Predicting Air Pollution Levels in Pune, India using Generative Adversarial Networks

Sneha Khedekar
DIT, Pimpri, Pune, Maharashtra, India | AISSMS College of Engineering, Pune, Maharashtra, India
snehakhedekar1@gmail.com (corresponding author)

Sunil Thakare
DIT, Pimpri, Pune, Maharashtra, India | Anantrao Pawar College of Engineering and Research, Pune, Maharashtra, India
prof_sbthakare@rediffmail.com

Received: 25 July 2024 | Revised: 12 August 2024 | Accepted: 29 August 2024

ABSTRACT

Fuel combustion, industrial and factory exhausts, and mining activities contribute to air pollution. Predicting and evaluating the quality of air is a field of study that is growing in importance. This research builds a Generative Adversarial Network (GAN) air quality prediction model. A pre-trained accurate model was applied to predict pollutant levels in air at a given location based on historical data. The prediction GAN model utilized pollutants datasets of Particulate matter ($PM_{2.5}$ and PM_{10}), Nitrogen dioxide (NO_2), Carbon monoxide (CO), and Ozone (O_3) between 2016 and 2021 in Pune, India. The Root Mean Square Error (RMSE) statistical measure was used to assess the model's performance accuracy. The close alignment between real and predicted values underscores the high precision of the GAN model in forecasting air pollutant levels.

Keywords-air quality prediction; GAN; deep learning; Pune

I. INTRODUCTION

In recent years, there has been a growing global alarm regarding air pollution levels. The elevated levels of air pollution pose a significant threat to the health of millions of individuals. India is among the countries most severely affected by air-pollution. According to [1], India hosts 22 of the 30 most polluted cities globally. Because of its widespread negative effects on population, human health, the environment, the economy, and social well-being, the Indian government has launched a number of programs to curb air pollution. Authors in [2] define air pollution as any substance or particle in the atmosphere capable of causing harm to human health or the environment. Those substances could be made up of solid particles, liquid droplets, or gases. Pollutants come in two varieties: primary and secondary. Primary pollutants are generally generated directly from various activities, such as ash from a volcanic eruption, while secondary pollutants are not. The six most critical pollutants are lead, sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), fine particulate matter ($PM_{2.5}$ and PM_{10}), and ground level ozone (O_3). The most dangerous to health are O_3, $PM_{2.5}$, and NO_2 (the main ingredient in NO_x). On average, air pollution in Indian cities was 500% higher than the WHO-recommended safe limit threshold [3]. Twenty two Indian cities were identified to be among the 30 most severely polluted cities [4]. Based on $PM_{2.5}$ concentrations, WHO classified India as the fifth most impacted country by air pollution. An estimated 1.7 million deaths in India in 2019 were attributed to air pollution alone [1, 5]. Approximately 18% of all deaths that occurred during that time can be attributed to this. The current state of air quality in India was presented by [5] with a special focus on particulate matter. In the recent decades, scientists have faced numerous obstacles in their efforts to minimize air pollution. These are still big worldwide issues, though. The National AQI (Air Quality Index) was introduced by the Indian government in collaboration with IIT Kanpur in 2015. For the governments of developing and populated countries, air control is an essential duty. Due to rising energy and aspirational expenses, cities like Pune, which experience a spike in both industry and population, are especially vulnerable to an increase in pollution. An accurate and reliable model for predicting and evaluating atmospheric pollution condition becomes essential since it can provide advanced information on air pollutants and therefore direct control and protection activities in advance. The need to create approaches and find a solution to this problem is essential. For researchers studying air quality, predicting the quantity of air pollutants in metropolitan areas is crucial. Long-term air pollution control is necessary to stop the issue from getting worse. Techniques used in air quality forecasting and their improvements have received a lot of attention. In developing nations, air pollution is worse than in developed nations, and furthermore, limited research has been done to lower emissions. Because of this, the issues of air pollution in developing nations warrant greater focus. Thus, recording, evaluating, and forecasting the ambient air pollution levels in cities have emerged as essential components of successful urban air quality management strategies. The availability of sensor data and environmental sensing networks, along with improvements in artificial intelligence applications, have caused researchers to employ artificial intelligence techniques more frequently. With higher concentration of enterprises, industries, and people in metropolitan areas, it has become more challenging to maintain acceptable standards of air quality. With the increasing population comes a rise in transportation, fuel use, and power consumption. To ensure that to comprehend how both humans and the natural world contribute to air pollution, attention must be paid to the chemistry and emissions of air-pollutants. Therefore, developing an analytical model and making an accurate prediction are essential.

II. RELATED STUDIES

To project the future condition of the air, authors in [6] utilized data analysis techniques. The most recent advancements in AI-based technology can be advantageous for the administration and control of pollution reduction and mitigation activities, as explored in [7]. They looked at literature that was pertinent to the field of using AI to monitor and control processes that minimize and mitigate pollution. Using the proposed ICEEMD-ISCA-LSSVM model, authors in [8] predicted the concentration of six air pollutants. Three Chinese cities were used to assess the model's performance. Artificial Neural Networks (ANNs) are frequently used to estimate or forecast air pollution levels [9]. Air pollutants, such as NO_x, SO_2, CO, O_3, and suspended PM, are typically measured by their concentrations in the atmosphere of the affected areas. ANNs are complex numerical models that do not directly provide specific details on air pollution levels based on pollution indicators (such as traffic and meteorological elements), even if they are capable of reliably estimating air pollution levels. Authors in [10] developed a model that has the ability to predict daily levels of air pollutants. For air quality predictions, they employed the AI-based Neuro-Fuzzy (NF) model, and they selected the NO_2 pollutant concentration for examination. Applying Deep Learning (DL) techniques with a Recurrent Neural Network (RNN), authors in [11] created an intelligent prediction for the next two days' air pollution concentrations. Authors in [12] used Support Vector Regression (SVR), which is a variation of Support Vector Machines (SVMs) and is considered one of the most potent Machine Learning (ML) techniques, to create models for hourly air quality forecasting for the state of California. An econometric model forecast of the pollutant emissions from domestic flights at Salvador Airport was given in [13]. A derivative analysis of demand is included in this model up to 2020, taking into account the quantity of flights and emissions produced throughout the takeoff and landing cycles. The findings provide a forecasting model for the quantity of pollutants released. One of the first applications of DL algorithms to forecast air pollution time series is presented in [14]. Authors in [15] provide a comprehensive review on modeling with DL architectures on actual data of air pollution. They used this research to create DL air pollution architectures in the future and improve the outcomes even more by applying new insights from DL research. Authors in [16] investigated the possibility and usefulness to use data from Internet of Things (IoT) sensors to create a real-time air quality prediction system that can be used by individuals and public organizations to monitor and control air pollution. Regression, classic Long Short-Term Memory (LTSM), and bidirectional LSTM (BLSTM) models were used to assess the models' performance on multivariate air quality variables using the standard dataset that the Indian government provided. The suggested BLSTM model performed better than the other models in reducing RMSE errors and preventing overfitting. In order to forecast vehicle-induced air pollution levels, authors in [17] used sophisticated tree-based ML models, paying special attention to fine PM ($PM_{2.5}$). The models used were Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGBoost), Extra Tree (ET) which fared better than the others, and Random Forest (RF), in addition to a benchmark statistical model. In order to overcome the ET model's lack of explainability, the best ET models were then interpreted using SHAP analysis. Temperature, humidity, and wind speed were found to be the main factors influencing $PM_{2.5}$ level predictions based on the SHAP research. The current status of Generative Adversarial Network (GAN) technology was compiled in [18] along with future directions. The authors conducted a survey on the background of the GAN proposal, theoretical and practical models, and application domains. They discussed the benefits and drawbacks of GANs as well as their current development trends. GANs were examined and reviewed in [19]. The authors looked into the many kinds of GANs, their uses, and upcoming advancements in the field. They suggested that GANs represent one of the most promising areas within the specialized domain of ANNs, which remains underdeveloped and has yet to fully realize its scientific potential.

III. AREA OF STUDY AND DATA COLLECTION

The study carried out considered the Pune City region of Maharashtra State of India. With an estimated 7.4 million residents as of 2020, Pune is the second-largest metropolis in the Maharashtra state and India's seventh most populated city overall. Pune, an Indian metropolis, has seen a sharp decline in air quality in the last few years with 2018 being the second most polluted year since 2013, according to data gathered by the Indian Institute of Tropical Meteorology (IITM). In 2013, the average annual concentration of $PM_{2.5}$ was 29 $\mu g/m^3$, but by 2018, it had climbed by almost 60% to 47 $\mu g/m^3$, surpassing the yearly threshold for national ambient air quality for $PM_{2.5}$, which is 40 $\mu g/m^3$. Approximately 73 days in 2017 and 94 days in 2018, respectively, surpassed the yearly pollution standard, according to statistics from IITM's continuous air quality monitoring program. Pune's annual air quality trends, as analyzed by the Urban Emissions Air Pollution Knowledge Assessment (APnA) program, were recently published. This research shows that Pune's air quality is still a serious public health concern. Pune currently has 15 air pollution surveillance stations, which are run by three different systems. Five monitoring stations are run by the Maharashtra Pollution

Control Board (MPCB), while the other 10 are run by the IITM as part of its System of Air Quality and Weather Forecasting and Research (SAFAR) program [20]. In addition, the Pune Smart City Development Corporation keeps an autonomous network of 50 environmental monitoring stations that track radiation, temperature, humidity, noise, and air pollution.

In this study, five monitoring stations from SAFAR were considered. 5 monitoring stations were considered: Pashan, Shivajinagar, Hadapsar, Katraj and Bhumkar Chowk. The data collected were air pollutant data (daily) for years 2016, 2017, 2018, 2019, 2020, 2021 regarding $PM_{2.5}$, PM_{10}, O_3, NO_2, and CO. The continuous, multivariate time series of air quality data are made up of readings that represent a fixed measurement of time, with each subsequent reading being dependent on the previous one in some manner. Pre-processing was conducted on the gathered dataset to remove duplicate, null, and missing values.

IV. GENERATIVE ADVERSARIAL NETWORK

A GAN, which is built on DL, has been utilized to forecast the pollution levels in a given area for a given time frame. The Time-series GAN (TimeGAN) was introduced in 2019 and is capable of producing genuine time-series data, or sequential data with distinct observable behaviors, in a range of diverse areas. The TimeGAN architecture incorporates supervised loss, leveraging original data as supervision to better capture time-dependent distributions compared to other GAN architectures like WGAN, which rely solely on unsupervised adversarial loss across real and synthetic data. A learnt embedding space with supervised and unsupervised losses is used to train TimeGAN, a generative time-series model, both cooperatively and adversarially. As such, the proposed approach incorporates concepts from autoregressive models for sequence prediction, thereby straddling the boundaries of several research streams.

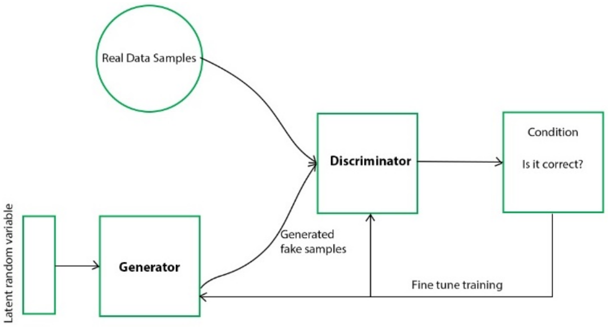

Fig. 1. Workflow of the GAN model.

The GAN is optimized using back-propagation to minimize the Generative Loss (GL):

$$\mathrm{GL} = -\left(\frac{1}{m}\right)\sum_{i=1}^{m}\log((x_i - z_i))^2 \qquad (1)$$

where x_i is actual value and z_i is the predicted value.

The loss is minimized when the log probability of predicting values closer to actual values is maximized. The Discriminative Loss (DL) is the simple sum of log-likelihood of classifying a given prediction as genuine or fake. The DL function is defined as the binary cross-entropy loss:

$$\mathrm{DL} = -\left(\frac{1}{m}\right)((\textstyle\sum_{i=1}^{m} logD(x_i)) - \sum_{i=1}^{m}\log(1 - D(G(z_i)) \qquad (2)$$

where $D(x_i)$ represents the probability of a given sample classified as real and (1-$D(G(z_i))$ is the probability of the predicted sample being classified as fake.

The steps adopted for the development of the model are:

1. Initialize Generator and Discriminator networks with random weights.
2. Generate a new prediction z_i using the Generator network based on current weights and a normal random noise sample.
3. Predict the probability outputs of the Discriminator network by inputting a real sample x_i and a predicted value from previous step z_i.
4. Calculate the GL and the DL.
5. If the probability of x_i being real is closer to 1 and the probability of z_i being real is closer to 0, then the combined loss should be minimal.
6. If the probability of x_i being real is closer to 1 and the probability of z_i being real is closer to 1, then the DL is large and the GL should be minimal.
7. If the probability of x_i being real is closer to 0 and the probability of z_i being real is closer to 1, then the DL is minimal and the GL is large.
8. Update the weights of the Generator network using gradient descent over GL and update the weights of the Discriminator network using the DL.
9. Repeat steps 2-6 m times for all input data samples.
10. Repeat steps 2-7 n times (set using hyper-parameter tuning).

V. RESULTS AND DISCUSSION

The GAN model for prediction was constructed by considering all pollutants for the considered stations of Pune region. Prediction for the last 7 days for the data considered, that means from 25th December 2021 to 31st December 2021. Figures 2 to 26 show the actual and the predicted values of $PM_{2.5}$, PM_{10}, O_3, NO_2, and CO for the regions Pashan, Shivajinagar, Hadapsar, Katraj, and Bhumkar Chowk. The blue line shows the original/observed values, whereas the orange line shows the produced/predicted values. Based on the figures, it is evident that the predicted values closely align with the actual values. The results demonstrate that the proposed GAN model is the most precise in predicting the concentration of air pollutants. This capability will empower policymakers and government agencies to make informed decisions by leveraging data analysis and interpretation of pollutant concentration forecasts, which are inherently driven by data. The generated model will be very helpful in informing the public about the expected level of air quality in metropolitan areas and shielding them from the negative effects of low ambient air quality. It can also be utilized to determine the appropriate operational measures and mitigation techniques.

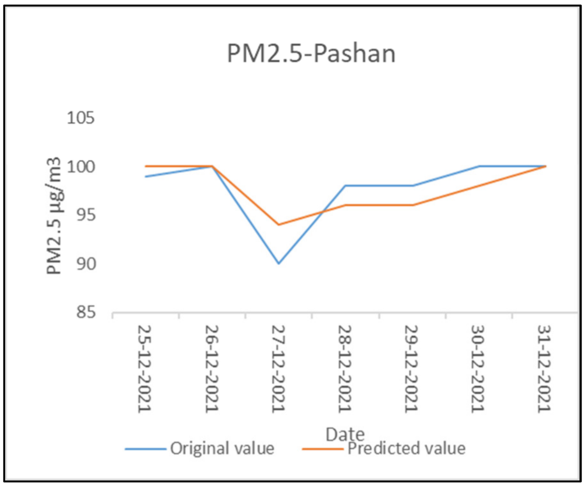

Fig. 2. Comparison between the original value and the GAN model prediction for $PM_{2.5}$ in Pashan.

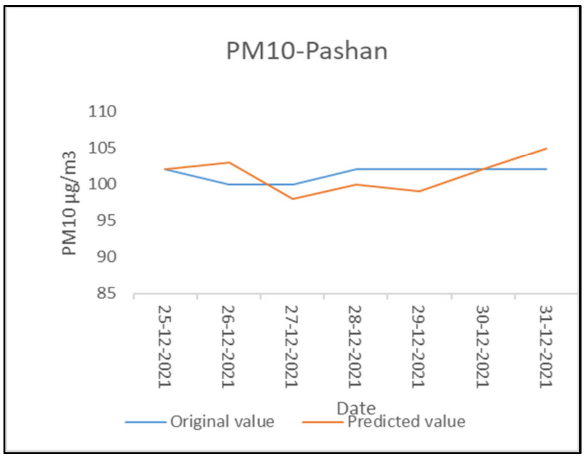

Fig. 3. Comparison between the original value and the GAN model prediction for PM_{10} in Pashan.

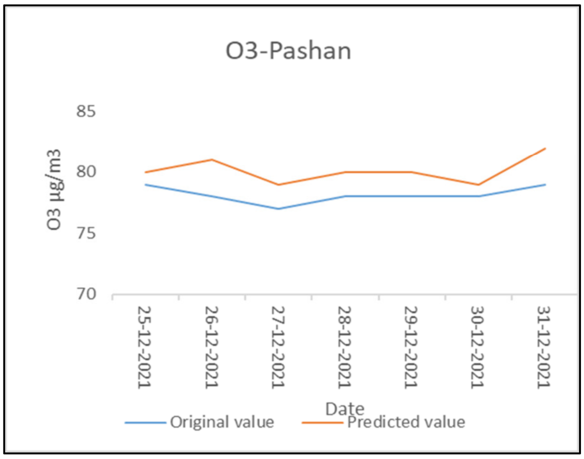

Fig. 4. Comparison between the original value and the GAN model prediction for O_3 in Pashan.

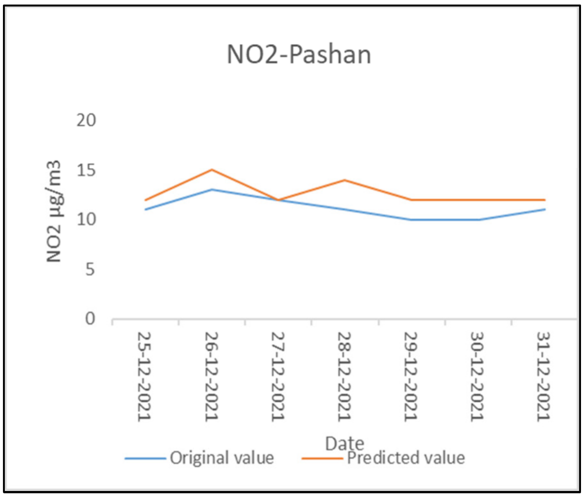

Fig. 5. Comparison between the original value and the GAN model prediction for No_2 in Pashan.

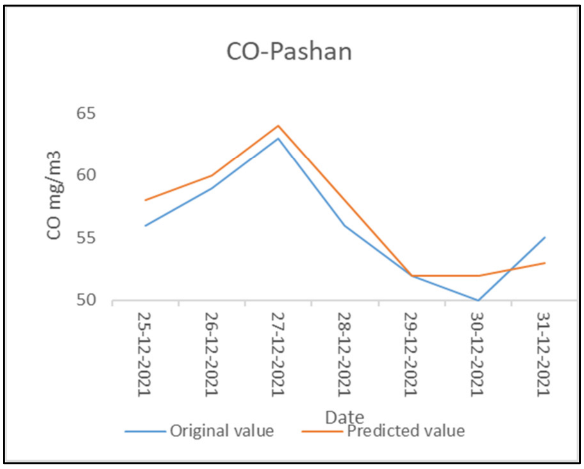

Fig. 6. Comparison between the original value and the GAN model prediction for CO in Pashan.

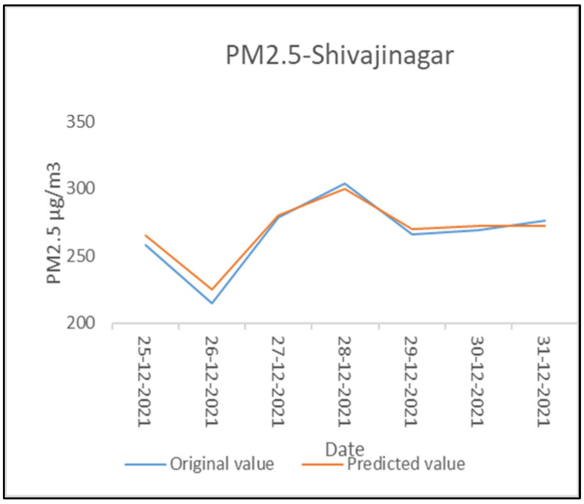

Fig. 7. Comparison between the original value and the GAN model prediction for $PM_{2.5}$ in Shivajinagar.

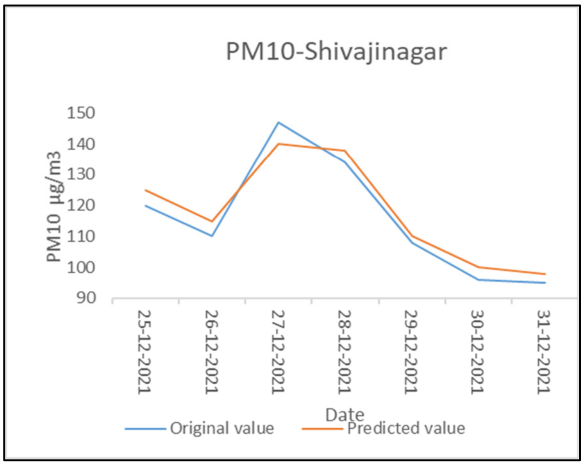

Fig. 8. Comparison between the original value and the GAN model prediction for PM_{10} in Shivajinagar.

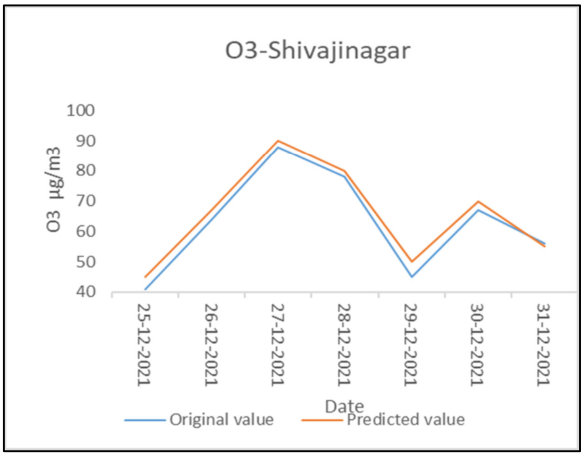

Fig. 9. Comparison between the original value and the GAN model prediction for O_3 in Shivajinagar.

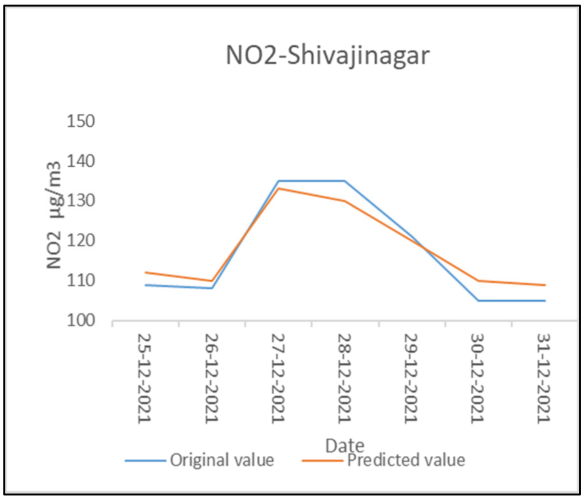

Fig. 10. Comparison between the original value and the GAN model prediction for NO_2 in Shivajinagar.

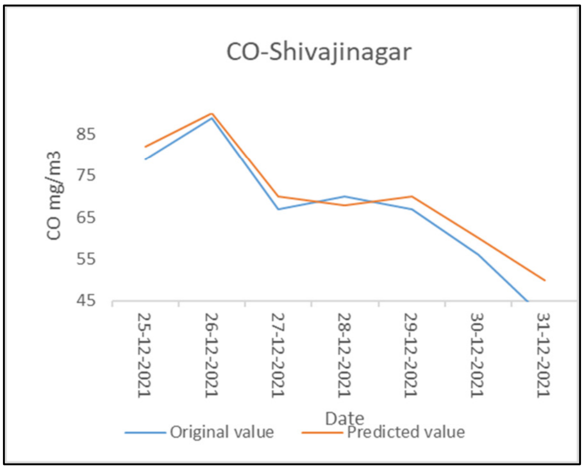

Fig. 11. Comparison between the original value and the GAN model prediction for CO in Shivajinagar.

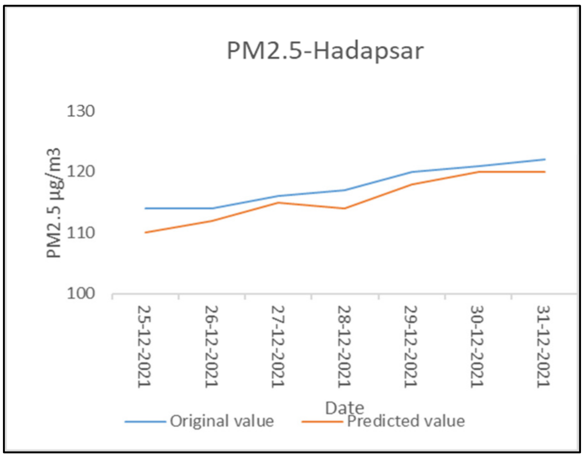

Fig. 12. Comparison between the original value and the GAN model prediction for $PM_{2.5}$ in Hadapsar.

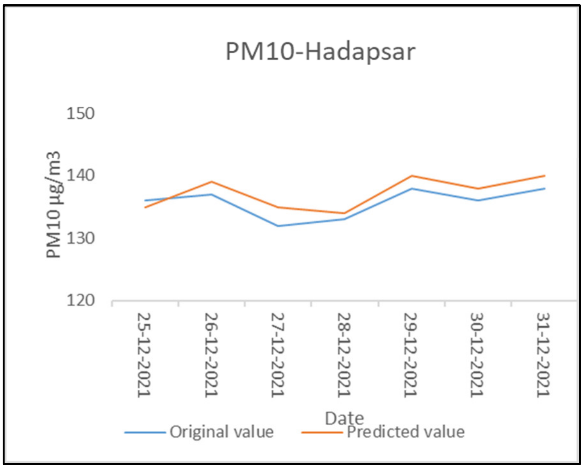

Fig. 13. Comparison between the original value and the GAN model prediction for PM_{10} in Hadapsar.

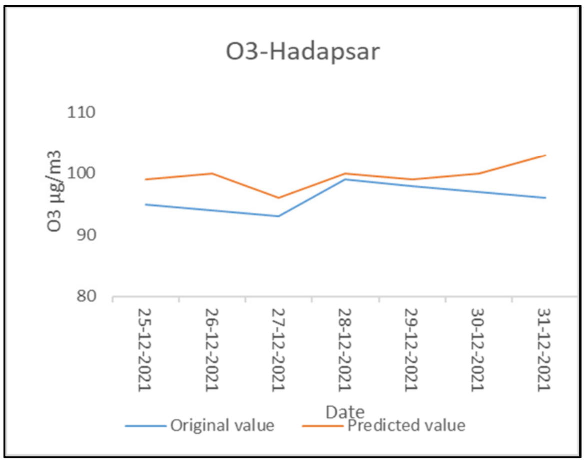

Fig. 14. Comparison between the original value and the GAN model prediction for O_3 in Hadapsar.

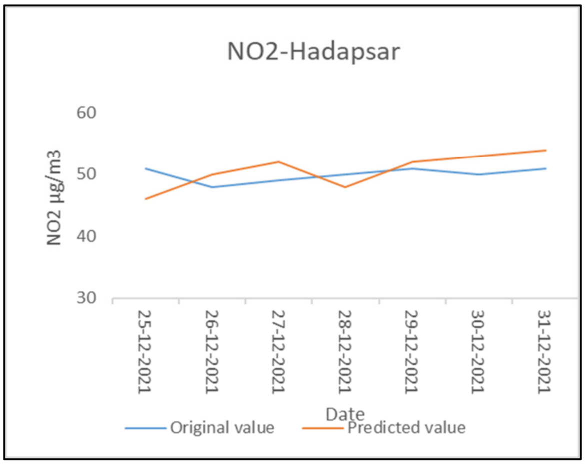

Fig. 15. Comparison between the original value and the GAN model prediction for NO_2 in Hadapsar.

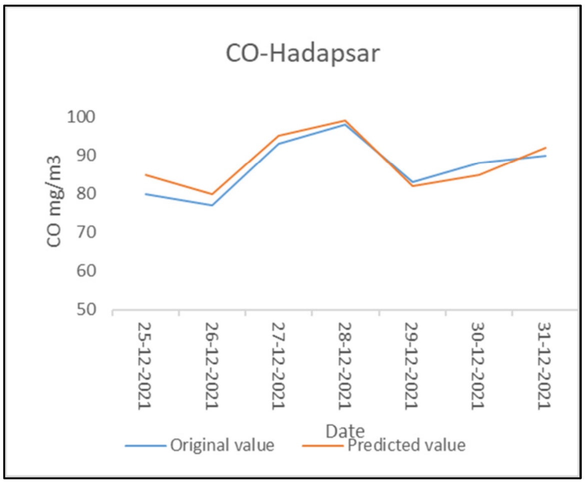

Fig. 16. Comparison between the original value and the GAN model prediction for CO in Hadapsar.

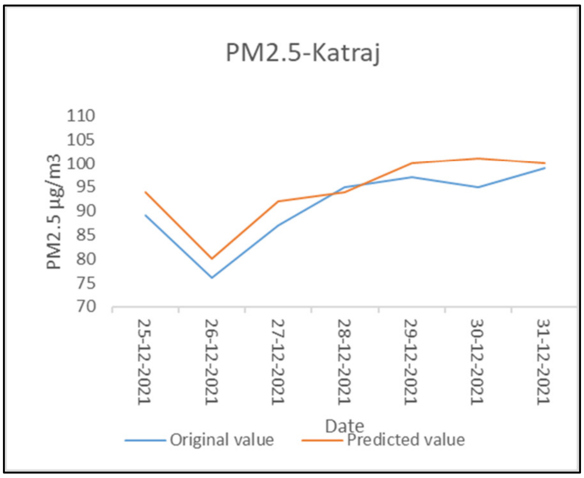

Fig. 17. Comparison between the original value and the GAN model prediction for $PM_{2.5}$ in Katraj.

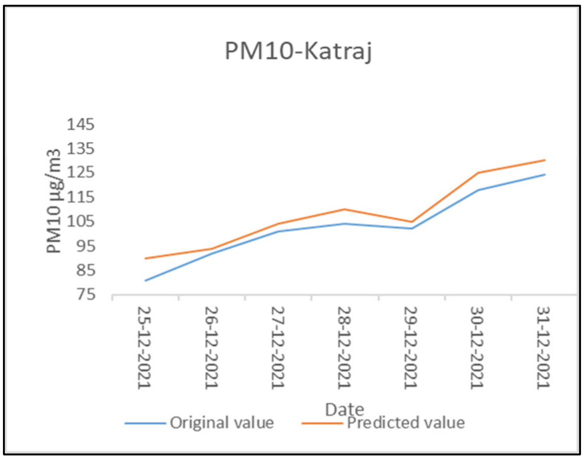

Fig. 18. Comparison between the original value and the GAN model prediction for PM_{10} in Katraj.

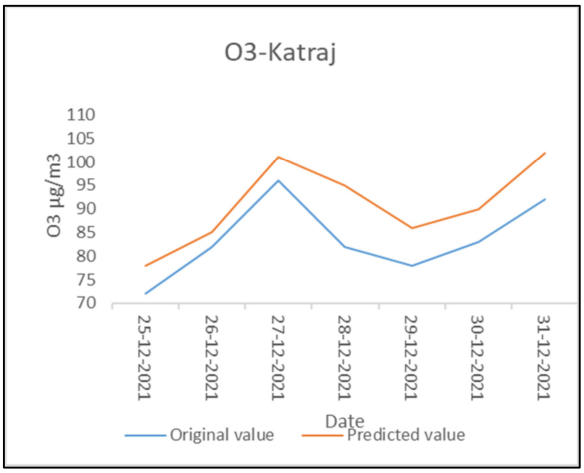

Fig. 19. Comparison between the original value and the GAN model prediction for O_3 in Katraj.

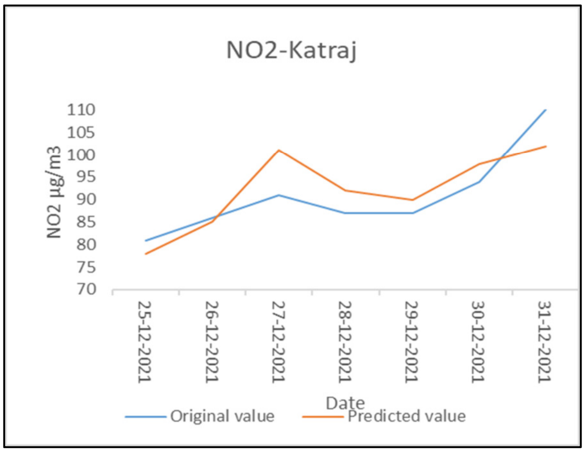

Fig. 20. Comparison between the original value and the GAN model prediction for NO_2 in Katraj.

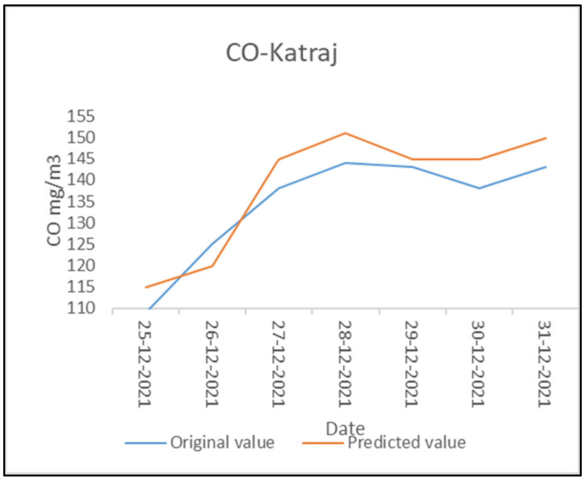

Fig. 21. Comparison between the original value and the GAN model prediction for CO in Katraj.

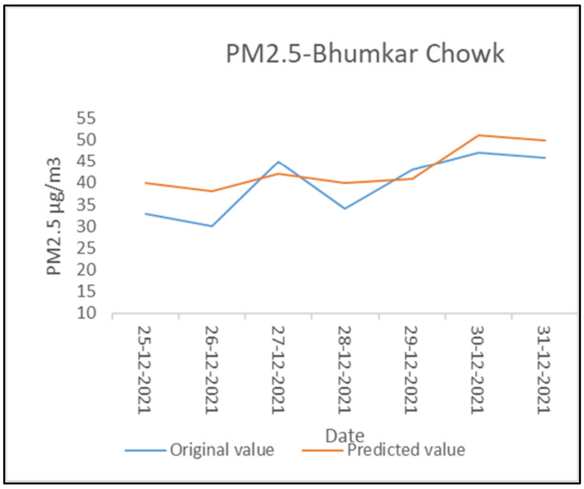

Fig. 22. Comparison between the original value and the GAN model prediction for $PM_{2.5}$ in Bhumkar Chowk.

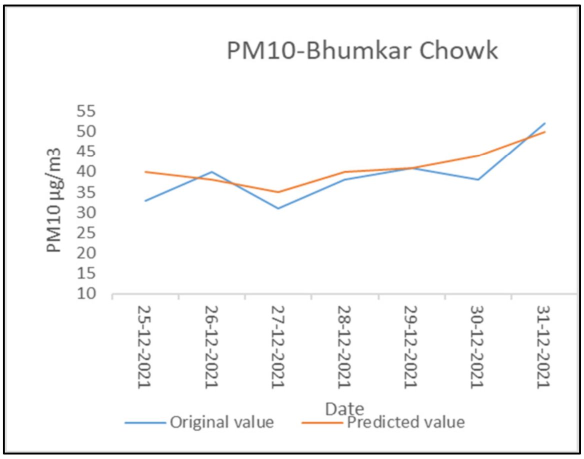

Fig. 23. Comparison between the original value and the GAN model prediction for PM_{10} in Bhumkar Chowk.

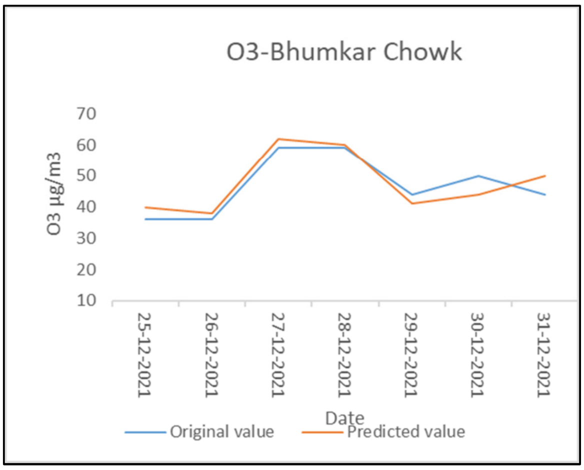

Fig. 24. Comparison between the original value and the GAN model prediction for O_3 in Bhumkar Chowk.

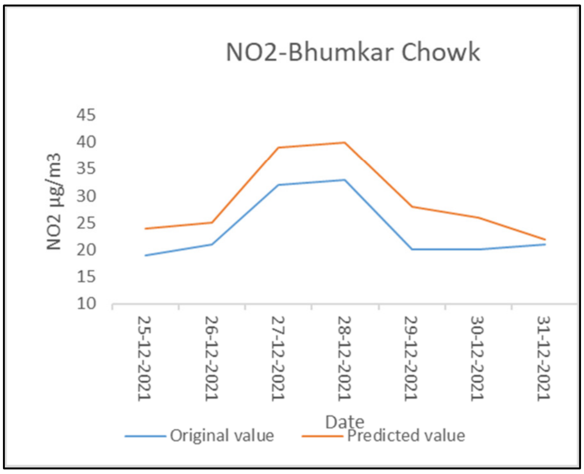

Fig. 25. Comparison between the original value and the GAN model prediction for NO_2 in Bhumkar Chowk.

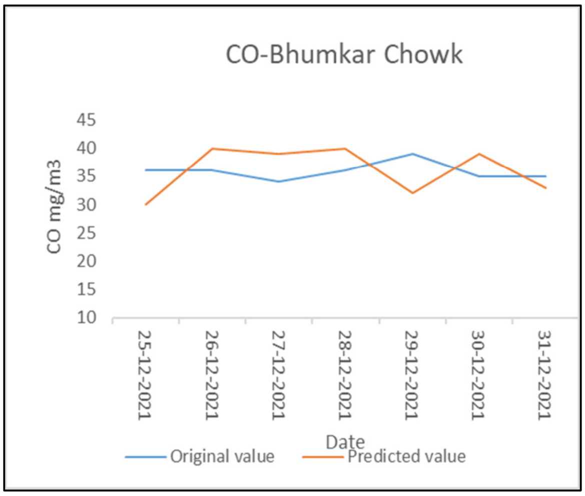

Fig. 26. Comparison between the original value and the GAN model prediction for CO in Bhumkar Chowk.

VI. EXPERIMENTAL ANALYSIS

Our approach utilizes Root Mean Square Error (RMSE) as a metric to assess predictive accuracy. The accuracy of the GAN model is evaluated as the RMSE between original and predicted values. The results can be seen in Table I and indicate high accuracy.

TABLE I. RMSE VALUES OF THE GAN MODEL

Pollutant	RMSE
$PM_{2.5}$	0.1099
PM_{10}	0.1445
O_3	0.4157
NO_2	0.0870
CO	0.1786

VII. CONCLUSION

This study presents a system that uses a GAN model based on deep learning to predict the air quality utilizing a publicly available air pollution dataset for Pune. The proposed deep learning model was used to predict the concentrations of pollutants $PM_{2.5}$, PM_{10}, NO_2, CO, and O_3. The outcomes demonstrate both the superior performance of the proposed model and the proximity of the predicted and the actual values. This is one of the first attempts of using a GAN model in the prediction of air pollution levels. Therefore, this GAN model can be extended to other heavily polluted Indian cities and its predictions can be utilized to develop strategies for pollution control and public health advisories. In this study, the meteorological parameters were not considered in the prediction process. In a next study, the authors will include these parameters to find their effect on the prediction.

ACKNOWLEDGMENT

The authors wish to thank to the faculties and staff of Department of Civil Engineering at D.Y Patil Institute of Technology, Pimpri and AISSMS COE, Pune, for the technical support.

REFERENCES

[1] A. Kanaujia, M. Bhati, L. Sandhiya, S. N. Nishad, and S. Bhattacharya, *Air Pollution in India: A Critical Assessment and Suggestive Pathways for Clean Air*. New Delhi, India: National Institute of Science Communication and Policy Research, 2022.

[2] A. Jaiswal, C. Samuel, and V. M. Kadabgaon, "Statistical trend analysis and forecast modeling of air pollutants," *Global Journal of Environmental Science and Management*, vol. 4, no. 4, pp. 427–438, Oct. 2018, https://doi.org/10.22034/gjesm.2018.04.004.

[3] B. R. Gurjar, "Air Pollution In India: Major Issues And Challenges." https://www.magzter.com/stories/Education/Energy-Future/Air-Pollution-In-India-Major-Issues-And-Challenges.

[4] WHO, *World Air Quality Report*. Geneva, Switzerland: World Health Organization, 2020.

[5] K. Ravindra, M. K. Sidhu, S. Mor, S. John, and S. Pyne, "Air Pollution in India: Bridging the Gap between Science and Policy," *Journal of Hazardous, Toxic, and Radioactive Waste*, vol. 20, no. 4, Oct. 2016, Art. no. A4015003, https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000303.

[6] N. Sharma, S. Taneja, V. Sagar, and A. Bhatt, "Forecasting air pollution load in Delhi using data analysis tools," *Procedia Computer Science*, vol. 132, pp. 1077–1085, Jan. 2018, https://doi.org/10.1016/j.procs.2018.05.023.

[7] C. W. Chan and G. H. Huang, "Artificial intelligence for management and control of pollution minimization and mitigation processes," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 2, pp. 75–90, Mar. 2003, https://doi.org/10.1016/S0952-1976(03)00062-9.

[8] R. Li, Y. Dong, Z. Zhu, C. Li, and H. Yang, "A dynamic evaluation framework for ambient air pollution monitoring," *Applied Mathematical Modelling*, vol. 65, pp. 52–71, Jan. 2019, https://doi.org/10.1016/j.apm.2018.07.052.

[9] K. Yan Chan and L. Jian, "Identification of significant factors for air pollution levels using a neural network based knowledge discovery system," *Neurocomputing*, vol. 99, pp. 564–569, Jan. 2013, https://doi.org/10.1016/j.neucom.2012.06.003.

[10] D. Mishra and P. Goyal, "Neuro-Fuzzy Approach to Forecast NO2 Pollutants Addressed to Air Quality Dispersion Model over Delhi, India," *Aerosol and Air Quality Research*, vol. 16, no. 1, pp. 166–174, 2016, https://doi.org/10.4209/aaqr.2015.04.0249.

[11] S. Al-Janabi, M. Mohammad, and A. Al-Sultan, "A new method for prediction of air pollution based on intelligent computation," *Soft Computing*, vol. 24, no. 1, pp. 661–680, Jan. 2020, https://doi.org/10.1007/s00500-019-04495-1.

[12] 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, no. 1, 2020, Art. no. 8049504, https://doi.org/10.1155/2020/8049504.

[13] D. A. Pamplona and C. J. P. Alves, "Civil Aircraft Emissions Study and Pollutant Forecasting at a Brazilian Airport," *Engineering, Technology & Applied Science Research*, vol. 10, no. 1, pp. 5217–5220, Feb. 2020, https://doi.org/10.48084/etasr.3227.

[14] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, Aug. 2018, https://doi.org/10.1080/10962247.2018.1459956.

[15] Y. A. Ayturan, Z. C. Ayturan, and H. O. Altun, "Air Pollution Modelling with Deep Learning: A Review," *International Journal of Environmental Pollution and Environmental Modelling*, vol. 1, no. 3, pp. 58–62, 2018.

[16] P. Mullangi *et al.*, "Assessing Real-Time Health Impacts of outdoor Air Pollution through IoT Integration," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13796–13803, Apr. 2024, https://doi.org/10.48084/etasr.6981.

[17] C. Matara, S. Osano, A. O. Yusuf, and E. O. Aketch, "Prediction of Vehicle-induced Air Pollution based on Advanced Machine Learning Models," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12837–12843, Feb. 2024, https://doi.org/10.48084/etasr.6678.

[18] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," *Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017, https://doi.org/10.1109/JAS.2017.7510583.

[19] J. Patel, M. Pandya, and V. Shah, "Review on Generative Adversarial Networks," *International Journal of Technical Innovation in Modern Engineering & Science*, vol. 4, no. 7, pp. 1230–1235, Jul. 2018.

[20] "Air Quality Early Warning System For Delhi," *IITM*. https://ews.*tropmet*.res.in/.