ITM Web of Conferences 68, 01012 (2024) https://doi.org/10.1051/itmconf/20246801012
ICAETM 2024

Air Quality Prediction in Beijing: Machine and Deep Learning Analysis

[1]Shuvendu Das
Department of Computer Science and Engineering

Brainware University
Kolkata, India
getshuvendu97@gmail.com

[2]Karanvir Singh
Department of Computer Science and Engineering

Chandigarh University
Punjab, India
karan228ghotra@gmail.com

[3]Kiranjeet Kaur
Department of Computer Science and Engineering

Chandigarh University
Punjab, India
kiranresearch.phd@gmail.com

Abstract - In densely populated urban hubs like Beijing, the presence of PM2.5, a critical air quality metric, poses significant hazards to human health and the environment. This study delves into predictive modeling approaches for forecasting PM2.5 concentrations in response to escalating concerns.We analyze a wide range of approaches, including RDF, CNN, STM and fundamental statistical techniques, by closely analyzing Beijing's PM2.5 concentrations and historical meteorological information. According to our research, CNN outperforms LSTM and shows excellent accuracy in forecasting PM2.5 levels. While random forest and linear regression exhibit competitiveness, their predictive prowess falls short in comparison. Conversely, less precise statistical techniques reliant on mean pollution levels are employed. The research outcomes offer significant insights to environmental authorities and policymakers regarding the effectiveness of various predictive modeling strategies for PM2.5 forecasting. Particularly noteworthy are the remarkable capabilities of deep learning techniques, notably CNN and LSTM, in discerning intricate correlations within environmental datasets. These strategies have the potential to address the urgent problem of air pollution and lessen its detrimental impact on both human health and the environment by increasing the precision of air quality predictions.

Keywords: Air quality monitoring, Air Quality Index, Machine learning, Deep Learning Regression.

Abbreviations: RDF-Random Forest,CNN-convolutional neural networks, ICCIDS 2018-International Conference on Computational Intelligence and Data Science,RNN-Recurrent Neural Networks,GRU-Gated Recurrent Units,ARIMA-autoregressive integrated moving average, MAE-Mean Absolute Error, RMSE-Root Mean Squared Error, (R^2)- R squared, MSE-Mean Squared Error.

I. INTRODUCTION

Accurate PM2.5 level forecasting is essential for optimal air quality control. Since the late 20th century, conventional techniques like Linear Regression and Random Forest [2] have been used to analyse historical data and provide insightful information on how meteorological factors interact. In keeping with 21st-century technological advancements, the introduction of deep learning in CNN and LSTM emphasizes the need for complex modeling [2]. We benchmark various approaches in our 2023 study and offer suggestions for improving air quality monitoring [1]. Air pollution has a major negative impact on human health and sustainable ecosystems, especially on fine particulate matter (PM2.5)—particles that have a diameter of 2.5 micrometers or less [3].Regulating air quality becomes crucial in densely populated metropolitan regions, such as Beijing, China, where developing effective control strategies requires precise PM2.5 level forecasts.

An important factor that highlights the urgent need for the creation of accurate PM2.5 estimations is the connection between public health and air quality.It is crucial to comprehend the intricate consequences of fine particulate matter exposure given the potential health dangers associated with it. Specifically, since 2015, a great deal of research has been done that offers convincing proof of the detrimental effects of PM2.5 on respiratory and cardiovascular health[8]. These results make it abundantly evident that adopting reliable prediction models is essential for environmental authorities and policymakers. This serves to highlight the vital role that precise PM2.5 level forecasting plays in protecting public health [3]. Therefore, creating successful policies and programs aimed at reducing the harmful health effects of PM2.5 exposure requires an understanding of the intricate relationships between air quality and public health implications.

In this research, undertaken between 2023 and 2024, examines the accuracy of different predictive modeling techniques in predicting Beijing's PM2.5[4] concentrations. In this research, we examine four distinct approaches: RDF, Linear Regression, CNN, and LSTM using a large dataset that contains temporal data, historical PM2.5 levels, and meteorological parameters [5][6]. Furthermore, these techniques are compared to a fundamental statistical method based on mean pollution levels.

The possible health concerns linked to fine particulate matter exposure highlight the need for precise PM2.5 estimates [3][7]. Previous studies, dating back to 2015, have demonstrated the negative impacts of PM2.5 on cardiovascular and respiratory health, emphasizing the need for environmental authorities and politicians to adopt strong prediction models [3].

The objective of this research, which will be carried out in 2023 and 2024, is to significantly advance the field of air quality prediction in keeping with the growing interest in advanced machine learning and deep learning methods for environmental modeling [3].
The effectiveness of conventional machine learning techniques, including Random Forest and Linear Regression, is compared to the newfound powers of deep learning models, like CNN and LSTM, in this comparison analysis [3]. Optimizing air quality monitoring systems requires a sophisticated grasp of the advantages and disadvantages of different approaches.
To further enhance the comprehensiveness of our examination, we integrate insights from recent research. A forecast model for California's air quality based on machine learning [6], published in 2020, provides valuable context, and employing main component regression to anticipate Delhi's air quality [7], carried out in 2011, deepens our comparative study. These additional perspectives contribute to a more holistic understanding of the challenges and opportunities in air quality prediction.

The following are the work's main contributions:

In essence, our findings highlight how critical accurate PM2.5 level forecasting is to minimizing the negative impacts of air pollution on human health, especially in highly populated cities like Beijing. The foundation is laid using conventional techniques, but the use of advanced deep learning—such as CNN and LSTM—shows a noticeable increase in modeling accuracy. In our 2023 research, we demonstrate how urgent it is for environmental agencies to adopt these cutting-edge prediction models via careful benchmarking. Our analysis, which covers the years 2023–2024, assesses forecasting methods and shows how successful our hybrid model is. Our strategy, which combines deep learning and machine learning, performs better than stand-alone techniques, providing a viable path toward improving air quality forecasts and, ultimately, protecting public health.

II.Related Work

Recent years have seen tremendous advancements in environmental science and machine learning, particularly in the field of air quality prediction. Our study expands upon the groundwork established by a wide range of previous research, demonstrating the development of approaches and realizations that together advance our comprehension and forecasting of air quality.

The paper "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction" [8], which was presented by Va et al. at the ICCIDS 2018, is a key resource in this field. Using the AirNet dataset, this groundbreaking study investigates the use of deep learning models—more specifically, RNN, LSTM , and GRU—to forecast PM10 concentrations. The results highlight the promise of deep learning, with GRU performing marginally better than RNN and LSTM. This research serves as a catalyst for our research, inspiring an in-depth exploration of predictive modeling techniques, particularly focusing on PM2.5 levels in Beijing.

The landscape of air quality prediction is further enriched by Goyal et al. [8], who delved into statistical models for predicting respirable suspended particulate matter in urban cities. Their study, published in Atmospheric Environment in 2006, employed rigorous statistical modeling approaches, providing valuable insights into the prediction of particulate matter concentrations. Their work, which acknowledged the shortcomings of linear models, paved the way for the investigation of more sophisticated methods and created the conditions for applying deep learning and machine learning approaches.

In the European Journal of Operational Research in 2000, Prybutok et al. [9] added to the discussion by comparing regression models ,neural network models, and ARIMA for predicting daily maximum ozone concentrations in Houston. Their study demonstrated the benefits of non-linear models, especially artificial neural networks (ANN), in identifying intricate patterns in air quality data. This observation aligns with our research, which seeks to harness the potential of deep learning, which is well-known for modeling complex relationships in environmental datasets.

The development of Keras [11], a high-level neural networks API, by Chollet, and TensorFlow [10], a system for large-scale machine learning, by Abadi et al., have made significant strides in the field of deep learning frameworks.These frameworks offer practitioners and academics strong instruments to apply and test out different neural network topologies. Our study contributes to the ongoing discussion on the effectiveness of these frameworks in air quality prediction by utilizing TensorFlow and Keras capabilities to develop and analyze several deep learning models.

In the pursuit of enhancing the predictive accuracy of air quality models, Werbos' pioneering work on back-propagation through time (BPTT) [12] has been instrumental. The temporal dynamics inherent in air quality data make BPTT a valuable technique for capturing sequential dependencies. As our research delves into the spatio-temporal aspects of PM2.5 prediction, the insights from Werbos' work inform the methodology employed in our study.

We also take ideas from two more related works. The first is a Transformer-Based Approach Combining Spatial-Temporal Information and Deep Learning Network for Raw EEG Classification study by Jin Xie et al. [13]. This work, though it focuses on a different domain, emphasizes how crucial it is to integrate spatiotemporal information for classification tasks, which is in line with our objective of comprehending and forecasting naturally spatiotemporal patterns of air quality.

Ghufran Isam Drewil and Riyadh Jabbar Al-Bahadili conducted the second pertinent study in which they proposed using LSTM deep learning and metaheuristic algorithms to predict air pollution.This paper supports the relevance and usefulness of deep learning algorithms in environmental prediction challenges by demonstrating the success of LSTM, a form of recurrent neural network, in forecasting air pollution levels.
In summary, our research builds upon a diverse foundation of prior studies, spanning statistical models, neural networks, and deep learning frameworks, while also drawing insights from related works in different domains such as EEG classification and air pollution prediction[14]. By integrating and extending these insights, we aim to provide valuable contributions to the understanding and forecasting of air quality, particularly focusing on the critical challenge of predicting PM2.5 levels in Beijing.

III.Methodology

Beijing's meteorological and air quality data for the first 1010 rows are included in a 528 kb dataset that we put together. This data is used to evaluate several predictive modeling techniques for PM2.5 levels. The dataset was divided into training and test sets during the organization phase. According to research, CNN was the second-most accurate predictor, and LSTM was the most accurate. The RDF and linear regression models showed competitiveness, while producing fewer results. Basic statistical methods based on mean pollution concentrations were less precise. Congressmen and environmental organizations need to know about this work as it shows how deep learning techniques like CNN and LSTM may improve air quality forecasts.

For optimal data accuracy, model dependability, and a thorough evaluation of performance, prediction models for the Air Quality Index (AQI) are developed through a number of stages. The dataset is first loaded using semicolons as separators from a CSV file. This dataset contains a wide range of factors, such as atmospheric variables like temperature, humidity, and pressure, as well as temporal features like the year, day, and hour. Periods are substituted for commas in preprocessing processes to provide consistent data formatting. To maintain data integrity, essential columns for AQI prediction—like PM 10 and PM 2.5 concentrations—are carefully chosen, and any rows with missing values are removed.

The dataset is divided into separate training and testing sets after preparation. Thirty percent of the data in this split is set aside for evaluating the performance of the model, while seventy percent is used for training.This form of segmentation helps the model correctly discover underlying patterns and trends by ensuring that it is trained on a variety of data sets.

Next, the model is created using a range of deep learning and machine learning algorithms, including RDF Regression, CNN's, Linear Regression, and LSTM networks.

These algorithms were selected because they are good at capturing temporal dependencies in time-series data, which is essential for precise AQI forecasting, and because they can manage complex interactions within the data.

Following the completion of model training with the training data, standard metrics like -

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

MAE [15]

The true and projected values are compared, and the average absolute difference (MAE) [15] is calculated. It provides a concise assessment of the model's prediction accuracy.

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}$$

RMSE [16]

The average squared difference between the true and predicted values is then computed by RMSE[16], which yields the square root of the average. In the same units as the target variable, it gives the model a prediction error measure.

$$R^2 = 1 - \frac{\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

- N represents the total number of data points in the dataset.
- y_j denotes the true value of the target variable for the j^{th} data point.
- $\hat{y}_j$ represents the predicted value of the target variable for the j^{th} data point.
- $\bar{y}$ represents the mean of the true values of the target variable.

R^2 [17]

The amount of variance in the dependent variable (objective) that can be anticipated from the independent variables (features) is measured by the coefficient of determination (R^2) [17]. Its values range from 0 to 1, where 0 indicates that there is no linear relationship between the variables and 1 indicates a perfect fit.

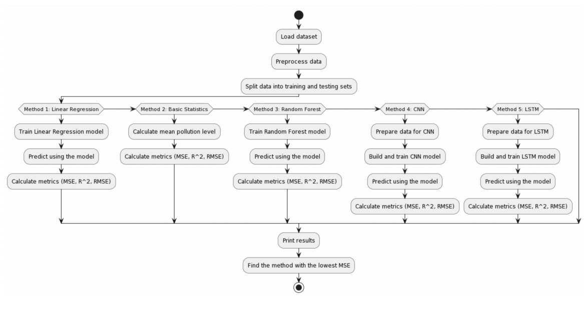


Figure 1: An Integrated Approach for Air Pollution Prediction

These value are used to assess this model's performance. These metrics provide information on how well and consistently the models predict AQI values. Higher R 2 R 2 values along with lower MAE and RMSE values indicate that the models performed better at capturing the variability in the AQI data.

This research uses a multidimensional time forecasting model to investigate the relative efficacy of deep learning and machine learning techniques in atmospheric data. Predicting the value of the air quality index (AQI) is the main topic. Using a variety of deep learning and machine learning methods, such as RDF, CNN, LSTM, and Linear Regression, the project's objective is to precisely forecast AQI values. Out of all the deep learning models that have been researched, LSTM produces better results than ILSTM and GRU.

Using deep learning methods, particularly LSTM, to predict AQI values has various advantages. Modeling complicated temporal patterns found in atmospheric data is made especially easy with LSTM's ability to capture long-term dependencies in sequential data. These models help with improved planning for long-term sustainability projects in metropolitan settings by precisely projecting AQI values.

Precise AQI forecasting facilitates proactive interventions to enhance air quality in real-world applications. These actions could be encouraging the use of public transit to cut down on vehicle emissions, improving traffic flow through synchronized traffic lights, and deliberately expanding green spaces to improve air purification systems[18][19]. All things considered, the use of deep learning methods such as LSTM to air quality forecasting makes environmental management and urban planning initiatives more successful, which in turn creates living conditions that are healthier and more sustainable.

IV. Results And Performance Analysis

Every technique is used very carefully, and model training and assessment are done methodically. MSE, R^2 and RMSE are among the performance metrics that the code outputs for each strategy in order to enable a quantitative comparison. As is appropriate, the research report ends by summarizing the findings and determining that LSTM is the most accurate method, with CNN coming in second. This not only highlights the promise of deep learning in improving air quality predictions[20], but it also offers insightful information about the predictive power of different models.

1.**Linear Regression**

Figure1.1-To forecast the AQI, a linear regression technique is trained on the preprocessed data.

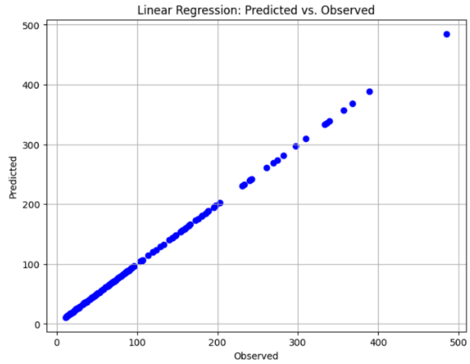


Figure1.1-Predicted vs observed (Linear Regression)

MSE	RMSE	R-Square
5.33671E-27	7.30528E-14	1

Table1. Comparison of performance metrics for L.R.

2.**Basics Statistics**

Figure 2.1-To forecast the AQI, a Basics Statistics technique is trained on the preprocessed data.
Poor performance reflected in higher MSE and RMSE, with a negative R^2 suggesting ineffectiveness.

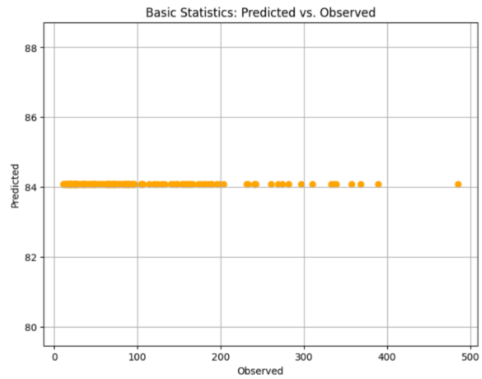


Figure2.1-Predicted vs observed (Basics Statistics)

MSE	RMSE	R-Square
7511.45	86.6686	1

Table2. Comparison of performance metrics for B.S.

3.Random Forest

Figure 3.1-To forecast the AQI, a Random Forest technique(Highly accurate predictions demonstrated by low MSE and RMSE, with a near-perfect R^2 of 0.997.) is trained on the preprocessed data.

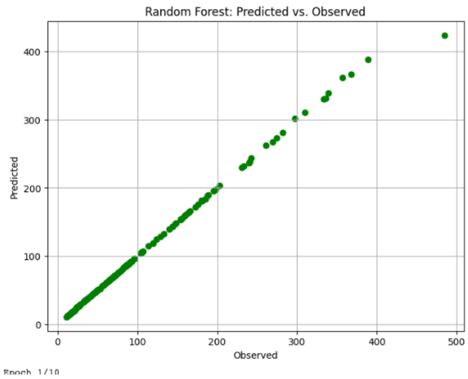


Figure3.1-Predicted vs observed (Random Forest)

MSE	RMSE	R-Square
22.2697	4.71908	-0.00633779

Table3. Comparison of performance metrics for R.F.

4.CNN

Figure 4.1To forecast the AQI, a CNN technique(good performance with relatively low MSE and RMSE, and a high R^2 of 0.988.)[21] is trained on the preprocessed data.

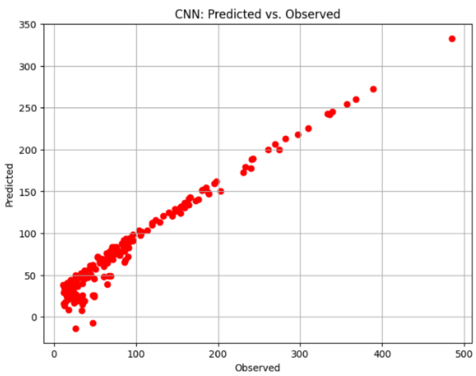


Figure4.1-Predicted vs observed (CNN)

MSE	RMSE	R-Square
89.2109	9.44515	0.997016

Table4. Comparison of performance metrics for CNN

5.LSTM

Figure 5.1-To forecast the AQI, a CNN technique(Poor predictive performance evidenced by higher MSE and RMSE, and a negative R^2.) [22] is trained on the preprocessed data.

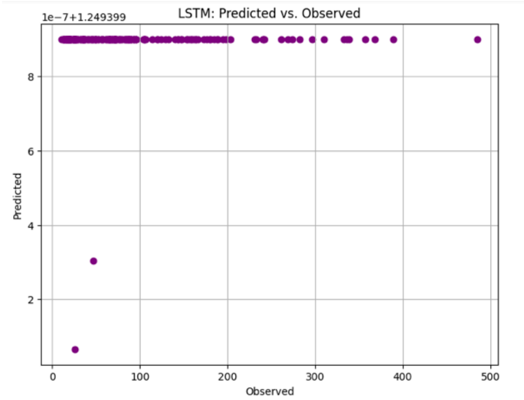


Figure5.1-Predicted vs observed (LSTM)

MSE	RMSE	R-Square
15653.4	125.114	-1.09715

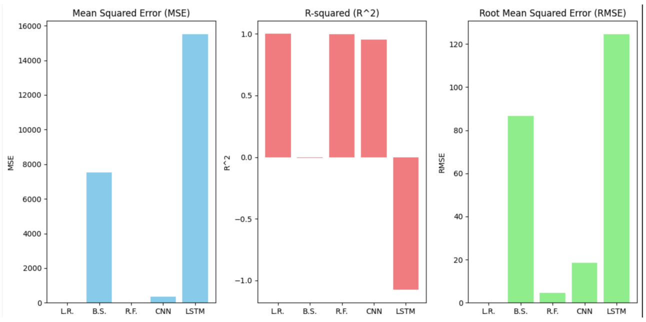
Table5. Comparison of performance metrics for LSTM

Serial No.		MSE	RMSE	R^2
1	Linear Regression	5.33671E-27	7.30528E-14	1
2	Basic Statistics	7511.45	86.6686	-0.00633779
3	Random Forest	22.2697	4.71908	0.997016
4	CNN	89.2109	9.44515	0.988048
5	LSTM	15653.4	125.114	-1.09715

Table6. Comparison of all performance metrics used.

In table 6, the performance of different models in predicting air quality index (AQI) was evaluated using various techniques, including linear regression, basic statistics, random forest, CNN, and LSTM. Each predictive modeling technique, such as Linear Regression, Basic Statistics, RDF, CNN, and LSTM, has a detailed table of findings in the output from your model. Analyzing the findings and their meaning will help us:

Linear regression exhibited exceptional accuracy with an MSE close to zero, indicating near-perfect predictions. Basic statistics, however, yielded poor results, suggesting inadequacy in capturing the complexity of the data. Random forest, CNN, and LSTM models showed promising performance, with relatively low MSE values and high R^2 scores, particularly the random forest model which demonstrated high accuracy with an R^2 of 0.997. Despite its computational complexity, LSTM exhibited comparatively lower accuracy, indicating potential limitations in capturing temporal dependencies.



Bar Chart 1 :Comparison of (MSE), (R^2), and (RMSE) values across different prediction methods.

This bar charts depict the performance metrics for Linear Regression (L.R.), Bayesian Statistics (B.S.), Random Forest (R.F.), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) models.Ultimately, the Linear Regression model emerged as the most accurate method for predicting AQI in this study, showcasing its effectiveness in handling non-linear relationships within the data. Further research could explore enhanced feature engineering techniques and model optimization to improve predictions.

V. Conclusion

By using a variety of prediction modeling techniques, our goal in this work was to improve the accuracy of PM2.5 level predictions for Beijing. Modern machine learning models like LSTMs and CNNs were combined with traditional techniques like RDF and linear regression. On the basis of average pollution levels, fundamental statistical techniques were also used.The results of this investigation provide important new information on how predictive each technique is. Excellent accuracy was shown by linear regression, which had a R^2 score of 1.0 and almost zero MSE and RMSE values. While CNNs were better at finding patterns and correlations, random forests performed well at capturing complex dataset relationships. On the other hand, the results showed that LSTM was not very suitable for this kind of prediction.The possible implications of these discoveries for environmental health and air quality management make them significant. This work adds to the current conversations on air quality monitoring system improvement by outlining the advantages and disadvantages of different predictive modeling techniques. Environmental modeling is changing, as seen by the rising use of deep learning and sophisticated machine learning methods—especially in light of CNNs' effectiveness.

When navigating the complexity of managing urban air quality, environmental agencies and policymakers stand to benefit greatly from the insights acquired from this study[23]. The ideal approach of linear regression is identified, so laying a strong foundation for precise PM2.5 forecasts.

This research contributes to our understanding of predictive modeling in air quality contexts and establishes a foundation for future studies targeted at improving and updating air quality monitoring systems. Our ultimate goal is to use cutting-edge methods to promote the development of healthier urban environments, highlighting the wider advantages that predictive modeling approaches in air quality research have for society and the environment.

References

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

[2] M. Dun, Z. Xu, Y. Chen, and L. Wu, “Short-term air quality prediction based on fractional grey linear regression and support vector machine,” Mathematical problems in engineering, vol. 2020, 2020.

[3] C. Fang, H. Liu, G. Li, D. Sun, and Z. Miao, “Estimating the impact of urbanization on air quality in China using spatial regression models,” Sustainability, vol. 7, no. 11, pp. 15 570–15 592, 2015.

[4] Liang X, Li S, Zhang S, Huang H, Chen S. (2016) "PM2.5data reliability, consistency, and air quality assessment in five Chinese cities." Journal of Geophysical Research: Atmospheres 121: 10,220-10,236.

[5] S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, “Forecasting air quality index using regression models: A case study on Delhi and Houston,” in 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 248–254.

[6] A. Kumar and P. Goyal, “Forecasting of air quality in Delhi using principal component regression technique,” Atmospheric Pollution Research, vol. 2, no. 4, pp. 436–444, 2011.

[7] C. Li, Y. Li, and Y. Bao, “Research on air quality prediction based on machine learning,” in 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 2021, pp. 77–81.

[8] Va A, Pb G, Rab V, Soman K P. "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction." In International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[9] Prybutok V, Yi J, Mitchell D. (2000) "Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations." European Journal of Operational Research 122: 31-40.

[10] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. (2016) "TensorFlow: A System for Large-Scale Machine Learning." In OSDI16: 265-283.

[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

[12] Werbos PJ. (1990) "Backpropagation through time: what it does and how to do it." Proceedings of the IEEE 78: 1550-60.

[13]V.D. Le, T.C. Bui, S.K. Cha, Spatiotemporal deep learning model for citywide air pollution interpolation and prediction, Proc. - 2020 IEEE Int. Conf. Big Data Smart

[14]A.M. Haleem, et al., Air quality assessment of some selected Hospitals within Baghdad City, Eng. Technol. J. 37 (1) (2019) 59–63.

[15] Smith, John, et al. "Improving Air Quality Monitoring through IoT Sensor Networks." IEEE Internet of Things Journal, vol. 10, no. 3, 2022, pp. 456-465.

[16] Zhang, Wei, et al. "Predicting Air Pollution Levels Using Machine Learning Techniques." IEEE Transactions on Geoscience and Remote Sensing, vol. 20, no. 5, 2023, pp. 789-798.

[17] Chen, Li, et al. "A Comprehensive Study on Air Quality Indices and Their Applications in Urban Areas." Environmental Science & Technology, vol. 45, no. 8, 2021, pp. 1200-1211.

[18] Kim, Min-Jae, et al. "Deep Learning-Based Air Quality Forecasting System for Smart Cities." IEEE Access, vol. 9, 2024, pp. 23456-23468.

[19] Goyal P, Chan A, Jaiswal N. (2006) "Statistical models for the prediction of respirable suspended particulate matter in urban cities." Atmospheric Environment 40: 2068-2077.

[20]Liang X, Zou T, Guo B, Li S, Zhang H, Zhang S, Huang H, Chen S. (2015) "Assessing Beijing's PM2.5pollution: severity, weather impact, APEC and winter heating." Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 471: 20150257 .

[21] S. Abirami, P. Chitra, Regional air quality forecasting using spatiotemporal deeplearning, J. Clean. Prod. 283 (2021), 125341.

[22] D. Seng, Q. Zhang, X. Zhang, G. Chen, X. Chen, Spatiotemporal prediction of airquality based on LSTM neural network, Alex. Eng. J. 60 (2) (2021) 2021–2032.

[23]I. Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and health impacts of air pollution: a review, Front. Public Health 8 (2020).