ADVANCED MACHINE LEARNING TECHNIQUES FOR PREDICTING NO$_x$ LEVELS

by

Randa ALHARBI[a] ***and Abeer D. ALGARNI***[b*]

[a] Department of Statistics, Faculty of Science, University of Tabuk, Tabuk, Saudi Arabia
[b] Department of Information Technology, College of Computer and Information Sciences,
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Original scientific paper
https://doi.org/10.2298/TSCI2406979A

This study explores the application of machine learning techniques to forecast atmospheric pollutant concentrations, focusing on NO$_x$, NO$_2$, and NO over the period from January 1, 2017, to December 1, 2017. Accurate prediction of air pollutant levels is crucial for effective environmental monitoring and public health protection. The research employs the Gaussian mixture model and decision tree model to analyze and predict pollutant data. The methodology encompasses rigorous data preprocessing steps, including cleaning and normalization, followed by model training and validation using cross-validation techniques to enhance robustness. Model performance is assessed through multiple metrics, including entropy, log-likelihood, normalized entropy criterion, integrated completed likelihood, akaike information criterion, and Bayesian information criterion. Results demonstrate that the Gaussian mixture model outperforms other approaches in predicting air pollutant levels, offering improved accuracy and reliability for environmental forecasting.

Key words: *air pollutants, machine learning, prediction, decision tree model, Gaussian mixture model, statistical metrics*

Introduction

Air pollution, particularly from NO$_x$, NO$_2$, and NO, poses a significant threat to both environmental sustainability and public health. These pollutants contribute to respiratory and cardiovascular diseases and worsen climate-related challenges. Accurate predictions of their levels are essential for effective mitigation strategies, informed policy-making, and public health protection.

Recently, machine learning (ML) techniques have proven effective in analyzing complex environmental data and enhancing air quality forecasts. This study utilizes Gaussian mixture models (GMM) and decision tree models (DTM) to address the variability and uncertainty in air pollution data. These models are evaluated through various performance metrics, including entropy, log-likelihood, normalized entropy criterion (NEC), integrated completed likelihood (ICL), Akaike information criterion (AIC), and Bayesian information criterion (BIC), ensuring a comprehensive assessment of model reliability and accuracy.

The goal of this research is to improve the accuracy and dependability of air quality predictions, thereby supporting more effective environmental monitoring, timely interventions,

[*] Corresponding author, e-mail: adalqarni@pnu.edu.sa

and greater public awareness of pollution risks. The findings also highlight the value of combining advanced ML models with robust evaluation frameworks to address air quality forecasting challenges.

Literature survey

Accurately predicting air quality and pollutant levels is crucial for mitigating environmental risks and protecting public health. As concerns over air pollution grow, many studies have adopted ML techniques to model the air quality index (AQI) and forecast pollutant concentrations, offering promising solutions to handle the complexities of environmental data. This section reviews the key ML approaches used in air quality prediction.

The study [1] explored various methods for AQI modeling and pollution forecasting, comparing support vector machines (SVM), long short-term memory (LSTM), and seasonal autoregressive integrated moving average (SARIMA). The SVM with radial basis function (RBF) kernel outperformed the others, and outliers were addressed using the Z-Score method [2]. In another study, air pollution data from Delhi, India (2009-2017), revealed a significant rise in pollutants like PM10, NO_2, and PM2.5 between 2016 and 2017, indicating worsening pollution [3, 4]. Regression-based ML models were used to predict AQI, evaluated with MAE, mean absolute percentage error (MAPE), correlation coefficient (R^2), and root mean square error (RMSE). A study in Stuttgart, Germany, used ML models to estimate pollutant levels at Marienplatz and Am Neckartor, aiming to replace traditional monitoring stations with virtual ones [5]. The accuracy of predictions was improved by incorporating data from nearby stations. In Jakarta, Indonesia, the CatBoost algorithm was applied to predict urban air quality from 2010 to 2021, achieving 0.9781 accuracy, demonstrating its ability to handle environmental data and missing values [6, 7].

Rapid urbanization contributes to air pollution, with vehicular congestion and industrial activities worsening air quality [8]. A study in Makkah used data from 2016 to 2018 to develop predictive algorithms, with the ensemble boosting tree model achieving 97.4% accuracy, surpassing fuzzy decision tree and ensemble bagging tree models [9, 10]. The environmental impact of pollutants like SO_2 and NO_x is well-documented, contributing to acid rain and smog [11]. A study focused on coking facilities in China proposed a quantitative approach to forecast SO_2 emissions and set regulations for industrial pollutants [12].

Support vector regression and random forest regression have demonstrated superior performance in predicting pollutant concentrations [13]. The ImDFR model was also successful in predicting dioxin emissions from municipal waste incineration, optimizing the flue gas purification process [14]. Efforts to reduce maritime air pollution focus on improving ship design, operational efficiency, and adopting cleaner fuels like LNG and biofuels. Stronger regulation by the international maritime organization is essential for addressing pollution in under-studied regions [15]. Finally, a study in Chennai used a multivariate time series model and a real-time autoregressive approach to predict PM2.5 levels, with a weighted ensemble technique outperforming classical models like ARIMA and VAR [16].

In conclusion, ML techniques are increasingly valuable for predicting and managing air quality. Models like SVM, CatBoost, and ensemble learning approaches show promise across various urban and industrial contexts. Local data, such as nearby monitoring stations or historical trends, can improve prediction accuracy. However, challenges remain, including managing missing data and optimizing algorithms for complex environments. This study builds on this progress by using GMM and DTM to refine pollutant level predictions, with the aim of advancing forecasting methods.

The used machine learning models

This section presents the two ML models employed in this study to predict the levels of air pollutants: GMM and DTM. Both models have been selected due to their ability to handle complex environmental data and their proven effectiveness in predictive tasks.

Gaussian mixture model

The GMM is a powerful statistical tool for modeling complex data distributions by combining multiple Gaussian distributions, each with its own mean, variance, and weight. This flexibility allows GMM to capture multi-modal patterns, making it ideal for clustering data and forecasting distributions. In environmental science, pollutants like NO_x, NO_2, and NO often exhibit complex, multi-modal behavior, especially with seasonal variations or pollution peaks from specific sources. The GMM can effectively model these patterns by accounting for distinct clusters in the data, such as traffic emissions, industrial sources, or seasonal effects [17].

The probability density function of a GMM is a weighted sum of several Gaussian distributions. Each component represents a different cluster within the data, with specific means, variances, and weights reflecting the central tendency, dispersion, and relative importance of each cluster. This is particularly useful when air pollution data shows distinct groups corresponding to various pollution sources.

The expectation-maximization (EM) algorithm is an iterative method for estimating model parameters when data contains missing or latent variables. This is particularly relevant for environmental data, where gaps may occur due to sensor malfunctions or unobservable factors affecting pollutant concentrations [18]. The EM algorithm is composed of two main steps:

- *Expectation step* (E-step): This step estimates missing or latent data based on available observations and current model estimates. In air pollution forecasting, this might involve predicting the likelihood that an observation belongs to a specific pollutant cluster, based on existing data and model assumptions [19].
- *Maximization step* (M-step): After filling in the missing data, the model parameters (such as means, variances, and weights of the Gaussian distributions) are updated to maximize the likelihood of the observed data, improving the model's fit. This step refines the estimates of pollutant concentration patterns over time, enhancing the accuracy of the model's predictions [20].

These steps are repeated iteratively until the algorithm converges to optimal parameter values, refining the model's predictions and making it well-suited for complex, incomplete environmental datasets.

By combining GMM with the EM algorithm, this study improves the accuracy and reliability of air pollution forecasts, capturing the underlying distribution of pollutant concentrations. These methods enhance environmental monitoring, decision-making, and public health strategies.

Decision tree

Decision trees (DT) are a widely-used ML technique that excels in both classification and regression tasks. They are particularly valuable for analyzing complex datasets where the goal is to partition the data into distinct categories or predict numerical outcomes based on input features. The main strength of DT lies in their ability to model relationships between variables in a hierarchical structure, making them intuitive and easy to interpret. The process of building a decision tree involves recursively partitioning the dataset into subsets based on the values of different features, thereby creating a tree-like model [21, 22]. Each internal node of the tree represents a decision based on a feature, while the leaf nodes represent the outcomes or predictions. These trees are effective in capturing both linear and non-linear relationships within the data.

Structure of a decision tree

- *Root node*: The root node is the starting point of the decision tree. It contains the entire dataset and represents the first feature split that divides the data into subsets. This initial division is based on a feature that best separates the data, according to some criterion, such as information gain or Gini impurity [23]. The root node is critical because it lays the foundation for the entire structure of the tree.
- *Decision node*s: After the root node, each subsequent node is called a decision node. These nodes represent a decision based on one of the input features. At each decision node, the dataset is split into subsets according to the feature values, aiming to maximize the distinction between the resulting groups. The feature chosen for each split is selected in such a way that the resulting subsets are as pure as possible, meaning they consist mostly of data points that belong to the same class or have similar outcomes in the case of regression [24].
- *Leaf nodes*: The leaf nodes are the terminal points of the decision tree. These nodes do not lead to further splits but instead represent the final predicted outcome. In classification tasks, the leaf node typically contains the most frequent class label, while in regression, it contains the predicted value for the outcome. The accuracy and effectiveness of a decision tree depend largely on how well the leaf nodes reflect the true distribution of the data in their respective partitions [25].
- *Internal nodes*: Internal nodes are any nodes in the tree that perform a split based on feature values. These nodes are neither the root nor the leaf nodes but play a vital role in the tree's decision-making process. Each internal node divides the data into smaller subsets, further refining the predictions made by the tree. The deeper the internal nodes are in the tree, the more specific the predictions they make are, but this can also lead to overfitting if the tree becomes too complex. Effective pruning techniques are often used to simplify the tree and enhance its generalization capability [26].

Model evaluation and validation

Both models, GMM and DTM, are evaluated and validated using cross-validation techniques to assess their performance and ensure the robustness of the predictions. The models are trained on a portion of the data and tested on another portion to evaluate their accuracy, generalization ability, and resilience to overfitting. Several evaluation metrics are used, including MAE, RMSE, and R^2, to measure the models' predictive accuracy. In addition, the performance of the models is compared using various information criteria, such as the AIC and the BIC, which provide insights into the models' complexity and goodness-of-fit.

In the next section, we present the numerical results and discussion of the air pollutant levels (NO_x, NO_2, NO) based on the ML models employed in this study.

Numerical results

The time series data for NO_x, NO_2, and NO, which significantly impact air pollution, covers the period from January 1, 2017, to December 31, 2017, fig. 1. The data is sourced from the national center of meteorology, Kingdom of Saudi Arabia website. Figure 1 illustrates the fluctuations in air pollution levels during this period, highlighting the marked changes and trends, particularly during the months of November and December, which experience the highest levels of pollution. These months are notably affected by seasonal weather patterns. For instance, during temperature inversions in the winter, smog can become trapped near the ground, leading to elevated pollution levels. Additionally, the speed and direction of the wind during these months can hinder the dispersion of pollutants, exacerbating the accumulation of harmful emissions.

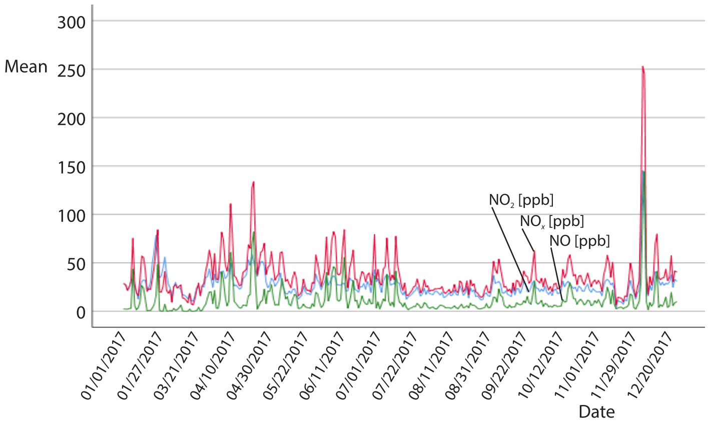

Figure 1. Pollution level data for the city of Jeddah

Table 1 presents the minimum and maximum values, along with the mean and standard deviation, for the air pollutants NO_x, NO_2, and NO. In metropolitan areas, the average annual concentrations of NO_2 typically range between 20 μg/m³ and 50 μg/m³. Similarly, NO concentrations in urban environments generally fluctuate between 10 ppb and 50 ppb, with variations depending on traffic density and industrial activities. The concentration of NO_x in urban regions can range from 50 ppb to 100 ppb or higher, influenced by various pollution sources, including vehicle emissions and industrial processes.

Table 1. Descriptive statistics for air pollutant levels

Variable	Observations	Minimum	Maximum	Mean	Standard deviation
NO_x	312	7.000	253.000	37.180	25.231
NO_2	312	4.200	145.300	26.701	12.950
NO	312	0.000	144.000	11.376	14.717

Decision tree models results

The DT are flexible and intuitive tools for making predictions. Starting at the root node, decisions are made at each subsequent node based on feature values, allowing for an easy path to a forecast.

When building a ML model, it is crucial to split the data into training and test sets. This ensures that the model is evaluated on data it has not encountered during training, which helps assess its ability to generalize to new, unseen data. This approach prevents the model from merely memorizing the training set and ensures that it can handle real-world data effectively.

As shown in fig. 2, the distribution of the training and test datasets is: for NO_2, 171 samples (54.81%) were used for the training set, and 141 samples (45.19%) were used for the test set, totaling 312 samples; for NO, 159 samples (50.96%) were used for training, and 153 samples (49.03%) for testing, with 312 samples in total; for NO_x, 145 samples (46.37%) were allocated to

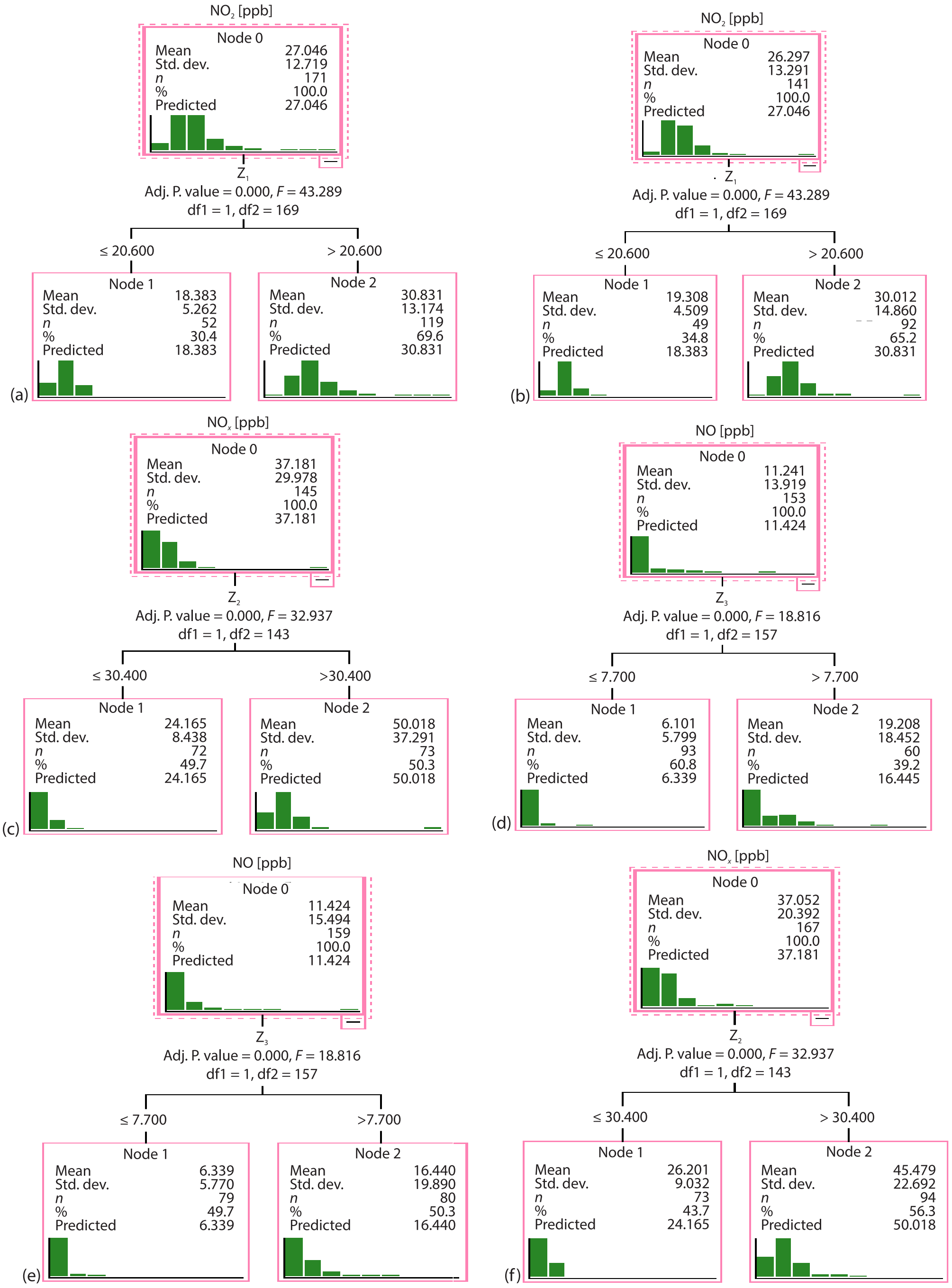

Figure 2. Training and testing sample sets for air pollutant levels using the DT diagram: (a), (b) correspond to the (training, testing) sets for NO_2, (c), (d) correspond to the (training, testing) sets for NO, and (e), (f) correspond to the (training, testing) sets for NO_x

the training set, and 167 samples (53.53%) to the test set, out of 312 samples. These panels also display the mean, standard deviation, and predicted values at nodes zero, one, and two.

Gaussian mixture model results

The GMM consists of multiple Gaussian distributions, each characterized by its own mean and variance. It is used to estimate the probability density of a set of data points and to cluster them into groups that are likely to have originated from different Gaussian distributions. The model assigns each data point to the distribution it is most likely to belong to. To fit a GMM to the data, the EM algorithm is employed. This iterative method identifies the parameters that maximize the likelihood of the data under the model.

The data presented in tab. 2 shows the proportions, mean, and variability for each class of the pollutant (NO).

Table 2. The proportions, the mean, and the variance by class (NO)

Class	1	2	3	4	5
Proportions	0.021	0.377	0.148	0.374	0.080
Mean	79.510	7.977	16.804	3.649	35.667
Variance	79.510	7.977	16.804	3.649	35.667

Figure 3 illustrates the most probable values for each class based on the NO pollutant data. It provides a visual comparison between the observed data and the model's predictions, allowing for an assessment of the models' accuracy and reliability in capturing the patterns and fluctuations of pollutant levels. The cumulative distribution functions (CDF) for NO are also presented, highlighting the probability distribution and variability over time, and demonstrating a strong alignment between the predicted values and the empirical data.

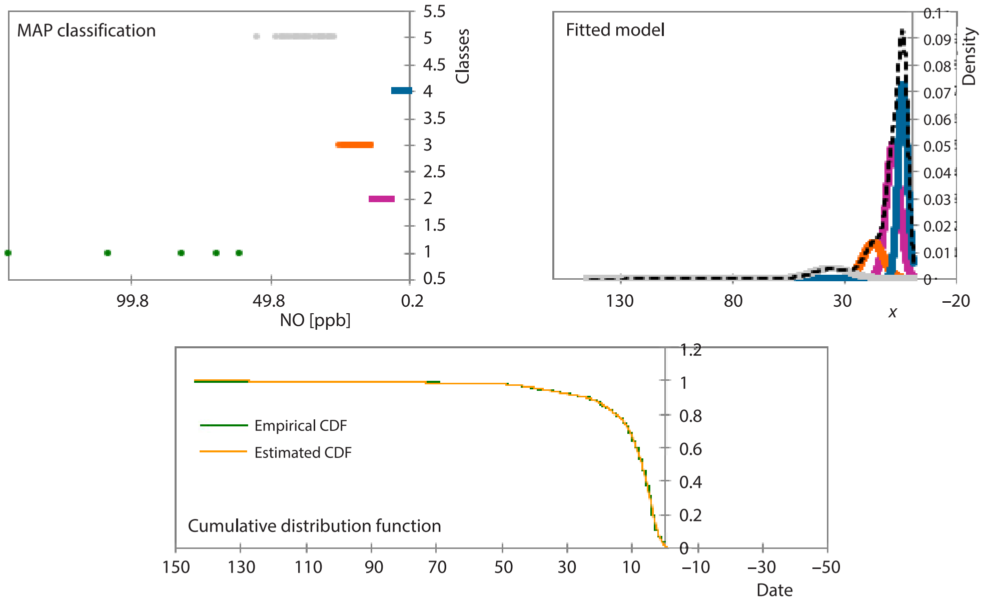

Figure 3. The MAP classification, fitted model, and CDF of NO

The data shown in tab. 3 presents the proportions, mean, and variation for each class of the NO_x pollutant.

Figure 4 presents the most probable values associated with each class based on the NO_x pollutant data. The second panel shows the alignment of the observed data with the model predictions for NO_x, allowing for a clear visual comparison between the actual and predicted values. Additionally, the third panel displays the CDF for NO_x, which depict the probability distribution and variability over time. These CDF highlight the strong agreement between the predicted and empirical data, enabling an evaluation of the models' accuracy and reliability in capturing the patterns and fluctuations in pollutant levels.

Table 3. The Proportions, the mean, and the variance by class (NO_x)

Class	1	2	3	4	5
Proportions	0.414	0.408	0.112	0.043	0.024
Mean	23.651	34.894	57.418	78.225	144.252
Variance	49.387	68.585	76.981	13.578	4822.066

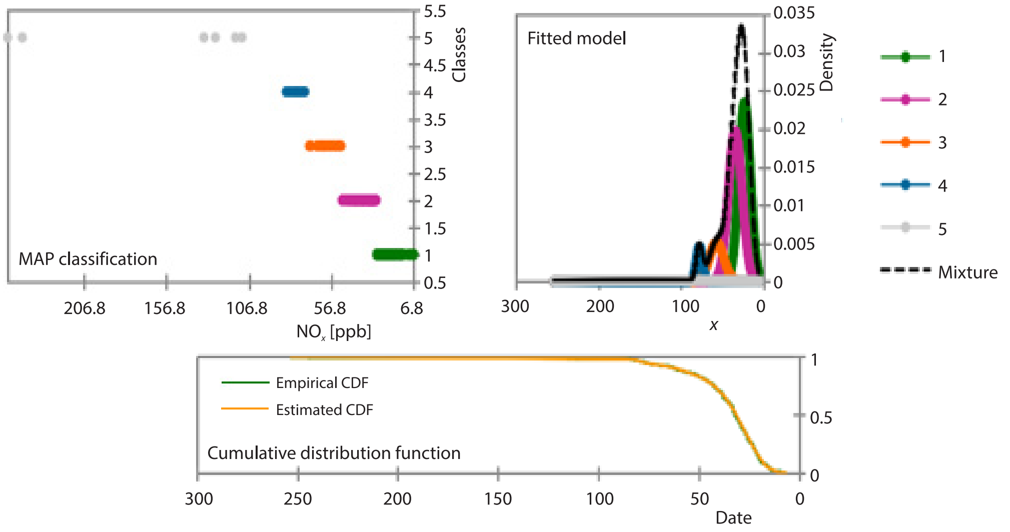

Figure 4. The MAP classification, fitted model, and CDF for NO_x

The data shown in tab. 4 presents the proportions, mean, and variation for each class of the NO_2 pollutant.

Table 4. The Proportions, the mean, and the variance by class (NO_2)

Class	1	2	3	4	5
Proportions	0.179	0.405	0.364	0.035	0.017
Mean	18.655	23.707	28.619	52.153	88.370
Variance	13.739	44.750	60.292	22.849	1159.919

The data presented in tab. 4 provides the PMV for each class of the of pollutant (NO_2).

Based on the data for the NO_2 pollutant, the first panel of fig. 5 shows the most probable values associated with each class. The second panel displays the fitted model, aligning the data with the model's predictions for NO_2, allowing for a visual comparison between the observed data and the model's forecasts. This comparison provides insight into the accuracy and reliability of the models in capturing the patterns and fluctuations in pollutant levels. The third panel presents the CDF for NO_2, which illustrate the probability distribution and variability over time, while also demonstrating a strong correlation between the predicted and empirical data.

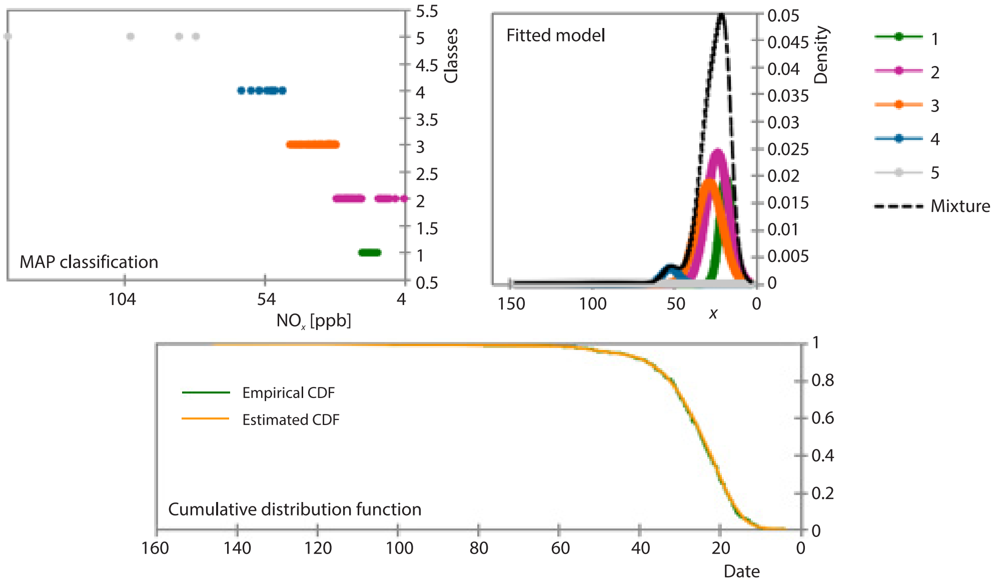

Figure 5. The MAP Classification, fitted model, and CDF for NO_2

Air pollution levels from January 1 to December 1, 2017, reveal trends, with the worst months being November and December, likely influenced by weather changes. For instance, smog can accumulate during winter temperature inversions, and low wind speeds can hinder pollution dispersion, fig. 1. Table 1 shows the range, mean, and standard deviation for the pollutants NO_x, NO_2, and NO, while tabs. 2-4 provide percentages, means, and variations for each pollutant category. Figures 3, 4, and 5 visually present fitted models and CDF for NO_x, NO_2, and NO, highlighting the maximum a posteriori (MAP) classification for each class. These CDF help understand the probability distribution of air pollution over time, showing a strong match between the predicted and observed data.

Table 5 outlines the selection criteria for evaluating the models, using BIC, AIC, ICL, log-likelihood, NEC, and entropy. Models with lower BIC, AIC, and ICL values indicate a better fit, while lower entropy and NEC values reflect more reliable predictions. The NO_x show the best model fit, having the lowest BIC, AIC, and ICL values. According to the ICL criteria, GMM models outperform DTM, with results of –117661634.4 for NO, –117661634.4 for NO_2, and –142259279.6 for NO_x.

Table 5. The BIC, AIC, ICL, Log-likelihood, NEC, and Entropy criteria

	Entropy	NEC	ICL	Log-likelihood	AIC	BIC
NO_x	7074687.67	1.079	–125522893.75	–55686644.33	–111373316.66	–111373518.40
NO_2	11659304.65	2.582	–120907361.72	–48794261.34	–97588550.67	–97588752.42
NO	6701482.57	0.683	–103819089.16	–45207947.13	–90415922.27	–90416124.01

Conclusion

In recent years, air pollution has become a critical global issue, garnering increasing attention from policymakers, environmentalists, and various stakeholders. In response to this growing concern, this study has focused on leveraging ML techniques to predict air pollution levels, aiming to enhance our understanding and management of this environmental challenge. The findings of this research underscore the complex and dynamic nature of air pollution, while also demonstrating the effectiveness of advanced ML models in accurately forecasting pollutant concentrations. Specifically, the results reveal that the GMM outperforms the DTM in capturing the underlying patterns of air pollution data. The lower ICL values observed in the GMM models indicate a better fit to the data, showcasing their superior efficiency compared to DTM. In particular, GMM models have provided more consistent results with reduced uncertainty, establishing them as a reliable tool for air quality forecasting. As air pollution continues to pose significant environmental and public health risks, the application of these predictive models is crucial for informed decision-making and the development of effective pollution control strategies, ultimately contributing to a healthier and more sustainable future.

Acknowledgment

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] Patil, R. M., *et al.*, A Literature Review on Prediction of Air Quality Index and Forecasting Ambient Air Pollutants Using Machine Learning Algorithms, *Int. J. Innov. Sci. Res. Technol.*, *5* (2020), 8, pp. 1148-52
[2] Maltare, N. N., Vahora, S., Air Quality Index Prediction Using Machine Learning for Ahmedabad City, *Digital Chemical Engineering*, *7* (2023), 100093
[3] Sharma, N., *et al.*, Forecasting Air Pollution Load in Delhi Using Data Analysis Tools, *Procedia Computer Science*, *132* (2018), Jan., pp.1077-1085
[4] Dun, M., *et al.*, Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine, *Mathematical Problems in Engineering*, *1* (2020), 8914501
[5] Samad, A., *et al.*, Air Pollution Prediction Using Machine Learning Techniques – An Approach to Replace Existing Monitoring Stations with Virtual Monitoring Stations, *Atmospheric Environment*, *310* (2023), 119987
[6] Mani, G., Viswanadhapalli, J. K., Prediction and Forecasting of Air Quality Index in Chennai Using Regression and ARIMA Time Series Models, *Journal of Engineering Research*, *10* (2022), 2A, pp. 179-94
[7] Liu, H., et al., Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms, *Applied Sciences*, *9* (2019), 4069
[8] Idroes, G. M., *et al.*, Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring, *Leuser Journal of Environmental Studies*, *6* (2023), 2, pp. 62-68
[9] Noviandy, T. R., *et al.*, Ensemble Machine Learning Approach for Quantitative Structure Activity Relationship-Based Drug Discovery: A Review, *Infolitika Journal of Data Science*, *25* (2023), 1, pp. 32-41
[10] Almaliki, A., Abdessamed, Derdour, A., Earning Methods, *Sustainability*, *15* (2023), 13168
[11] Menezes, F., Popowicz, G. M., Acid Rain and Flue Gas: Quantum Chemical Hydrolysis of NO_2, *Chem. Phys. Chem.*, *23* (2022), 202200395
[12] Ju, T., *et al.*, A New Prediction Method of Industrial Atmospheric Pollutant Emission Intensity Based on Pollutant Emission Standard Quantification, *Frontiers of Environmental Science and Engineering*, *17* (2023), 8
[13] Gomez, D., *et al.*, A New Approach to Monitor Water Quality in the Menor Sea (Spain) Using Satellite Data and Machine Learning Methods, *Environmental Pollution*, *286* (2021), 117489
[14] Xia, H., *et al.*, Dioxin Emission Prediction Based on Improved Deep Forest Regression for Municipal Solid waste Incineration Process, *Chemosphere*, *294* (2022), 133716
[15] Mueller, N., *et al.*, Health Impact Assessments of Shipping and Port-Sourced Air Pollution on a Global Scale: A Scoping Literature Review, *Environmental Research*, *216* (2023), 114460

[16] Muruganandam, N. S., Arumugam, U., Dynamic Ensemble Multivariate Time Series Forecasting Model for PM2. 5, *Computer Systems Science and Engineering*, *44* (2023), 979

[17] Saraiva, E. F., *et al*., An Integrated Approach for Making Inference on the Number of Clusters in a Mixture Model, *Entropy*, *21* (2019), 1063

[18] Carlsson, K. C., *et al*., Modelling Subpopulations with the $ MIXTURE Subroutine in NON-MEM: Finding the Individual Probability of Belonging to a Subpopulation for the Use in Model Analysis and Improved Decision Making, *The AAPS Journal*, *11* (2009), 1, pp. 148-154

[19] Arshad, U., *et al.*, Development of Visual Predictive Checks Accounting for Multimodal Parameter Distributions in Mixture Models, *Journal of Pharmacokinetics and Pharmacodynamic*s, *46* (2019), Apr., pp. 241-250

[20] Pallathadka, H., *et al*., Classification and Prediction of Student Performance Data Using Various Machine Learning Algorithms, *Materials Today: Proceedings*, *80* (2023), Part 3, pp. 3782-3785

[21] Hamdi, M., *et al*., Forecasting and Classification of New Cases of COVID 19 Before Vaccination Using Decision Trees and Gaussian Mixture Model, *Alexandria Engineering Journal*, *62* (2023), Jan., pp. 327-333

[22] Ranka, S., Singh, V., CLOUDS: A Decision Tree Classifier for Large Datasets, *Proceedings*, 4th Knowledge Discovery and Data Mining Conference, Manchester, UK, 1998, Vol. 2. pp. 2-8

[23] Zhao, L., *et al*., Decision Tree Application Classification Problems with Boosting Algorithm, Electronics, *10* (2021), 1903

[24] Sun, R., *et al*., A Gradient Boosting Decision Tree-Based GPS Signal Reception Classification Algorithm, *Appl. Soft Comput*., *86* (2020), 105942

[25] Cheng, K. C., *et al.*, Establishing a Multiple-Criteria Decision-making Model for Stock Investment Decisions Using Data Mining Techniques, *Sustainability*, *13* (2021), 3100

[26] Li, Y., Predicting Materials Properties and Behavior Using Classification and Regression Trees, *Materials Science and Engineering A*, *433* (2006), 1-2, pp. 261-268

Paper submitted: June 24, 2024
Paper revised: September 10, 2024
Paper accepted: October 12, 2024