Jurnal Kejuruteraan 37(6) 2025: 2851-2859
https://doi.org/10.17576/jkukm-2025-37(6)-23

K-Nearest Neighbors for Predicting Ozone Concentrations: A Machine Learning Approach for Air Quality Assessment

Emi Amielda Ahmad Mokhtar[a*] , Nuryazmin Ahmat Zainuri[b], Muhamad Alias Md Jedi[c] & Nur Arzilah Ismail[b]

[a]*Department of Engineering Education,Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*

[b]*Department of Engineering Education, Engineering Mathematics Research Group, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*

[c]*Department of Mechanical and Manufacturing Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*

**Corresponding author: P134103@siswa.ukm.edu.my*

Received 1 November 2024, Received in revised form 10 June 2025

Accepted 10 July 2025, Available online 30 September 2025

ABSTRACT

Ozone (O_3) is a significant air contaminant that poses severe health risks, particularly in urban areas. Accurate prediction of the ozone concentration level is crucial for increasing public consciousness and giving important data to governments for public health alerts and air quality management. This study explores the application of machine learning techniques, the K-Nearest Neighbors (KNN) method, for predicting ozone (O_3) concentrations based on meteorological variables collected from three monitoring stations in the Klang Valley region. The research involves data preprocessing procedure that includes handling missing values through imputation and applying the KNN algorithm to predict ozone concentrations. The model was trained and tested using cross-validation and its performance was assessed using evaluation metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The KNN model achieved the most accurate predictions at Petaling Jaya station with an MAE of 0.00350 and RMSE of 0.00447, followed by Cheras station (MAE: 0.00402, RMSE: 0.00520) and Batu Muda station (MAE: 0.00406, RMSE: 0.00527). These results indicate that the KNN model achieved relatively low prediction error in this study, demonstrating its potential reliability in forecasting ozone concentrations. These findings suggest that the KNN model can be applied to predict ozone concentrations across different urban locations, offering valuable insights for public health planning and supporting strategies for mitigating ozone pollution.

Keywords: Ozone concentration; machine learning; K-Nearest Neighbors (KNN); air quality prediction

INTRODUCTION

Air pollution refers to the existence of harmful pollutants in the atmosphere, with tropospheric ozone being particularly concerning due to its negative impacts on human health and ecosystems. In Malaysia, ozone concentrations have been rising, especially in urban and industrial areas. Research by Ahamad et al. (2020) identified significant ozone increases in Peninsular Malaysia, driven by urbanization and industrial activities, while Mohd Napi et al. (2020) noted diurnal variations in industrial regions, with ozone concentrations peaking at midday. Other studies, such as Baidrulhisham et al. (2022), linked ozone formation to meteorological factors like temperature and nitrogen oxide levels.

At the ground level, ozone is created by photochemical processes involving precursors like nitrogen oxides (NOx) and volatile organic compounds (VOCs) in the presence of sunlight rather than being released directly. Many respiratory problems, including asthma and chronic bronchitis, have been related to this secondary pollutant. The World Health Organization (WHO) reports that prolonged exposure to high ozone concentrations is a major cause of preventable deaths globally. Being a tropical nation, Malaysia is subject to strong sun radiation, which speeds up the creation of ozone, especially in crowded urban areas like the Klang Valley. This region is a priority focus for air quality prediction research due to its fast rate of urbanization, elevated vehicle emissions and industrial activity.

Ozone concentration prediction has been a significant research topic due to its implications for public health and environmental management. Traditional statistical models, including Multiple Linear Regression (MLR), have been commonly used to forecast ozone concentrations. However, the nonlinear correlation between ozone and its precursors, such as nitrogen oxides, volatile organic compounds and meteorological variables, limits the effectiveness of linear models (Jumin et al. 2020). Moreover, studies by Bekesiene et al. (2021) conclude that the regression modelling approach was not highly effective since the relationship between weather factors and ozone concentrations is more complex than a simple straight line and lacks capacity. Similarly, Su et al. (2020) demonstrated that stepwise regression underperforms when dealing with complex, nonlinear and temporally variable data such as ozone concentrations.

In recent years, many researchers have increasingly turned to a machine learning approach to enhance the precision of ozone concentration prediction. The machine learning algorithms, which are better suited to capturing these complex relationships, have gained traction. Several studies have applied models like Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN) and Random Forests (RF) to predict ozone concentrations, with promising results (Li & Ren 2022; Alomar et al. 2020; Tanaskuli et al. 2020; Lu et al. 2021; de Oliveira et al. 2021) due to their ability to handle complex, nonlinear relationships in meteorological data.

The KNN algorithm, a non-parametric and straightforward machine learning technique, has been widely applied in various environmental studies due to its simplicity and effectiveness in handling complex datasets. Its adaptability allows it to be used across different areas, such as air quality monitoring, climate modeling, and ecological data analysis (Marchang & Tripathi 2021). In particular, the KNN algorithm has shown promise for handling non-linear relationships, especially in environmental datasets (Evitania 2023; Sarkar et al. 2021). Unlike regression-based models, KNN does not rely on predefined assumptions about the relationship between variables, making it better suited for datasets with complex interactions, such as those involved in ozone prediction.

Despite this progress, K-Nearest Neighbors (KNN), a non-parametric algorithm, has been relatively underexplored in ozone prediction, especially in Malaysia. KNN's simplicity and ability to model nonlinear relationships make it particularly suitable for environmental datasets, where patterns are often irregular (Jeyabalan et al. 2022; Sarkar et al. 2021). Previous studies have shown KNN to be effective in air quality forecasting (Juarez & Petersen 2023), demonstrating its potential to outperform traditional models.

This study aims to predict daily ozone concentrations using the KNN algorithm, employing meteorological data focusing on three monitoring stations in Klang Valley region, which include the Batu Muda, Cheras and Petaling Jaya stations. The framework illustrates in Figure 1 involves preprocessing the dataset, applying the KNN algorithm and validating its accuracy using cross-validation. This framework provides a systematic structure for understanding how environmental factors influence ozone concentrations, beginning with data collection and data preprocessing, including missing value imputation and normalization.

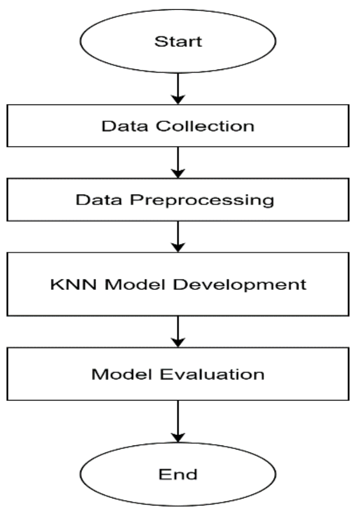


FIGURE 1. Framework for KNN Ozone Prediction

By evaluating the model's performance using cross-validation and metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), this research contributes to the ongoing effort to improve air quality predictions using machine learning techniques. This study presents a model that environmental authorities could adapt for better ozone forecasting in addition to its technical achievements. KNN is suitable to more complicated models in facilitating data-driven environmental decision-making because of its transparency and simplicity, which make it appropriate for applications that demand interpretability.

METHODOLOGY

The data for this study was obtained from three monitoring stations in the Klang Valley region, specifically located in Petaling Jaya, Batu Muda and Cheras stations. These locations are situated in densely populated areas with significant industries, making them ideal for studying ozone concentration in an urban environment. The monitoring stations were chosen based on their documented exceedances of ozone quality standards in 2022, as illustrated in Figure 2 from the Environmental Quality Report 2022.

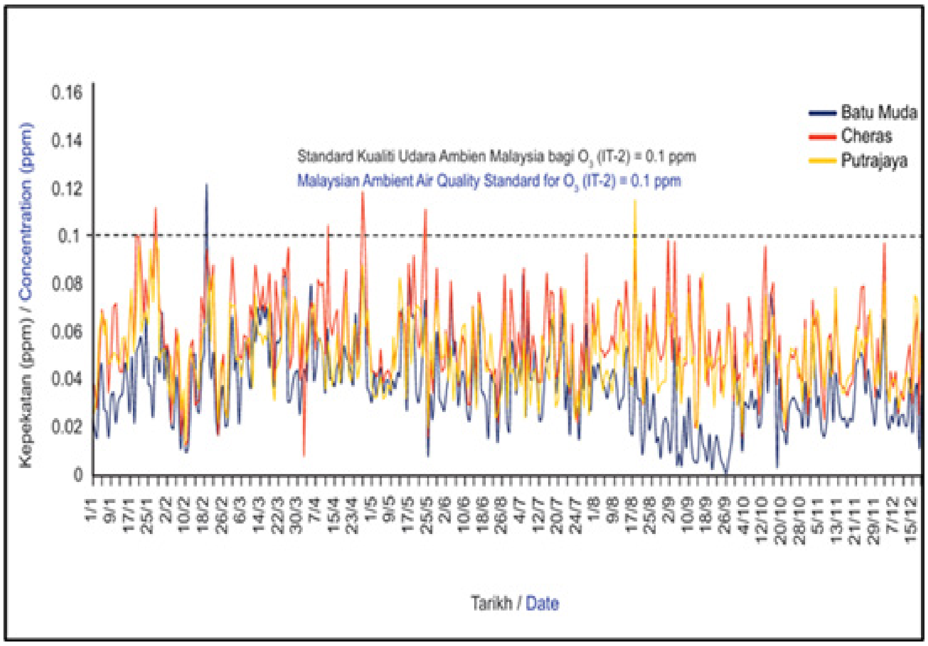


FIGURE 2. Trend of Daily Maximum 1-hour Concentration of Ozone (O_3), Klang Valley 2022 (Environmental Quality Report 2022)

The dataset was sourced from the Department of Environment Malaysia from the period of 2017 to 2022. The data consists of daily measurements of ozone concentrations (O_3), and meteorological variables such as Particulate Matter ($PM_{2.5}$ and PM_{10}), Sulphur Dioxide(SO_2), Nitrogen Dioxide (NO_2), Nitric Oxide (NO), Carbon Monoxide (CO), wind speed, wind direction, humidity levels and solar radiation. Table 1 exhibits a summary of the variables.

TABLE 1. Summary of variables and units

Variables	Unit
Ozone (O_3)	ppm
Particulate Matter ($PM_{2.5}$ and PM_{10})	$\frac{\mu g}{m^2}$
Sulphur Dioxide(SO_2)	ppm
Nitrogen Dioxide (NO_2)	ppm
Nitric Oxide (NO)	ppm
Carbon Monoxide (CO)	ppm
Wind speed	m/s
Wind direction	(°)
Humidity levels	%
Solar radiation	$\frac{W}{m^2}$

In addition to data cleaning, the study included basic preprocessing steps such as ensuring data completeness and making sure consistent variable units across all records. Time series consistency was maintained by aligning dates and ensuring that no days were missing across all stations. Missing values were imputed through KNN imputation, while all other variables were verified for proper data entry. These steps are important to prepare the dataset for meaningful training and testing phases in the predictive modeling process.

The model was constructed using original meteorological variables without employing feature engineering techniques such as deriving lagged values or derivation of indicators. This decision was made to maintain model interpretability and ensure alignment with the actual observed variables. The selected variables such as ozone, nitrogen dioxide, carbon monoxide, temperature and wind variables were selected based on their documented influence on ozone formation in prior research. This approach helps preserve data authenticity and prevents overfitting due to synthetic transformations.

Ozone (O_3) concentration at ground level is influenced by various environmental factors, including temperature, humidity, solar radiation, wind speed, and co-pollutants such as NO_2 and CO (Xiong et al. 2024). These variables often interact in nonlinear and complex ways, which makes traditional linear modeling approaches less effective. To address this, the study employs the K-Nearest Neighbors (KNN) algorithm, a non-parametric machine learning method grounded in instance-based learning theory. For instance, a study by Evitania (2023) implemented the KNN algorithm using RapidMiner to predict air pollution levels, achieving an accuracy of 93.94%. This study highlights the effectiveness in handling complex environmental data, including ozone concentrations.

This approach is suitable for ozone prediction because it can handle complex relationships without needing to define them explicitly. Thus, the theoretical foundation combines environmental science principles with machine learning theory to support the use of KNN in modeling ozone concentration patterns. To ensure data quality, the missing values were imputed using K-Nearest Neighbors (KNN) imputation. This imputation was chosen due to its ability to maintain the dataset's original distribution and avoid bias that might arise from simpler methods like mean imputation. Saeipourdizaj et al. (2021), Sarbakhsh and Gholampour (2021) applied the KNN method to impute missing values in PM_{10} and O_3 data, demonstrating its effectiveness alongside interpolation and moving average methods. KNN imputation does not assume a specific distribution of the data, making it robust for complex datasets. The K-Nearest Neighbors of a data point with a missing value are calculated using the Euclidean distance metric as follows.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where $d(x,y)$ is the distance of Euclidean between two datasets, x_i and y_i are the values for points respectively. After the K Nearest Neighbors are identified, the missing values are computed. The KNN method is given as follows:

$$\hat{y}_{\text{missing}} = \frac{1}{K}\sum_{i=1}^{K} y_i \quad (2)$$

where $\hat{y}_{\text{missing}}$ is the imputed value for the missing observation, y_i are the values of the i-th nearest neighbors, and k is the number of nearest neighbors.

The KNN predictive algorithm for estimating ozone concentration follows these steps: (1) For each test observation, calculate the Euclidean distance between the test data point and all training data points using relevant meteorological variables; (2) Identify the K Nearest Neighbors based on the smallest distances; (3) Retrieve the ozone values corresponding to these neighbors; (4) Compute the predicted ozone concentration by averaging the ozone values of the k neighbors. This method leverages historical patterns in the training data to generate reliable predictions without relying on a predefined model structure.

The appropriate number of neighbors, k, was determined through hyperparameter tuning. Various k values ranging from 1 to 20 were tested, and the optimal k for each station was selected based on the lowest Root Mean Squared Error (RMSE) obtained during cross-validation.

To ensure accurate distance calculations in the KNN algorithm, min-max normalization is applied to the dataset, as it includes variables with different scales. This normalization is important to ensure that no feature has a disproportionate influence on the model. Min-max scaling transforms all features within the range [0,1]. The transformation was performed using the following formula:

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \quad (3)$$

where y is the original data, $y_{\min}$ and $y_{\max}$ are the minimum and the maximum values of the dataset respectively and y' is the normalized value.

The model is developed based on 70% portioned data. Next, the model was tested using the remaining 30%. This approach is commonly used in machine learning and statistical modeling to ensure that the model is trained on a substantial portion of the data while still being validated on a separate, unseen dataset to assess its predictive accuracy. To further enhance model's ability to generalize and reduce the risk of overfitting, this partitioning method helps evaluate how well the model performs on new, unseen data. This procedure is particularly critical for environmental datasets, which often demonstrates seasonal fluctuations, inconsistent patterns, and external factors such as regional weather variability or emissions from nearby sources. Next, model performance was assessed using performance indicators: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics assess the model's accuracy by quantifying the average magnitude of errors between predicted and actual values. These metrics are frequently used in regression models to evaluate predictive accuracy (Wu et al. 2021; Yafouz et al. 2022). The formula for RMSE and MAE is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of data points. The RMSE measures the standard deviation of the variations between actual and predicted values. It demonstrates how close the predictions are from the actual data, particularly along the line of best fit. A lower RMSE indicates a more accurate model. Larger deviations or outliers have a significant impact on RMSE since it squares the errors before averaging. While MAE measures the average size of all errors between predictions and observational data. It provides a simple measure of precision by averaging absolute differences. Unlike RMSE, MAE is less influenced by big errors, giving it a more reliable indicator for assessing overall performance. Lower values of both RMSE and MAE show better model performance.

RESULTS AND DISCUSSION

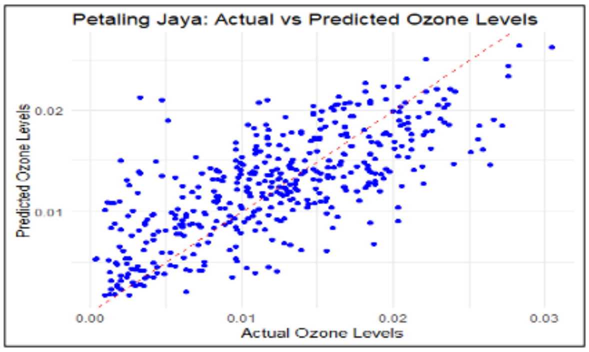
The performance of the KNN model was evaluated using the meteorological variables taken from three air quality monitoring stations: Batu Muda, Cheras and Petaling Jaya. The evaluation metrics used to assess the model's performance are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Batu Muda station achieved its best performance with *k*=13, resulting in an RMSE of 0.00527 and an MAE of 0.00406. Cheras station showed better performance compared to Batu Muda, with *k*=11, yielding an RMSE of 0.00520 and MAE of 0.00402. Among the three monitoring stations, the Petaling Jaya station showed the best performance of the KNN model with the lowest value of RMSE that is 0.00447 and MAE of 0.0035 with *k*=5. The accuracy observed at Petaling Jaya suggests that localized meteorological or environmental factors, potentially characterized by greater consistency, might have played a role in enhancing the KNN prediction performance. *k* values in the KNN demonstrate the number of neighbors which is nearest to be considered in performing a prediction. At Petaling Jaya station, a smaller *k*=5 was optimal, likely because smaller nearest neighbors were sufficient to capture the relatively consistent meteorological patterns. Different from Batu Muda station, a bigger *k* value, that is *k*=13 showed the best performance, indicating that averaging over a broader set of neighbors helped mitigate the effects of higher variability in local environmental surroundings. The summary of the evaluation metrics for each monitoring station is presented in Table 2 below.

TABLE 2. The performance metrics for each monitoring station

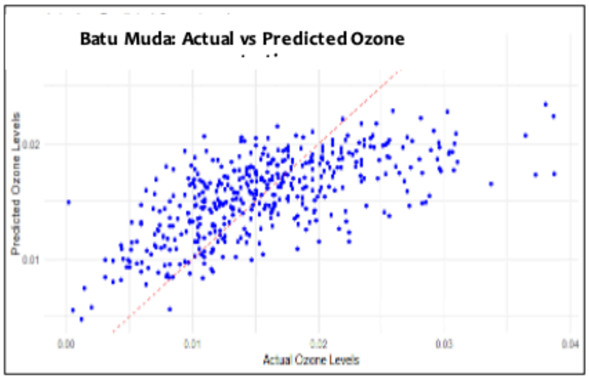
Monitoring station	RMSE	MAE	*k*
Batu Muda	0.00527	0.00406	13
Cheras	0.00520	0.00402	11
Petaling Jaya	0.00447	0.00350	5

The differing results of the KNN model across stations demonstrates the importance of local meteorological and environmental conditions. Surrounding land use, traffic density, industrial presence, and even topographical features can all have an impact on pollutant dispersion and ozone formation. For example, greater consistency in Petaling Jaya's weather and emissions could lead to more predictable ozone patterns, improving model performance. Meanwhile, greater variability or a combination of pollution sources in Batu Muda may explain why a higher *k*-value is required to stabilise predictions. These findings suggest that local context is important for model effectiveness and should be considered in future predictive studies.

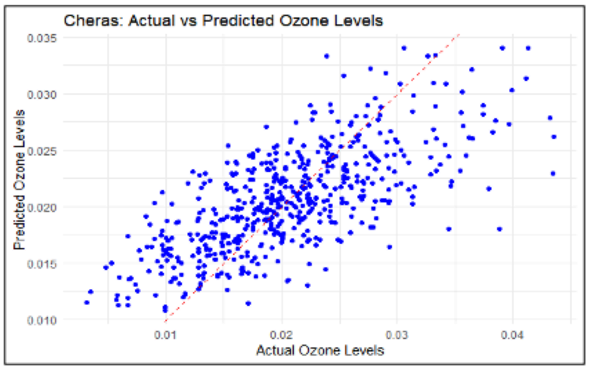
Figure 3 displays the relationship between actual and predicted ozone concentrations for each station. Scatter plots comparing actual and predicted ozone concentrations show that the model performs well for lower ozone concentrations, with predictions aligning closely to the 45-degree line. However, as actual ozone concentrations increase, the predicted data become more variable. In Cheras and Petaling Jaya monitoring stations, the model tends to underpredict higher ozone concentrations while maintaining relatively good accuracy for lower values. At the Batu Muda monitoring station, a wider spread of points indicates more significant prediction errors, particularly for medium and high ozone concentrations. Overall, the model demonstrates moderate accuracy but requires improvements for higher concentrations and regional variability.



(a) Petaling Jaya monitoring station

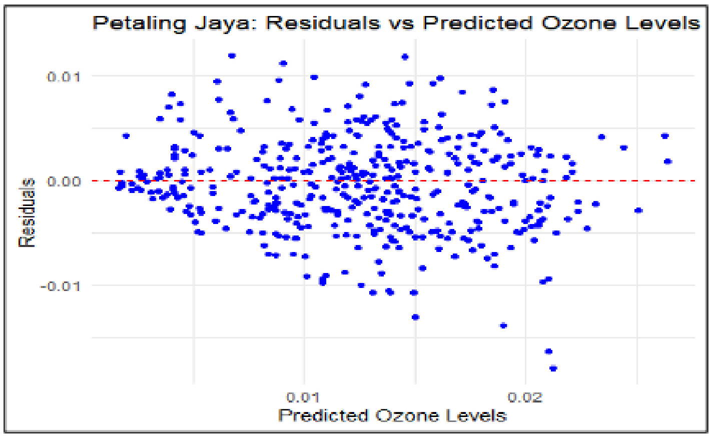


(b) Batu Muda monitoring station

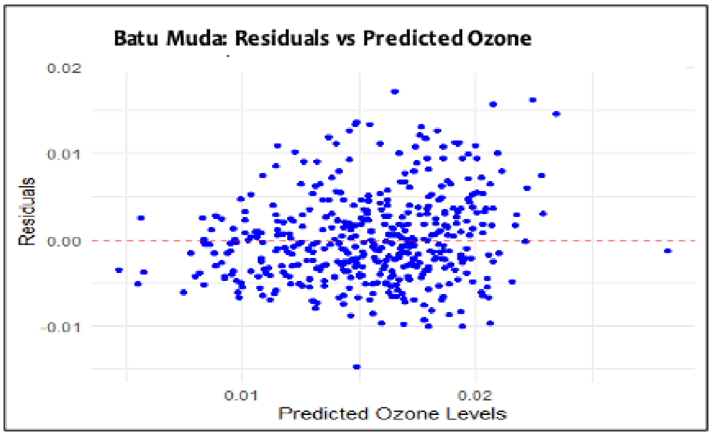


(c) Cheras monitoring station

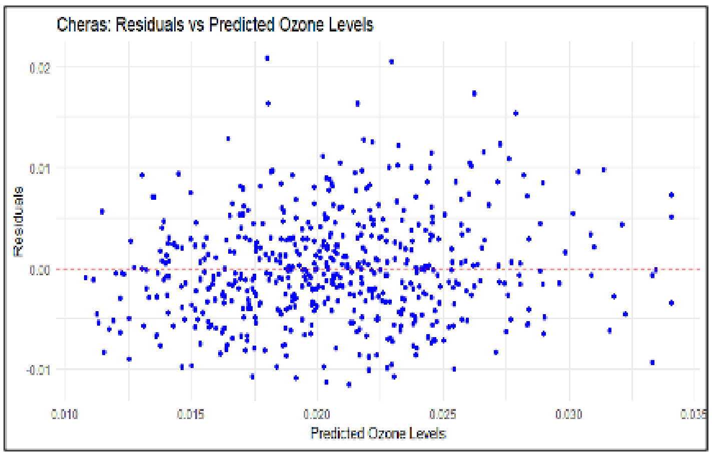
FIGURE 3. Relationship between actual and predicted ozone concentrations for each station



(a) Petaling Jaya monitoring station



(b) Batu Muda monitoring station



(c) Cheras monitoring station

FIGURE 4. Residuals vs. predicted ozone concentrations for each station

CONCLUSION

This study demonstrates the effectiveness of the *K*-Nearest Neighbors (KNN) algorithm in predicting ozone concentrations using meteorological data from three monitoring stations in the Klang Valley region of Malaysia. By applying KNN and evaluating the model using cross-validation, the optimal *k* values for each station were identified: *k*=13 for Batu Muda, *k*=11 for Cheras, and *k*=5 for Petaling Jaya. The model achieved favorable results, with small RMSE values of 0.00527, 0.00520, and 0.00447, and MAE values of 0.00406, 0.00402, and 0.00350 for Batu Muda, Cheras, and Petaling Jaya, respectively. These results demonstrate the model's accuracy in predicting daily ozone concentrations, particularly at the Petaling Jaya station, which exhibited the best performance. The findings highlight KNN's ability to handle nonlinear relationships in environmental datasets and support its application to localised air quality forecasting. However, some limitations were identified, such as the model's reduced precision at higher ozone concentrations and the variability of optimal k values across locations. These findings suggest that future applications of KNN for air pollution prediction could benefit from adaptive or hybrid strategies that take into account regional data characteristics.

Overall, the study confirms the suitability of KNN for air quality prediction, providing valuable insights for public health planning and strategies to mitigate ozone pollution in urban areas. By providing accurate ozone prediction, this research can contribute to better public health planning, timely alerts for vulnerable populations, and the formulation of effective strategies to mitigate the harmful effects of ozone pollution. From a policy standpoint, the implementation of data-driven ozone prediction tools such as the KNN model could support environmental agencies in making informed decisions on air quality control. Real-time prediction systems based on reliable machine learning models can be integrated into public advisory platforms, assisting local governments to issue warnings and promote healthier urban planning. The use of predictive analytics could further support early intervention efforts, reducing the burden on healthcare systems and enhancing the quality of life in urban areas.

Future research should focus on combining additional meteorological and emission-related variables, implementing feature selection techniques to identify the most influential predictors and comparing the performance of KNN with more advanced algorithms such as gradient boosting or deep learning to improve the accuracy and interpretability of ozone concentration estimates.

ACKNOWLEDGMENT

The authors would like to thank the Department of Environment Malaysia for providing information on air pollutants data.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Ahamad, F., Griffiths, P. T., Latif, M. T., Juneng, L., & Xiang, C. J. 2020. Ozone trends from two decades of ground level observation in Malaysia. *Atmosphere* 11(7): 755.

AlOmar, M. K., Hameed, M. M., & AlSaadi, M. A. 2020. Multi hours ahead prediction of surface ozone gas concentration: robust artificial intelligence approach. *Atmospheric Pollution Research* 11(9): 1572-1587.

Baidrulhisham, S. E., Noor, N. M., Hassan, Z., Sandu, A. V., Vizureanu, P., Ul-Saufie, A. Z., ... & Deák, G. 2022. Effects of weather and anthropogenic precursors on ground-level Ozone concentrations in Malaysian cities. *Atmosphere* 13(11): 1780.

Bekesiene, S., Meidute-Kavaliauskiene, I., & Vasiliauskiene, V. 2021. Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. *Mathematics* 9(4): 356.

de Oliveira, R. C., Cunha, C. L., Tôrres, A. R., & Corrêa, S. M. 2021. Forecasts of tropospheric ozone in the Metropolitan Area of Rio de Janeiro based on missing data imputation and multivariate calibration techniques. *Environmental Monitoring and Assessment* 193: 1-16.

Department of Environment. 2022. Environmental Quality Report 2022. Putrajaya.

Evitania, C. G. 2023. Implementation of the K-Nearest Neighbor Algorithm to Predict Air Pollution. *Information Technology and Systems* 1(1): 45-54

Jeyabalan, S. D., Yesudhas, N. J., Harichandran, K. N., & Sridharan, G. 2022. Multivariate temporal data classification framework for ozone concentration prediction. *Journal of Intelligent & Fuzzy Systems* 43(1): 143-157.

Juarez, E. K., & Petersen, M. R. 2021. A comparison of machine learning methods to forecast tropospheric ozone concentrations in Delhi. *Atmosphere* 13(1): 46.

Jumin, E., Zaini, N., Ahmed, A. N., Abdullah, S., Ismail, M., Sherif, M., ... & El-Shafie, A. 2020. Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. *Engineering Applications of Computational Fluid Mechanics* 14(1): 713-725.

Li, D., & Ren, X. 2022. Prediction of ozone hourly concentrations based on machine learning technology. *Sustainability* 14(10): 5964.

Lu, H., Xie, M., Liu, X., Liu, B., Jiang, M., Gao, Y., & Zhao, X. 2021. Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China. *Atmospheric Pollution Research* 12: 101066.

Marchang, N., & Tripathi, R. 2020. KNN-ST: Exploiting spatio-temporal correlation for missing data inference in environmental crowd sensing. *IEEE Sensors Journal* 21(3): 3429-3436.

Mohd Napi, N. M., Abdullah, S., Ahmed, A. N., Mansor, A. A., & Ismail, M. 2020, May. Annual and diurnal trend of surface ozone (O3) in industrial area. In IOP Conference Series: Earth and Environmental Science (Vol. 498, No. 1, p. 012062).

Saeipourdizaj, P., Sarbakhsh, P., & Gholampour, A. 2021. Application of imputation methods for missing values of PM_{10} and O_3 data: Interpolation, moving average and K-nearest neighbor methods. *Environmental Health Engineering and Management Journal* 8(3): 215-226.

Sarkar, S. S. Ray, A. Prasad and C. Pradhan, “A Novel Detection Approach of Ground Level Ozone using Machine Learning Classifiers,” 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 428-432

Su, X., An, J., Zhang, Y., Zhu, P., & Zhu, B. 2020. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. *Atmospheric Pollution Research* 11(6): 51-60

Tanaskuli, M., Ahmed, A. N., Zaini, N., Abdullah, S., Borhana, A. A., Mardhiah, N. A., & Mathivanan, M. 2020. Ozone prediction based on support vector machine. *Indonesian Journal of Electrical Engineering and Computer Science* 17(3): 1461-1

Wu, J., Wang, Y., Liang, J., & Yao, F. 2021. Exploring common factors influencing PM2. 5 and O3 concentrations in the Pearl River Delta: Tradeoffs and synergies. *Environmental Pollution* 285: 117138.

Xiong, G.-S., Liu, X.-Z., Li, Y., Ren, Y.-Z., Tang, Q.-Z., & Tang, X.-W. 2024. Spatiotemporal Changes in and Forces Driving Ozone Concentration in the Beijing–Tianjin–Hebei (Jing–Jin–Ji) Region from 2015 to 2022. *Atmosphere* 15(4): 416

Yafouz, A., AlDahoul, N., Birima, A. H., Ahmed, A. N., Sherif, M., Sefelnasr, A., ... & Elshafie, A. 2022. A comprehensive comparison of various machine learning algorithms for predicting short-term ozone concentrations. *Alexandria Engineering Journal* 61(6): 4607-4622.