

### 3.1. Material

To do this research we used a big database that had been collected from January 1st, 2010 to December 31st, 2014, which consisted of the most important atmospheric variables required in the forecasting of PM2.5 concentration. The dataset includes eight significant attributes contributing for a unique purpose to the model, namely: PM2.5 concentration (ug/m^3), dew point (°C), temperature (°C), pressure (hPa), cbwd – wind direction, wind speed (m/s), and cumulative counts of hours characterized by snowfall ("Is") and rainfall ("Ir"). Such diverse variables contribute towards understanding various atmospheric conditions that underlie fluctuations in PM2.5 levels. This dataset is comprised of data points each with a given year, month, day and hour thus allowing us to carry out temporal analysis of factors influencing PM2.5 concentrations within the specified period of time. In addition, it should be noted that missing values are represented as “NA,” and the time granularity provided by the temporal data supports our machine learning based pursuit of sustainable urban air quality management practices.

Figure 1 gives an overview through a diagram illustrating how different variables relate to one another over a given period with respect to the levels of PM2.5 in Beijing city. The visualization serves as a crucial exploratory step, providing a clear overview of the patterns, trends, and potential correlations within the data.

### 3.2. Method

In this study, we leverage the power of Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), renowned for its efficacy in capturing intricate temporal dependencies within sequential data. LSTMs are particularly well-suited for time-series forecasting tasks due to their ability to retain and selectively utilize information over extended periods. This becomes crucial in our context, as we aim to predict PM2.5 concentrations, which exhibit intricate patterns and dependencies influenced by various atmospheric factors.

To provide a solid theoretical foundation, we delve into the core components of LSTM networks. Memory cells, input gates, forget gates, and output gates collectively empower LSTMs to store and process information over time. We emphasize how LSTMs address the vanishing gradient problem, a common challenge in training deep neural networks, by enabling the effective capture of both short and long-term dependencies. Understanding these components is pivotal for grasping how the LSTM network transforms raw input data into meaningful predictions.

Moving to the practical application, we outline our approach to applying LSTM for PM2.5 concentration estimation. Our dataset is meticulously structured to facilitate time-series forecasting, ensuring that temporal relationships and dependencies are faithfully represented. Input features encompass critical atmospheric parameters such as PM2.5 concentrations, dew point, temperature, pressure, wind direction, wind speed, and cumulative hours of snow and rain. By incorporating these features, we enable the LSTM network to discern intricate patterns within the data, ultimately enhancing the accuracy of our PM2.5 predictions.

```
# Import necessary libraries
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dropout, Dense

# Design the LSTM network
model = Sequential()
# Add the first LSTM layer with 100 units, returning sequences, and specifying the
input shape
model.add(LSTM(100, return_sequences=True, input_shape=(train_X.shape[1],
train_X.shape[2])))
# Add dropout layer to prevent overfitting
model.add(Dropout(0.3))
# Add a second LSTM layer with 50 units and returning sequences
model.add(LSTM(units=50, return_sequences=True))
# Add dropout layer
model.add(Dropout(0.2))
# Add a third LSTM layer with 50 units and returning sequences
model.add(LSTM(units=50, return_sequences=True))
# Add dropout layer
model.add(Dropout(0.2))
```

```
# Add a fourth LSTM layer with 50 units
model.add(LSTM(units=50))
# Add dropout layer
model.add(Dropout(0.2))
# Add a Dense (fully connected) layer with 1 unit and linear activation function
model.add(Dense(1, activation='linear'))
# Compile the model using mean squared error as the loss function and the Adam op-
timizer
model.compile(loss='mse', optimizer='adam')
```

## 4. Empirical Findings and Analysis

This section serves as the cornerstone for understanding the outcomes of our research efforts, presenting a detailed analysis of the data generated through the application of machine learning algorithms. The empirical findings embrace the effectiveness and efficiency of the proposed methodology in predicting PM2.5 levels, revealing patterns, relationships and nuances in the data. In Figure 2, we have presented a correlation map that sum up all these related parameters affecting PM2.5 concentrations on a visual basis. This is a picture meaning that it can be used to quickly see how different atmospheric factors relate to each other completely. Through use of color codes, the correlation map offers important aspects about how closely associated things are which are significant when looking for possible drivers or influencers of PM2.5 conditions. This visualization is then an essential aid for our empirical analysis as it aids in understanding connections within the dataset as well as being used as a basis for future discussions regarding sustainable urban air quality management in relation to these variables' influences on it.

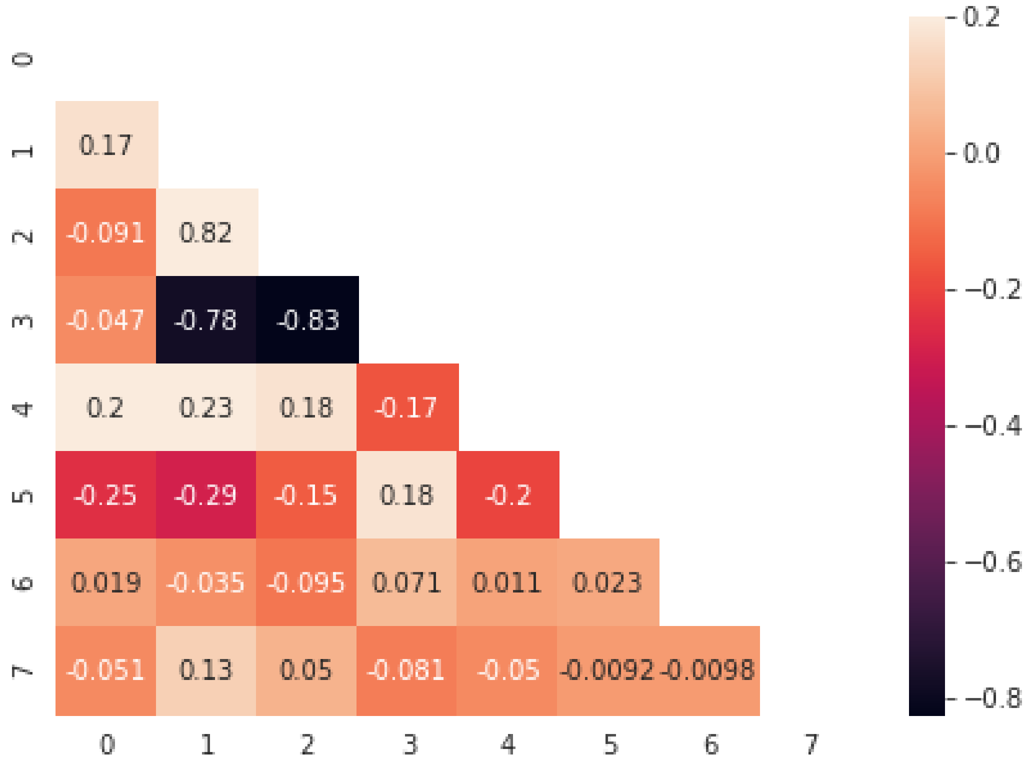


*Figure 2: Correlation Map of Key Atmospheric Parameters and PM2.5 Concentrations.*

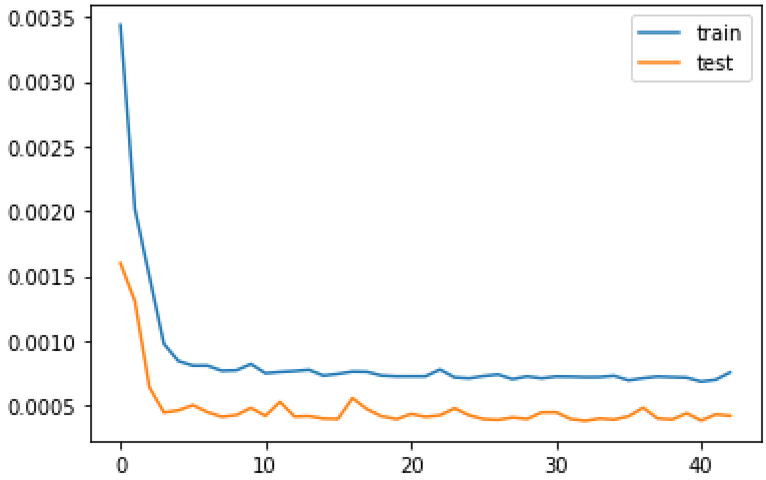


*Figure 3: Learning Curves for LSTM Models.*

In Figure 3, we present the learning curves for Long Short-Term Memory (LSTM) models, offering a visual representation of the model's training and validation performance over successive epochs. The curves depict the evolution of training and validation errors, providing a dynamic insight into the convergence and generalization capabilities of the LSTM algorithm. This graphical representation is crucial for assessing the model's training dynamics and understanding its ability to learn complex patterns within the dataset. The learning curves serve as a key diagnostic tool, aiding in the optimization of model hyperparameters and contributing to the overall evaluation of the LSTM's efficacy in estimating PM2.5 concentrations.

In Figure 4, we present the prediction curve versus the actual curve, offering a side-by-side comparison of the LSTM model's predictions against the observed PM2.5 concentrations. This visual representation allows for a direct assessment of the model's accuracy in capturing the temporal variations and trends present in the actual data. The alignment between the prediction and actual curves serves as a critical benchmark for evaluating the

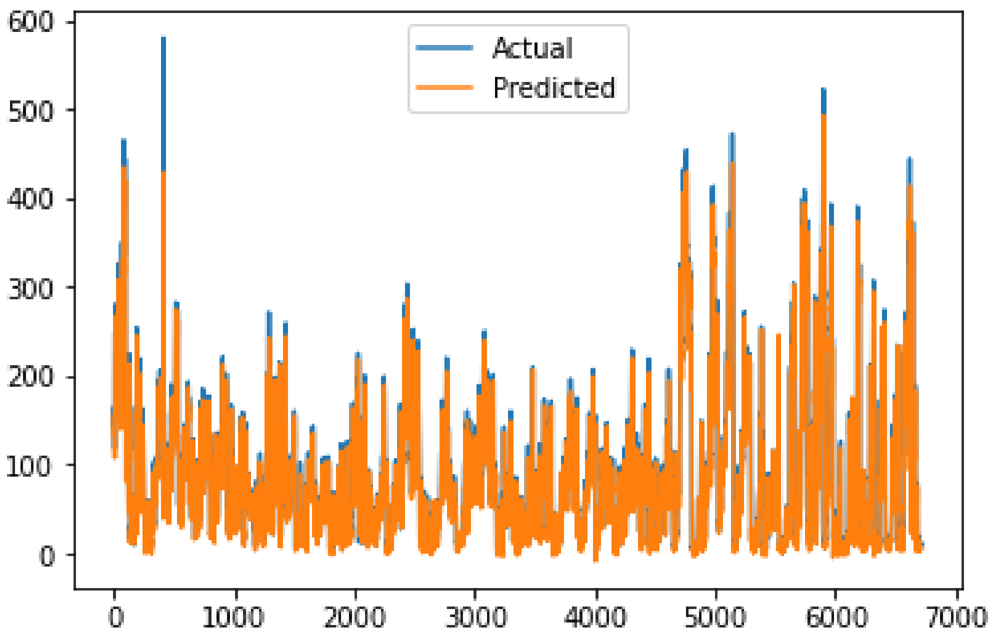


*Figure 4: Prediction Curve vs. Actual Curve for LSTM Model*

model's performance, highlighting its capacity to generate reliable estimates. Figure 4 plays a pivotal role in gauging the predictive power of the LSTM algorithm and contributes to the broader discussion on the effectiveness of machine intelligence in estimating PM2.5 concentrations.

## 5. Conclusions and Future Directions

Our study underscores the pivotal role of machine intelligence, specifically the application of Long Short-Term Memory (LSTM) networks, in advancing the estimation of PM2.5 concentrations for sustainable urban air quality management. Through a meticulous exploration of atmospheric parameters and their temporal dynamics, our findings reveal the efficacy of LSTM models in capturing intricate patterns and dependencies within the dataset. Exponentiation with public dataset, spanning from 2010 to 2014, facilitated a nuanced exploration of urban air quality dynamics. The outcomes not only showcase the practical viability of LSTM networks but also emphasize the importance of considering various meteorological factors for accurate PM2.5 predictions. These findings hold significant implications for environmental monitoring and policy-making, highlighting the potential of advanced machine learning techniques to enhance our capabilities in managing and mitigating air pollution in urban areas.

Towards the future, the study offers several future directions for integrating machine intelligence for PM2.5 concentration estimation and sustainable urban air quality management. For instance, further research could explore how to incorporate more data sources (i.e real-time traffic patterns, land-use data and demographic information) in order to enhance predictive accuracy of models. Furthermore, beyond LSTMs, it may be necessary to consider integrating superior machine learning techniques aimed towards comparing their effectiveness in dealing with inherent complexities of air quality dynamics. Moreover, creating explainable AI models would help make predictions more transparent and therefore get greater acceptance from stakeholders including decision-makers. Besides that, expanding the temporal range of the dataset and incorporating more recent data will help assess these models in terms of changes in urban environments over time.

## Funding

This research was conducted without external funding support.

## Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

## Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

## References

[1]. Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., ... & Sachdeva, S. (2019). Evaluation of different machine learning approaches to forecasting PM2. 5 mass concentrations. *Aerosol and Air Quality Research*, *19*(6), 1400-1410.

[2]. Peng, J., Han, H., Yi, Y., Huang, H., & Xie, L. (2022). Machine learning and deep learning modeling and simulation for predicting PM2. 5 concentrations. *Chemosphere*, *308*, 136353.

[3]. Xiao, Q., Chang, H. H., Geng, G., & Liu, Y. (2018). An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data. *Environmental science & technology*, *52*(22), 13260-13269.

[4]. Xiao, F., Yang, M., Fan, H., Fan, G., & Al-Qaness, M. A. (2020). An improved deep learning model for predicting daily PM2. 5 concentration. *Scientific reports*, *10*(1), 20988.

[5]. Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost machine learning method in PM2. 5 prediction: A case study of Shanghai. *Aerosol and Air Quality Research*, *20*(1), 128-138.

[6]. Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057-2066.

[7]. Chen, G., Li, S., Knibbs, L. D., Hamm, N. A., Cao, W., Li, T., ... & Guo, Y. (2018). A machine learning method to estimate PM2. 5 concentrations across China with remote sensing, meteorological and land use information. *Science of the Total Environment*, *636*, 52-60.
[8]. Masood, A., & Ahmad, K. (2020). A model for particulate matter (PM2. 5) prediction for Delhi based on machine learning approaches. *Procedia Computer Science*, *167*, 2101-2110.
[9]. Chang, F. J., Chang, L. C., Kang, C. C., Wang, Y. S., & Huang, A. (2020). Explore spatio-temporal PM2. 5 features in northern Taiwan using machine learning techniques. *Science of the Total Environment*, *736*, 139656.
[10]. Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., ... & Schwartz, J. (2020). Predicting fine particulate matter (PM2. 5) in the greater london area: An ensemble approach using machine learning methods. *Remote Sensing*, *12*(6), 914.
[11]. Lv, L., Wei, P., Li, J., & Hu, J. (2021). Application of machine learning algorithms to improve numerical simulation prediction of PM2. 5 and chemical components. *Atmospheric Pollution Research*, *12*(11), 101211.
[12]. Qiao, W., Tian, W., Tian, Y., Yang, Q., Wang, Y., & Zhang, J. (2019). The forecasting of PM2. 5 using a hybrid model based on wavelet transform and an improved deep learning algorithm. *IEEE Access*, *7*, 142814-142825.
[13]. Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), 373.
[14]. Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2. 5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment*, *699*, 133561.
[15]. Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... & Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric environment*, *155*, 129-139.
[16]. Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. *Heliyon*, *6*(11).
[17]. Choi, S. W., & Kim, B. H. (2021). Applying PCA to deep learning forecasting models for predicting PM2. 5. *Sustainability*, *13*(7), 3726.
[18]. Enebish, T., Chau, K., Jadamba, B., & Franklin, M. (2021). Predicting ambient PM2. 5 concentrations in Ulaanbaatar, Mongolia with machine learning approaches. *Journal of exposure science & environmental epidemiology*, *31*(4), 699-708.
[19]. Ejohwomu, O. A., Shamsideen Oshodi, O., Oladokun, M., Bukoye, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and forecasting temporal PM2. 5 concentration using ensemble machine learning methods. *Buildings*, *12*(1), 46.
[20]. Shen, J., Valagolam, D., & McCalla, S. (2020). Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2. 5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. *PeerJ*, *8*, e9961.