Nitrogen Dioxide Concentration Forecasting Based on Machine Learning Algorithms: New Borg El Arab City, Alexandria, Egypt as a Case Study

Mostafa M. Abdelmalek[1,2,a)], Hatem Mahmoud[1,3,b)], Hassan Shokry[1,c)]

[1]*Environmental Engineering Department, Egypt-Japan University of Science and Technology, Alexandria 21934, EGYPT.*
[2]*Mining and Metallurgical Engineering Department, Faculty of Engineering, Assiut University, Assiut 71516, EGYPT.*
[3]*Department of Architecture Engineering, Faculty of Engineering, Aswan University, Aswan 81542, EGYPT.*

[a)] Corresponding author: mostafa.abdelmalek@ejust.edu.eg.
[b)]hatem.mahmoud@ejust.edu.eg.
[c)]hassan.shokry@ejust.edu.eg.

Keywords: Air quality, Nitrogen Dioxide, Machine Learning, Random Forest, Support Vector Machine.

Abstract: Nitrogen dioxide (NO_2) is a significant air pollutant primarily emitted from traffic and industrial activities, posing health risks. Accurate predictions of urban NO_2 concentrations are essential for effectively controlling air pollution. In this study, we focus on forecasting NO_2 levels in New Borg El-Arab City, Alexandria, Egypt—a rapidly developing industrial area—to enhance air quality management and urban planning. This research employs comparative analysis of three machine learning (ML) models, including Artificial Neural Networks (ANN), Random Forest (RF), and Support Vector Machines (SVM). Hourly datasets were collected from the New Borg El-Arab City Weather Station and an IoT-based air quality monitoring system with Arduino from 2nd January 2021 to 30th May 2021. While Key environmental and meteorological variables, such as Sulfur Dioxide (SO2), Fine Particulate Matter (PM2.5), Temperature (T), Relative Humidity (RH), and Wind Direction (WD), were collected, only four variables were selected to forecast NO_2 concentration based on their higher correlation with NO_2 as determined using the Correlation Matrix. The study employed R^2, RMSE, MAE, and MSE as evaluation metrics to assess the model's performance, ensuring robust comparisons. The findings indicate that ANN, RF, and SVM achieved a high accuracy, exceeding 91% for NO_2 prediction. The comparative analysis revealed that the ANN surpassed the other ML models with an RMSE of .7350 during training and 1.2281 for testing. This study contributes to the ongoing efforts to achieve sustainable urban development and improve public health outcomes in Egypt.

INTRODUCTION

Air pollution poses a significant threat to human health and is being scrutinized as environmental awareness grows[1]. This often-overlooked hazard is responsible for countless fatalities each year. Air pollution claims over seven million lives annually, with outdoor pollutants alone accounting for roughly 4.2 million deaths[2]. Conversely, exposure to indoor air pollution contributes to approximately 3.8 million deaths yearly[3]. As a rapidly developing nation, Egypt is experiencing significant environmental challenges driven by accelerated industrial growth, population expansion, extensive construction and demolition activities, and a marked increase in traffic volume. These factors have contributed to a deterioration in air quality, positioning air pollution as one of Egypt's most pressing environmental concerns[4]. New Borg El-Arab, a prominent industrial city in Alexandria Governorate, encompasses approximately 1200 factories distributed across four industrial zones. These zones host a diverse array of industries, including engineering, electrical, food processing, timber, plastics, paper, textiles, building materials, mechanical, chemical, and pharmaceutical sectors. Such industrial diversity contributes significantly to the city's economic development. However, the concentration of these industries has led to environmental concerns, particularly regarding air quality. Industrial activities, particularly those involving combustion processes in the chemical and engineering sectors, are recognized as significant sources of NO_2 emissions. NO_2 is a harmful pollutant that can have adverse

effects on human health and the environment. Studies have shown that areas with dense industrial operations, such as New Borg El-Arab, may experience elevated levels of NO_2, highlighting the need for effective emission control and air quality management strategies. Nitrogen dioxide is recognized worldwide as a significant contributor to air pollution. Although natural events, such as dust storms, bushfires, and volcanic eruptions, contribute to air quality degradation[5]. NO_2 emissions originate from both indoor and outdoor sources. Elevated NO_2 concentrations primarily stem from outdoor activities, such as high-temperature combustion, vehicular emissions, and industrial operations, with additional contributions from indoor sources, including gas appliances and tobacco smoke[6]. Moreover, A significant portion of atmospheric NO_2 arises from secondary formation through photochemical reactions, wherein nitric oxide (NO), the predominant nitrogen oxide, is quickly converted to NO_2 when exposed to ozone[7]. NO_2 poses risks to human health and has detrimental effects on the environment and ecosystems, contributing to phenomena like acid rain, depletion of the ozone layer, and climate change[8]. The World Health Organization (WHO) reported that NO_2 is associated with adverse effects such as increases in respiratory symptoms, asthma prevalence, cancer incidence, adverse birth outcomes, and mortality[9]. Consequently, governments need robust detection and predictive models to provide early warnings and inform effective control measures on NO_2 concentrations. Applying machine learning techniques in air pollution analysis and air quality prediction has encompassed various methodologies. Regression-based models—such as SVM and RF—are frequently employed due to their ability to manage nonlinear relationships and relative ease of interpretation. These models have proven particularly effective in short-term air quality forecasting by capturing the temporal variations in pollutant concentrations[10]. Additionally, ANNs are extensively applied in both short-term and long-term pollutant prediction tasks[11]. [12]Employed the RF model to predict the concentration of NO_2 from traffic flow and meteorological information. The R^2 reached .82, demonstrating the effect of traffic on NO_2 levels.

MATERIALS AND METHODS

Three fundamental steps structured the progress of this research. Weather and air pollution data underwent processing to ensure quality and consistency. The second phase involved feature extraction to examine the relationship between NO_2 concentrations and other meteorological and pollution-related variables. In the final phase, machine learning models were developed and assessed. FIGURE 1 illustrates the overall framework used in this study.

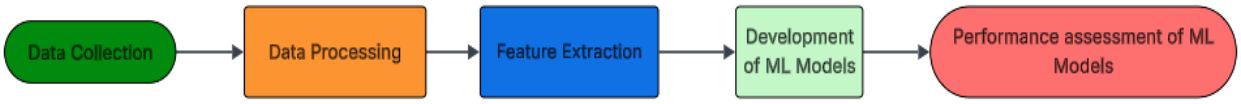


FIGURE 1. A flowchart illustrating the overall process of NO_2 forecasting

DATA PROCESSING

The dataset used in this study was collected from the New Borg El-Arab City Weather Station and an IoT-based air quality monitoring system with Arduino 2nd January 2021 to 30th May 2021. This data should undergo several phases to be ideal as input for ML models.

(a) Outliers and missing values removal using the data parameter and Box plotting.
(b) Data scaling is based on the min-max normalization technique.
(c) Data splitting into training and testing sets (90% for training and 10% for testing).

FEATURE EXTRACTION

ML models require independent variables (input features) that correlate highly with the target variable (NO_2) to achieve reasonable accuracy in the prediction process. We used the Correlation Matrix to investigate the relationship between input features and the target variable. The Correlation Matrix showed a strong positive association between PM2.5 and NO_2, as well as between SO_2 and NO_2, with correlation coefficients of 0.9714 and 0.7612, respectively. In contrast, WD, T, and RH show a negative correlation with NO_2, with coefficients of 0.6575, 0.3835, and 0.0827, respectively. As a result, the first four variables were selected as inputs for the ML models due to their strong correlation with NO_2 concentration. Conversely, the last variable was excluded because it exhibited a weak association with NO_2.

DEVELOPMENT OF ML MODELS

This study utilizes three state-of-the-art machine learning models—ANN, RF, and SVM—to predict NO_2 based on PM2.5, SO_2, WD, and T. Each model has unique advantages for regression tasks. ANN models represent a recent advancement in applying artificial intelligence to air pollution prediction and are widely employed for both forecasting and predictive analysis[13]. RF is an ensemble-based approach that generates predictions by aggregating the outputs of multiple decision trees, and SVM has gained considerable recognition over the past two decades as a novel and powerful statistical learning technique[14].

PERFORMANCE ASSESSMENT OF ML MODELS

The study evaluated all models based on their ability to forecast NO_2 levels using four performance indicators:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

$$MSE = \frac{1}{n}\sum_{n=1}^{n}(x_i - y_i)^2 \quad (4)$$

Where R^2 = Determination coefficient, MAE = Mean absolute error, RMSE = Root mean square error, MSE = Mean squared error, x_i = Actual values, y_i = Predicted values, $\bar{x}$ = The mean of actual values, and n = Number of data points.

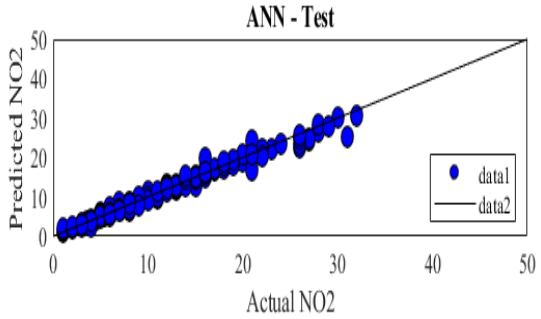
RESULTS AND DISCUSSION

TABLE 1. Model performance assessment during training and testing phases

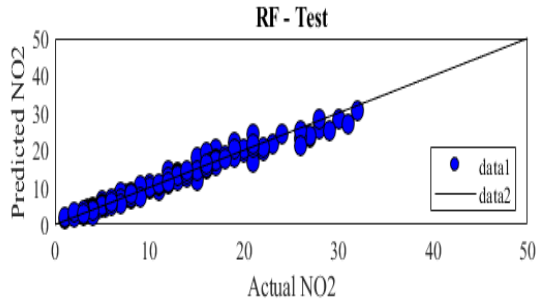
Parameter	Statistical index	Training	Testing
ANN	R^2	0.9893	0.9800
	RMSE	0.7352	1.0047
	MAE	0.5526	0.6503
	MSE	0.5403	1.5082
RF	R^2	0.9749	0.9717
	RMSE	1.1268	1.1943
	MAE	0.6938	0.9112
	MSE	1.2270	2.5701
SVM	R^2	0.9400	0.939
	RMSE	1.7428	1.7539
	MAE	0.8481	0.9345
	MSE	3.2455	5.0488

As shown in TABLE 1 and FIGURES 2 and 3, during the training phase, the three distinct algorithms (ANN, RF, and SVM) demonstrate different degrees of predictive precision and error-handling capability, underscoring their unique strengths in capturing the complexity of NO_2 data. ANN is a highly accurate model during the training phase, with an R^2 of 0.9893 and RMSE and MAE values of 0.7352 and 0.5526, respectively. This demonstrates its ability to achieve the lowest prediction error, establishing it as the most reliable model among the algorithms examined. On the other

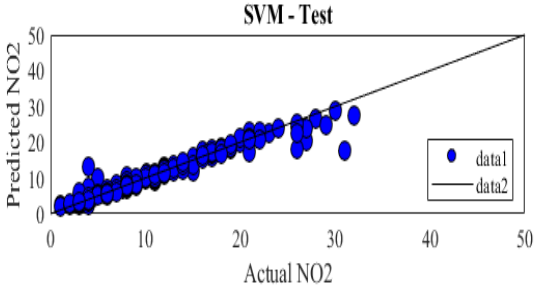
hand, SVM scored 0.9400, 1.7428, and 0.8481 for R^2, RMSE, and MAE, respectively, suggesting limitations in accurately learning the intricate patterns of the NO_2 dataset during the training process. In the case of RF, it achieved an acceptable value for R^2 and MAE, with 0.9749 and 0.6938, respectively. However, RMSE and MSE were higher compared to ANN, with values of 1.1268 and 1.2270, respectively. During the testing phase (see TABLE 1 and FIGURES 2 and 3), the ANN model exhibited the strongest predictive performance among the three algorithms. It achieved the highest R^2 value of 0.9800, indicating a strong correlation between the predicted and actual NO_2 concentrations. Furthermore, ANN recorded the lowest RMSE, MAE, and MSE values of 1.0047, 0.6503, and 1.5082, respectively, confirming its precision and reliability in generalizing unseen data. The RF model followed closely, with an R^2 value of 0.9717 and moderate error metrics (RMSE = 1.1943, MAE = 0.9112, MSE = 2.5701), indicating its solid capacity to capture data patterns during testing. In contrast, the SVM model demonstrated relatively lower performance, as evidenced by the lowest R^2 (0.9390) and the highest error rates across all indicators (RMSE = 1.7539, MAE = 0.9345, MSE = 5.0488), suggesting limitations in accurately modeling the complexity of NO_2 data during the testing stage.



a



b



c

FIGURE 2a,b,c. Comparison between predicted and actual NO_2 concentration during the ANN, RF, and SVM testing phase.

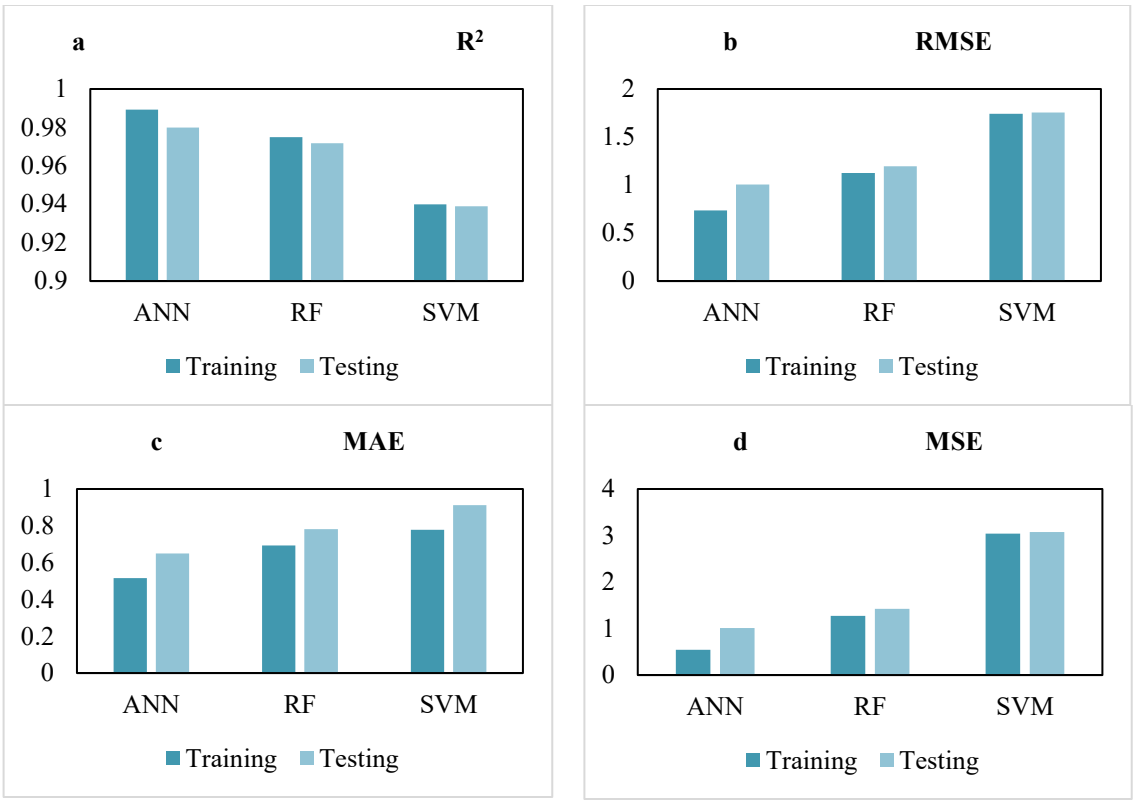


FIGURE 3a,b,c,d. Comparison between R^2, RMSE, MAE, and MSE during training and testing phases for ANN, RF, and SVM.

CONCLUSION

The prediction of NO_2 concentration in this study was in New Borg El-Arab City, Alexandria, Egypt, using a historical dataset from 2nd January 2021 to 30th May 2021. The employed ML models—ANN, RF, and SVM—exhibited varying degrees of effectiveness in predicting NO_2 concentrations. ANN demonstrated the highest accuracy across both phases, achieving an R^2 of 0.9893 in training and 0.9800 in testing. The corresponding error metrics were RMSE = 0.7352/1.0047, MAE = 0.5526/0.6503, and MSE = 0.5403/1.5082, confirming its strong generalization capability. RF followed with slightly lower but acceptable performance, recording R^2 values of 0.9749 (training) and 0.9717 (testing), and errors of RMSE = 1.1268/1.1943, MAE = 0.6938/0.9112, and MSE = 1.2270/2.5701. In contrast, SVM displayed the least effective performance, with R^2 values of 0.9400 (training) and 0.9390 (testing), and the highest error rates across both phases—RMSE = 1.7428/1.7539, MAE = 0.8481/0.9345, and MSE = 3.2455/5.0488. These findings highlight ANN as the most precise and reliable model for air quality prediction in this study. This analysis provides valuable insights into selecting suitable ML algorithms for environmental data modeling and can assist policymakers and urban planners in designing effective air pollution mitigation strategies.

REFERENCES

1. Du, X. *et al.* Integrated study of GIS and Remote Sensing to identify potential sites for rainwater harvesting structures. *Physics and Chemistry of the Earth, Parts A/B/C* **134**, 103574 (2024).
2. Zheng, X. *et al.* Coupling Remote Sensing Insights With Vegetation Dynamics and to Analyze NO2Concentrations: A Google Earth Engine-Driven Investigation. *IEEE J Sel Top Appl Earth Obs Remote Sens* **17**, 9858–9875 (2024).
3. Chen, G. *et al.* Numerical study on efficiency and robustness of wave energy converter-power take-off system for compressed air energy storage. *Renew Energy* **232**, 121080 (2024).
4. Hassan, S. K. & Khoder, M. I. Chemical characteristics of atmospheric PM2.5 loads during air pollution episodes in Giza, Egypt. *Atmos Environ* **150**, 346–355 (2017).

5. Rijnders, E., Janssen, N. A. H., van Vliet, P. H. N. & Brunekreef, B. Personal and outdoor nitrogen dioxide concentrations in relation to degree of urbanization and traffic density. *Environ Health Perspect* **109**, 411 (2001).
6. Mohammadi, M. J. *et al.* Dispersion Modeling of Nitrogen Dioxide in Ambient Air of Ahvaz City. *Health Scope* **5**, (2016).
7. Mavroidis, I. & Chaloulakou, A. Long-term trends of primary and secondary NO2 production in the Athens area. Variation of the NO2/NOx ratio. *Atmos Environ* **45**, 6872–6879 (2011).
8. Boningari, T. & Smirniotis, P. G. Impact of nitrogen oxides on the environment and human health: Mn-based materials for the NOx abatement. *Curr Opin Chem Eng* **13**, 133–141 (2016).
9. Samet, J. & Krewski, D. Health effects associated with exposure to ambient air pollution. *Journal of Toxicology and Environmental Health - Part A: Current Issues* **70**, 227–242 (2007).
10. Jaja-Wachuku, C., Garbagna, L., Saheer, L. B. & Oghaz, M. M. D. Improved NO2 Prediction Using Machine Learning Algorithms. *IFIP Adv Inf Commun Technol* **712**, 215–225 (2024).
11. Shams, S. R., Jahani, A., Kalantary, S., Moeinaddini, M. & Khorasani, N. Artificial intelligence accuracy assessment in NO2 concentration forecasting of metropolises air. *Sci Rep* **11**, 1805 (2021).
12. Kamińska, J. A. A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions. *Science of The Total Environment* **651**, 475–483 (2019).
13. Zhao, Y., Li, J., Wang, Y., Zhang, W. & Wen, D. Warming Climate-Induced Changes in Cloud Vertical Distribution Possibly Exacerbate Intra-Atmospheric Heating Over the Tibetan Plateau. *Geophys Res Lett* **51**, e2023GL107713 (2024).
14. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Min Knowl Discov* **9**, (2019).