Malaysian Journal of Computing, 11 (2): 2519-2536, 2026
eISSN: 2600-8238

COMPARATIVE MACHINE LEARNING PREDICTION OF CARBON MONOXIDE LEVELS AND CONTRIBUTING FACTORS IN SHAH ALAM

Farzana Zubaidi[1], Norshahida Shaadan[2*] and Septia Devi Prihastuti Yasmirullah[3]

[1,2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

[3]Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia

[1]farzana970824@gmail.com, [2*]norshahida588@uitm.edu.my, [3]septia.devi@ftmm.unair.ac.id

(indicates the corresponding author)*

ABSTRACT

Carbon monoxide (CO) pollution is a major environmental and public health concern because it is colourless and odourless and is produced mainly by incomplete combustion from vehicles, industries, and household sources. Accurate prediction of CO levels and identification of key contributing factors can support early warning and targeted mitigation strategies. This study compared eight machine-learning regression models—Support Vector Regression (SVR), Random Forest Regression (RFR), Decision Tree Regression (DTR), K-Nearest Neighbours (KNN), CatBoost, XGBoost, LightGBM, and Neural Network Regression (NNR)—using daily air-quality data from the Department of Environment Malaysia for 2018–2023. The dataset comprised 42 input features representing air pollutants, meteorological variables, and volatile organic compounds. Data cleaning, preprocessing, mutual-information-based feature selection, model training, and performance evaluation were conducted using the top 30, 20, and 10 features. CatBoost achieved the best performance across all feature-set sizes, with RMSE values of 0.520, 0.552, and 0.622 and corresponding R^2 values of 0.750, 0.718, and 0.642, respectively. SVR showed the weakest overall performance. The ten highest-ranked contributors were NO2, NOx, PM2.5, PM10, NO, OP_I_Butane, OP_I_Pentane, OP_N_Butane, OP_2_Methylpentane, and OP_Propane. These findings demonstrate the value of ensemble machine-learning methods for CO prediction and provide evidence on the environmental factors associated with CO variability in Shah Alam.

Keywords*: Air Quality Prediction, CatBoost, Feature Selection, Mutual Information, Volatile Organic Compounds.*

Received for review: 20-05-2026; Revised: 01-09-2026; Accepted: 01-09-2026; Published: 01-10-2026
DOI: 10.24191/mjoc.v11i2.11112

1. Introduction

Air pollution remains an important public health and environmental concern in rapidly urbanising cities. Among air pollutants, carbon monoxide (CO) is particularly hazardous because it is colourless and odourless, making exposure difficult to detect without monitoring devices. CO is produced mainly by the incomplete combustion of fossil fuels from vehicular



This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/).

and industrial activities (WHO, 2021). Prolonged exposure to CO can impair oxygen transport to the brain and heart, increasing the risk of respiratory and cardiovascular effects and premature death. In Malaysia, urban areas such as Shah Alam may face elevated CO exposure due to high traffic volumes, industrial activity, and continuing urban development. Urban CO concentrations are influenced by industrial emissions, vehicular traffic, and meteorological conditions. Industrial processes, including steel manufacturing and fossil-fuel combustion, remain important sources of CO emissions, particularly where combustion technologies are inefficient (Leguijt *et al*., 2025; Chen & Liu, 2022). Transportation emissions, especially from older vehicles and traffic congestion, can further increase urban CO levels (Latif *et al*., 2023; WHO, 2022). Meteorological factors, including temperature inversions, wind speed, humidity, and precipitation, affect pollutant dispersion and deposition and may contribute to seasonal or event-related concentration peaks (Chen *et al*., 2020; Wang & Zhang, 2021; Choi *et al*., 2023). Studies in Malaysia have reported elevated CO risks in urban areas such as Kuala Lumpur and Shah Alam in association with industrial activity and traffic congestion (Almubaidin & Mohd Zahari, 2024; Latif *et al*., 2023). Consequently, accurate data-driven prediction is important for urban air-quality management and early warning systems (Shaadan *et al*., 2019).

Traditional statistical models, including Multiple Linear Regression (MLR), have been widely used for pollutant prediction, while logistic regression and discriminant analysis have been applied to classification tasks (Ahmat *et al*., 2019). However, these models may rely on linearity assumptions and can inadequately represent nonlinear interactions between meteorological and pollutant variables, resulting in limited predictive performance (Chen *et al*., 2020; Xiong *et al*., 2023). Machine learning (ML) has therefore emerged as a viable alternative because of its ability to model complex nonlinear relationships. Support Vector Regression (SVR), Random Forest Regression (RFR), and ensemble boosting methods have demonstrated strong performance in air-pollution prediction (Latif *et al*., 2023; Halawa *et al*., 2022). Recent studies have also reported that boosting approaches and neural networks can outperform classical regressors (Yu *et al*., 2024). Unlike conventional regression models that impose stronger functional assumptions, ML methods can adapt to noisy, high-dimensional data, making them suitable for forecasting pollutants such as CO (Latif *et al*., 2023; Xiong *et al*., 2023). Machine-learning models used in air-quality research can broadly be categorised as classical regressors, ensemble or boosting models, and neural networks. Despite this progress, two gaps remain. First, limited research has explicitly focused on Shah Alam despite its distinctive industrial and residential characteristics within the Klang Valley (Almubaidin & Mohd Zahari, 2024; Latif *et al*., 2023). Second, many previous studies evaluated individual models without systematically comparing multiple ML methods using the same dataset and often considered a limited range of input variables for predicting CO levels.

Feature selection is important for identifying influential predictors and improving model interpretability. By retaining the most relevant variables, it can reduce redundancy, limit overfitting, and make findings easier to interpret for stakeholders. Filter methods, such as correlation- and mutual-information-based approaches, rank predictors independently of the final predictive model, whereas wrapper methods evaluate feature subsets according to model performance (Ul-Saufie *et al*., 2022). Jalali *et al*. (2024) introduced a two-step mutual information approach that improved pollution-prediction performance. Knowledge-driven approaches, such as that of Choy *et al*. (2021), used meteorological predictors for carbon monoxide forecasting to improve interpretability. Hybrid optimisation approaches, including genetic algorithms and swarm-based metaheuristics, have also shown promise in environmental modelling (Babaali *et al*., 2026). Despite these developments, few studies have systematically applied feature selection to carbon monoxide forecasting in Malaysia, indicating a continuing research gap.

In the Malaysian context, air-pollution research remains fragmented and often focuses on a single model or dominant pollutants such as PM10, PM2.5, and ozone (Azli *et al*., 2023), with comparatively fewer studies on carbon monoxide using ML models and diverse input features. This study therefore developed a CO-prediction framework that identifies both an

effective predictive model and the dominant variables associated with CO variability. Unlike prior studies, the proposed approach integrates three predictor domains: criteria pollutants and particulate matter, meteorological parameters, and chemically differentiated volatile organic compounds (VOCs). Eight regression models SVR, Decision Tree Regression (DTR), Random Forest Regression (RFR), KNN, XGBoost, LightGBM, CatBoost, and Neural Network Regression (NNR) were compared using six years of DOE Shah Alam data (2018–2023). Specifically, the study aimed to (i) analyse temporal patterns in CO levels, (ii) identify key contributing features using feature selection, and (iii) compare the predictive performance of the eight regression models. The study is intended to strengthen methodological rigour and understanding of CO pollution while providing evidence relevant to the Department of Environment (DOE) Malaysia.

2. Methodology

This section describes the research design and methodology, including the research framework, data, preprocessing and feature selection, model development, and evaluation procedures.

2.1 Research Framework

The research objectives were addressed using the systematic process flow shown in Figure 1. The procedure began with problem identification (Step 1), followed by data acquisition (Step 2) and data understanding (Step 3), during which pollutant and meteorological variables were explored to describe trends in CO levels.

Data preparation in Step 4 included data cleaning, transformation, and structuring to ensure suitability for modelling. Feature selection was then conducted in Step 5, followed by model development in Step 6 using eight regression algorithms: SVR, DTR, RFR, XGBoost, LightGBM, CatBoost, KNN, and NNR. The process concluded with model evaluation in Step 7. All eight algorithms were evaluated using a 70/30 train-test split under controlled baseline configurations to support reproducibility and fair comparison.

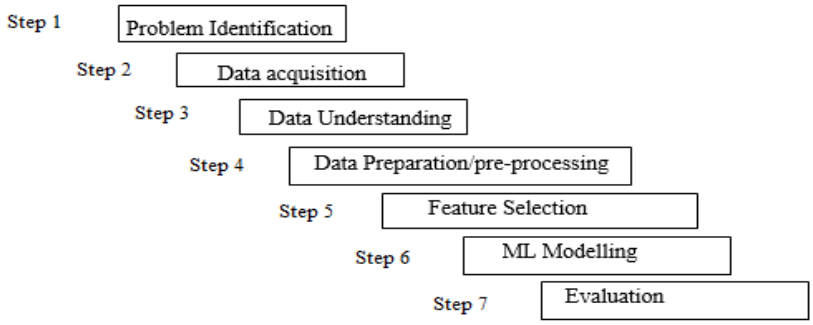


Figure 1. Research Process Flow

The research framework was designed for several audiences: academic researchers seeking to replicate or extend the modelling approach; policymakers and environmental agencies, including DOE Malaysia, that may use model outputs for air-quality management; and industry practitioners, such as data scientists and environmental engineers, seeking practical tools for pollution prediction.

2.2 Data and Description

This study used a dataset obtained from the Department of Environment (DOE) Malaysia comprising daily observations from 2018 to 2023 at the Shah Alam monitoring station. The dataset included gaseous and particulate air pollutants, meteorological variables, and an extensive set of volatile organic compounds (VOCs) associated with vehicular and other emission sources. It contained 2,192 records and 42 independent variables: seven air-pollutant variables, five meteorological variables, and 30 VOC/emission-related variables. The average CO concentration (ppm) was the dependent or target variable. Table 1 summarises the variables used in the study.

Table 1. Description of Study Variables

Category	Variables	Unit	Description
Air Pollutants	PM10	µg/m³	Particulate matter <10 µm
	PM2.5	µg/m³	Fine particulate matter <2.5 µm
	SO2	ppm	Sulphur dioxide concentration
	NO2	ppm	Nitrogen dioxide concentration
	NO	ppm	Nitric oxide concentration
	NOx	ppm	Total nitrogen oxides (NO + NO2)
	O_3	ppm	Ozone concentration
Meteorological	Wind Direction	(°)	Mean wind direction
	Wind Speed	m/s	Mean wind speed
	Relative Humidity	%	Atmospheric moisture
	Solar Radiation	W/m²	Solar energy intensity
	Ambient Temperature	(°C)	Air temperature
Volatile Organic Compounds (VOC)	Benzene	ppm	Vehicle exhaust marker
	Toluene	ppm	Traffic-related solvent emission
	Xylenes (o-, m-, p-)	ppm	Aromatic hydrocarbons from fuel combustion
	Ethane/Ethene/Acetylene	ppm	Combustion by-products from vehicles
	Propene (1-, 2-,CIS-2) Butane Butadiene Pentane (N-, 1-) Pentene (1-, Trans2-)	ppm	Indicators of vehicular fuel use
	Propane	ppm	Indicators of vehicular fuel use
	Isoprene	ppm	Natural + vehicle-related hydrocarbon
	Other VOCs: Trimethylbenzenes (1-2-3,1-2-4, 1-3-5) Hexane Heptane Octane Methylpentane Trimethylpentane Ethylbenzene Xylene (p-, m-, o-)	ppm	Additional traffic/industrial hydrocarbon emissions
Target Variable	CO	ppm	Carbon monoxide levels

2.3 Data Preprocessing and Feature Selection

Data preparation is a crucial first step in ensuring that a dataset is suitable for reliable modelling and analysis. Several preprocessing procedures were applied to improve data quality and consistency across variables. Missing values were handled using median imputation, which is suitable for skewed environmental data and reduces the influence of extreme values on the imputed estimates. Simple imputation methods such as median substitution are commonly used when the proportion of missing data is small, particularly below 5% of the dataset (Graham, 2009; Joel *et al*., 2025). This approach retained the completeness of the dataset for subsequent model training.

For regression-based predictive modelling, outliers in the target variable, CO, were considered during data cleaning. Such observations may arise from recording errors, sensor issues, or genuinely unusual events, including local fires or haze episodes. If retained without assessment, extreme values may influence model fitting and contribute to unstable predictions. Outliers were identified using the Local Outlier Factor (LOF) method, which accounts for multidimensional local-density variation and is therefore appropriate for datasets containing interacting environmental and chemical variables. Özüpak *et al*. (2025) similarly reported improved predictive stability after outlier treatment in air-quality forecasting. Min-Max scaling was then applied to normalise numerical variables to the range 0–1. This step was required for distance- and gradient-sensitive models such as SVR, KNN, and NNR, which can be affected by differences in feature scale.

Feature selection can be performed using filter, wrapper, or embedded approaches. This study used a filter-based approach because it is computationally efficient, model-agnostic, and suitable for high-dimensional data. Unlike wrapper methods, which repeatedly train models to evaluate feature subsets, or embedded methods, which are model-specific, filter methods rank variables independently of the predictive algorithm. Mutual Information (MI) was selected because it can capture nonlinear dependencies between continuous predictors and the target variable (Estévez *et al*., 2009).

Initially, the dataset contained 42 independent variables, not all of which were expected to contribute equally to CO prediction. Using the full feature set could increase model complexity and the risk of learning noise. MI scores were therefore calculated to rank the predictors according to their association with CO levels. Three feature subsets—the top 10, top 20, and top 30 variables were then evaluated to examine how dimensionality affected predictive performance. RapidMiner and Python were used for data analysis and model development.

2.4 Machine Learning Models

To predict Shah Alam's CO levels, this study employed multiple machine-learning algorithms to predict the target variable, including Decision Tree Regression, Random Forest Regression, K-Nearest Neighbours Regression, XGBoost Regression, Support Vector Regression, CatBoost Regression, LightGBM Regression, and Neural Network Regression. These models were selected for their complementary strengths in capturing linear, nonlinear, and interaction effects in complex datasets. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, each of the algorithms is applied to predict a new target observation, $\hat{y}_{i=n+1} = \hat{f}_{i=n+1}(x)$.

2.4.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) was selected for its ability to model nonlinear relationships between input variables and the target. SVR fits a function that approximates the data while keeping prediction errors within a defined tolerance margin (epsilon). The underlying optimisation problem is solved using convex quadratic programming, which yields a global optimum for the specified formulation (Zhang *et al*., 2024).

The method used a radial basis function (RBF) kernel, which can capture nonlinear relationships by mapping input features into a higher-dimensional feature space. The resulting prediction function is expressed as:

$$y = \omega^T \emptyset(x) + b \quad (1)$$

where y is the predicted output, ω is the weight vector, $\emptyset(x)$ is the kernel-based transformation of input features and b is the bias term.

The key hyperparameters were the regularisation parameter C, kernel coefficient (γ), and epsilon. These parameters were tuned using grid search and random search to balance bias and variance. C controls the trade-off between model complexity and training error, γ determines the influence of individual training samples, and epsilon defines the error-insensitive margin. The optimised model was then trained and evaluated using RMSE, MAE, and R^2 to assess generalisation to unseen data. Finally, the trained SVR model was used to predict CO levels from the selected environmental variables.

2.4.2 Random Forest Regression (RFR)

Random Forest Regression (RFR), introduced by Breiman (2001), is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. RFR uses bootstrap sampling (bagging) to generate multiple subsets of the training data. Unlike a single decision tree, which can exhibit high variance, RFR constructs many trees using different samples and feature subsets.

The final RFR prediction is obtained by averaging the outputs of the individual trees. Each tree is trained on a bootstrap sample drawn with replacement from the original training data, introducing diversity and improving robustness. Random feature selection further increases diversity among trees and can reduce variance compared with a single decision tree. Further details are provided by Zhou *et al*. (2023). RFR is non-parametric, can capture nonlinear relationships and feature interactions, and is generally robust to noise. These properties make it suitable for high-dimensional environmental data.

2.4.3 Decision Tree Regression (DTR)

Decision Tree Regression is a non-parametric supervised learning method that splits the feature space into a set of disjoint regions R_m (where m = 1,2, …, M) and fits a constant model within each region. The algorithm recursively partitions the training data based on input predictor variables to minimise prediction error.

For the given region R_m, the predicted target value y_m is calculated as the mean of the actual target values y_i of all N_m observations falling within that terminal node:

$$y_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i \quad (2)$$

At each internal node, the optimal split feature j and split point s are determined by selecting the boundary that minimises the sum of squared residuals across the resulting sub-regions (R_1 and R_2):

$$min_{j,s} = \left[\sum_{x_i \in R_1(j,s)} (y_i - y_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - y_2)^2\right] \quad (3)$$

The recursive partitioning process continues until a predefined stopping criterion, such as maximum tree depth (max_depth) or minimum samples per leaf (min_samples_leaf), is reached.

DTR can capture nonlinear relationships and feature interactions without requiring strong assumptions about data distribution. However, unconstrained deep trees may overfit the training data, leading to high variance on unseen observations. Despite this limitation, its interpretability makes it a useful baseline model. Further details are provided by Zahirsyah and Setiyaningsih (2025) and Mahamat *et al*. (2024).

2.4.4 K-Nearest Neighbour (KNN)

K-Nearest Neighbours Regression (KNN) is a non-parametric method that predicts a new target observation based on neighbouring data in the available input features. The method relies on a distance metric (commonly Euclidean distance) to identify the most similar observations. The prediction for a new input x is computed as the average of the responses of its k-nearest neighbours:

$$\hat{f}(x) = \frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} y_i \tag{4}$$

where $\mathcal{N}_k(x)$denotes the set of $k$nearest neighbours of x. The distance between observations is typically measured using Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p}(x_j - x_{ij})^2} \tag{5}$$

KNN regression is data-driven and is capable of capturing nonlinear relationships between the target and the input features. However, its performance is sensitive to the choice of k, feature scaling, and the presence of irrelevant variables. According to Zhang *et al*. (2024), despite these limitations, it serves as a simple yet effective baseline model, particularly in low- to moderate-dimensional data.

2.4.5 XGBoost

XGBoost Regression is an efficient and scalable implementation of gradient boosting. The method builds an ensemble of regression trees sequentially, where each new tree is trained to minimise the residual errors of the preceding ensemble using gradient-based optimisation. The prediction at iteration $t$is updated as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathrm{x_i}) \tag{6}$$

where $f_t(\cdot)$denotes the newly added regression tree. The model is trained by minimising a regularised objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T}\Omega(f_t) \tag{7}$$

where $l(\cdot)$is a differentiable loss function (e.g., squared error), and $\Omega(f_t)$is a regularisation term controlling model complexity.

XGBoost uses first- and second-order gradient information to optimise the loss efficiently, while regularisation helps reduce overfitting and improve generalisation. It can capture complex nonlinear relationships and feature interactions and is robust to multicollinearity and noisy data, making it suitable for high-dimensional predictive modelling.

2.4.6 LightGBM

LightGBM Regression is a gradient-boosting framework designed for efficient and scalable tree-based learning. It builds an ensemble of decision trees sequentially using a leaf-wise growth strategy with depth constraints, where each new tree is fitted to reduce residual error from the preceding ensemble. LightGBM can efficiently handle large datasets and high-dimensional features and can capture nonlinear interactions (Ke *et al*., 2017).

For the given training dataset, the prediction at iteration t is given as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{8}$$

where $f_t(\cdot)$represents the tree added at iteration t. The model minimises a regularised objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t) \tag{9}$$

where $l(\cdot)$is a differentiable loss function (e.g., squared error), and $\Omega(f_t)$penalises tree complexity to prevent overfitting.

2.4.7 CatBoost

CatBoost is a gradient boosting framework specifically designed to handle categorical features efficiently while mitigating prediction bias. Like other boosting algorithms, CatBoost builds an ensemble of decision trees sequentially, where each tree is trained to correct the residual errors of the previous trees. It uses ordered boosting and permutation-driven techniques to reduce overfitting and improve generalisation. The prediction at the iteration t is as follows.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{10}$$

where $f_t(\cdot)$represents the newly added tree. The model minimises a regularised loss function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t) \tag{11}$$

where $l(\cdot)$is the differentiable loss function, and $\Omega(f_t)$is a regularisation term that penalises tree complexity.

CatBoost can capture nonlinear interactions and complex patterns in high-dimensional datasets and is particularly effective for data containing categorical or mixed feature types (Prokhorenkova *et al*., 2018).

2.4.8 Neural Network Regression (NNR)

Neural Network Regression is a nonlinear function approximation method inspired by the structure of the human brain. It models complex relationships by transforming input features through multiple layers of interconnected neurons. Each neuron applies a weighted sum of its inputs followed by a nonlinear activation function, allowing the network to capture highly nonlinear patterns and interactions. For a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, the prediction of a feedforward neural network with L layers can be expressed as follows.

$$\hat{y}_i = f(x_i; \theta) = f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}(x_i) \tag{12}$$

where $f^{(l)}(\cdot)$denotes the transformation at the layer $l$and θrepresents all network weights and biases learned via optimisation to minimise a loss function such as Mean Squared Error (MSE).

NNR is particularly suitable for modelling highly nonlinear relationships and interactions in high-dimensional data. However, it requires careful hyperparameter tuning, sufficient data, and regularisation techniques to prevent overfitting (Huang *et al*., 2024).

2.5 Model Evaluation

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) were used to evaluate model performance. RMSE measures the square root of the mean squared difference between observed and predicted values and therefore penalises larger errors more heavily than smaller ones. Because it is expressed in the same units as the target variable, lower RMSE values indicate closer agreement between predictions and observations. MAE measures the average absolute prediction error and is less sensitive to extreme errors than RMSE. R^2 quantifies the proportion of variance in the observed data explained by the model; higher values indicate greater explanatory performance (Ibrahim *et al*., 2024).

3. Results and Discussion

This section presents the main findings according to the three research objectives: (i) analysis of temporal patterns in CO levels in Shah Alam (RO1), (ii) identification of dominant predictors using feature selection (RO2), and (iii) evaluation of regression models for CO prediction (RO3). The effects of preprocessing and feature reduction on model performance are also discussed.

3.1 CO Level Patterns in Shah Alam

Annual average CO concentrations (Figure 2) decreased from 0.973 ppm in 2018 to 0.544 ppm in 2022 before increasing to 0.724 ppm in 2023. The marked reduction during 2020 coincided with COVID-19 movement restrictions and reduced transportation and industrial activity. The increase in 2023 coincided with the resumption of transportation and industrial activity. Overall, the six-year pattern shows a decline followed by a post-pandemic rebound, underscoring the importance of continued monitoring as Shah Alam progresses towards its 2030 Low Carbon City goals.

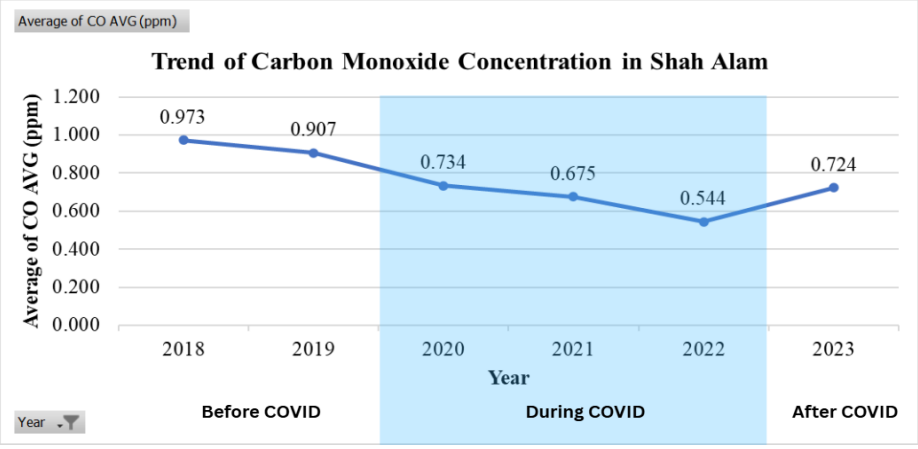


Figure 2. Annual Trend in CO Levels in Shah Alam, 2018–2023

The LOF method identified several outlying CO observations that may reflect unusual environmental events, sensor anomalies, or other atypical conditions (Figure 3). Removing these observations produced a more stable distribution with lower variance, thereby supporting

subsequent model fitting. Similar benefits of outlier treatment have been reported in air-quality forecasting studies (Özüpak *et al*., 2025).

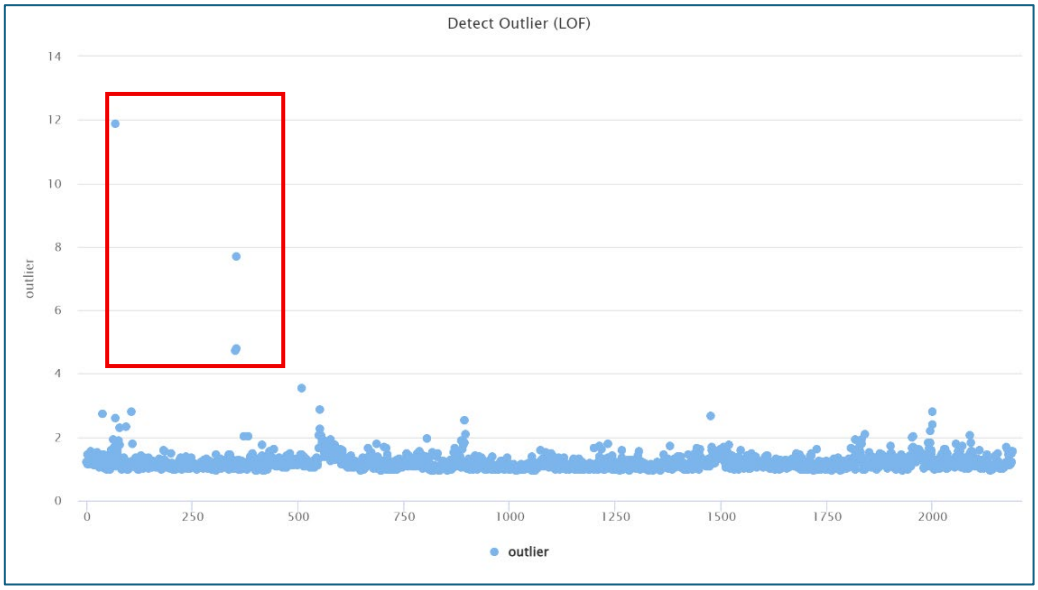


Figure 3. CO Observations Identified as Outliers Using the Local Outlier Factor (LOF), 2018–2023

3.2 Key Feature Identification

Mutual-information-based feature selection produced ranked lists of influential variables. Across the top-10, top-20, and top-30 feature sets, nitrogen oxides, particulate matter, and volatile organic compounds (VOCs) consistently appeared among the highest-ranked predictors, indicating that CO variability in Shah Alam is associated with a combination of combustion-related pollutants and atmospheric factors.

3.2.1 Top 10 Features by Weight

The highest-ranked predictors in Table 2 were NO2, NOx, PM2.5, PM10, and NO, all of which are associated with combustion and transportation emissions. The presence of VOCs such as OP_I_Butane, OP_I_Pentane, and OP_N_Butane among the top features suggests that CO variability may also be associated with fuel-related and industrial emission profiles. These results indicate that CO should be interpreted as part of a broader urban emission mixture rather than in isolation.

Table 2. Top 10 Features Ranked by Mutual Information Weight

Attribute	Weight
NO2	0.606
NOX	0.577
PM2.5	0.506
PM10	0.446
NO	0.446
OP_I_Butane	0.387
OP_I_Pentane	0.341
OP_N_Butane	0.309
OP_2_Methylpentane	0.283
OP_Propane	0.275

3.2.2 Top 20 Features by Weight

Expanding the feature set to 20 variables (Table 3) introduced additional VOCs, including propene, xylene species, and ethylbenzene, as well as meteorological variables such as wind direction and solar radiation. Wind patterns affect pollutant dispersion, while solar radiation can influence atmospheric photochemistry. The combined ranking of pollutant and meteorological variables indicates that CO modelling benefits from considering both emission-related and atmospheric factors.

Table 3. Top 20 Features Ranked by Mutual Information Weight

Attribute	Weight
NO2	0.606
NOX	0.577
PM2.5	0.506
PM10	0.446
NO	0.446
OP_I_Butane	0.387
OP_I_Pentane	0.341
OP_N_Butane	0.309
OP_2_Methylpentane	0.283
OP_Propane	0.275
OP_Propene	0.269
OP_O_Xylene	0.240
OP_Ethylbenzene	0.231
Wind direction	0.229
OP_Ethene	0.227
OP_N_Pentane	0.222
Solar radiation	0.221
OP_N_Heptane	0.217
OP_1_Pentene	0.210
OP_N_Octane	0.205

3.2.3 Top 30 Features by Weight

The 30-feature set (Table 4) included additional VOC species and meteorological variables, including ambient temperature. These variables may help the models capture secondary interactions among pollutants and atmospheric conditions. The repeated presence of several VOCs across the three feature sets indicates their potential relevance to CO variability in Shah Alam.

Table 4. Top 30 Features Ranked by Mutual Information Weight

Attribute	Weight
NO2	0.606
NOX	0.577
PM2.5	0.506
PM10	0.446
NO	0.446
OP_I_Butane	0.387
OP_I_Pentane	0.341

Attribute	Weight
OP_N_Butane	0.309
OP_2_Methylpentane	0.283
OP_Propane	0.275
OP_Propene	0.269
OP_O_Xylene	0.240
OP_Ethylbenzene	0.231
Wind direction	0.229
OP_Ethene	0.227
OP_N_Pentane	0.222
Solar radiation	0.221
OP_N_Heptane	0.217
OP_1_Pentene	0.210
OP_N_Octane	0.205
OP_1_3_5_Trimethylbenzene	0.173
OP_CIS_2_Butene	0.173
OP_TRANS_2_Pentene	0.170
OP_2_2_4_Trimethylpentane	0.155
OP_Trans_2_Butene	0.144
Ambient Temperature	0.135
OP_M_Xylene	0.128
OP_P_Xylene	0.117
OP_N_Hexane	0.108
OP_Benzene	0.1010

Overall, the feature-selection results indicate that a broader predictor set combining combustion-related pollutants, particulate matter, VOCs, and meteorological indicators provides more explanatory information than a set limited to conventional air pollutants.

3.3 Model Performance and Evaluation

Model performance was evaluated using RMSE, MAE, and R^2 across the top 30, top 20, and top 10 feature sets. All eight algorithms used a 70/30 train-test split under controlled baseline configurations to support reproducibility and fair comparison. Distance-sensitive models (SVR, KNN, and NNR) were fitted using scaled features, whereas tree-based and boosting models (RFR, DTR, XGBoost, LightGBM, and CatBoost) used the continuous input features directly.

3.3.1 Master Reproducibility Matrix of Evaluated Regression Models

To support reproducibility and comparability, the regression models were evaluated using controlled baseline configurations. Applying a consistent parameter setup across feature-set experiments helps ensure that differences in performance primarily reflect model structure and feature dimensionality rather than repeated experiment-specific tuning. Table 5 lists the final configuration values used for each algorithm.

Table 5. Master Reproducibility Matrix of Evaluated Regression Models

Model Name	**Hyperparameter / Configuration**	**Final Selected Value(s)**
SVR	kernel_type kernel_gamma C convergence_epsilon	radial (RBF) 1.0 0.0 0.001
RFR	number_of_trees criterion maximal_depth minimal_leaf_size minimal_size_for_split	100 least_square 10 2 4
KNN	k weighted_vote measure_types numerical_measure	5 true (distance-weighted) NumericalMeasures ManhattanDistance
NNR	training_cycles learning_rate momentum error_epsilon	200 0.01 0.9 1.0E-4
XGBoost	rounds learning_rate max_depth min_child_weight lambda	100 0.3 4 1.0 1.0
DTR	criterion maximal_depth apply_pruning apply_prepruning minimal_gain	least_square 10 true true 0.01
LightGBM	n_estimators learning_rate max_depth boosting_type	100 0.1 -1 (unlimited) gbdt
CatBoost	iterations learning_rate depth loss_function	1000 0.03 6 RMSE

All models used a 70/30 train-test split. SVR, KNN, and NNR were fitted using scaled input vectors, whereas RFR, XGBoost, DTR, LightGBM, and CatBoost were fitted directly to the continuous features because tree-based methods are generally scale-invariant. Maintaining the same configurations across the feature-set experiments provides a consistent basis for comparing the effects of feature reduction across the 30-, 20-, and 10-feature subsets.

3.3.2 Comparative Assessment

Table 6 compares model performance across the three feature sets. With the top 30 features, CatBoost achieved the best performance, with the lowest RMSE (0.520) and MAE (0.399) and the highest R^2 (0.750). Thus, CatBoost accounted for 75% of the variance in CO levels while producing the lowest prediction errors among the evaluated models. LightGBM and XGBoost were the next strongest performers: LightGBM achieved an RMSE of 0.559 and R^2 of 0.711, while XGBoost achieved an RMSE of 0.593 and R^2 of 0.632. Both gradient-boosting models showed strong predictive performance, although they were less accurate than CatBoost in this study.

Table 6. Performance of Machine-Learning Models Across Feature Sets

Model	Feature Selection	RMSE (ppm)	MAE (ppm)	R^2
CatBoost	30 Features	0.520	0.399	0.750
	20 Features	0.552	0.427	0.718
	10 Features	0.622	0.483	0.642
LightGBM	30 Features	0.559	0.429	0.711
	20 Features	0.587	0.457	0.681
	10 Features	0.628	0.492	0.635
XGBoost	30 Features	0.593	0.445	0.632
	20 Features	0.620	0.465	0.600
	10 Features	0.668	0.505	0.538
RFR	30 Features	0.635	0.491	0.598
	20 Features	0.647	0.503	0.573
	10 Features	0.682	0.535	0.512
NNR	30 Features	0.635	0.494	0.582
	20 Features	0.677	0.514	0.567
	10 Features	0.799	0.624	0.508
KNN	30 Features	0.638	0.469	0.588
	20 Features	0.663	0.496	0.561
	10 Features	0.721	0.543	0.472
DTR	30 Features	0.813	0.609	0.375
	20 Features	0.781	0.596	0.419
	10 Features	0.775	0.597	0.407
SVR	30 Features	0.921	0.705	0.168
	20 Features	0.865	0.656	0.259
	10 Features	0.762	0.585	0.391

To assess whether a smaller feature set could retain comparable predictive performance with lower dimensionality, the analysis was repeated using the top 20 features. CatBoost again achieved the lowest RMSE (0.552), followed by LightGBM (0.587) and XGBoost (0.620). These errors were slightly higher than those obtained with 30 features, but the relative ordering of the three leading models was unchanged. DTR (0.781) and SVR (0.865) had the highest RMSE values. For MAE, CatBoost again recorded the lowest value (0.427), followed by LightGBM (0.457) and XGBoost (0.465), while DTR (0.596) and SVR (0.656) produced larger errors.

For the 20-feature set, CatBoost also achieved the highest R^2 (0.718), followed by LightGBM (0.681) and XGBoost (0.600). These values were lower than those obtained with 30 features, indicating that the reduced models explained a smaller proportion of the variance in CO levels. DTR (0.419) and SVR (0.259) showed lower R^2 values. Overall, the 20-feature set retained substantial predictive information, while the additional variables in the 30-feature set improved model performance.

The top-10 feature set was then evaluated to determine whether CO levels could be predicted adequately using only the highest-ranked variables. Although this reduced input

dimensionality, it also decreased predictive performance. CatBoost remained the strongest model, with the lowest RMSE (0.622) and MAE (0.483) and the highest R^2 (0.642). LightGBM ranked next, followed by XGBoost. The results indicate that a relatively small set of important pollutants, particularly nitrogen oxides and particulate matter, can still support CO prediction, although accuracy was lower than with the 20- and 30-feature sets. SVR and DTR showed R^2 values below 0.40, highlighting the influence of model choice when fewer predictors are available.

Overall, Table 6 shows that the 30-feature set produced the highest predictive accuracy and lowest error values, particularly for CatBoost, LightGBM, and XGBoost. The 20-feature set provided a compromise between predictive performance and reduced dimensionality, while restricting the models to the top 10 features resulted in a clearer decline in generalisation performance. This progressive reduction in performance suggests that CO variability in Shah Alam is associated with a diverse set of interacting environmental factors and that excluding secondary but informative variables reduces the models' ability to represent these relationships.

4. Conclusion and Recommendations

This study evaluated eight machine-learning regression models for predicting CO concentrations in Shah Alam using pollutant measurements, meteorological variables, and VOCs. A structured workflow encompassing data preprocessing, feature selection, model development, and performance evaluation was used to address the three research objectives.

The temporal analysis showed a decline in annual CO concentrations from 2018 to 2022, followed by an increase in 2023. The reduction during the COVID-19 period coincided with lower mobility and activity levels, while concentrations increased as transportation and industrial activity resumed. Mutual-information-based feature selection identified nitrogen oxides (NO, NO2, and NOx), particulate matter (PM2.5 and PM10), VOCs, and meteorological variables among the key predictors of CO levels. The inclusion of VOCs and meteorological variables in the 20- and 30-feature sets improved predictive performance, indicating that CO variability is associated with multiple interacting environmental factors.

Gradient-boosting algorithms, particularly CatBoost, outperformed the other modelling approaches in predictive accuracy. For the 30-feature set, CatBoost achieved the highest R^2 and the lowest RMSE and MAE among the evaluated models. Performance declined as the number of input features was reduced, indicating that the models benefited from the broader environmental information captured by the larger feature sets. Overall, the results show that combining pollutant, VOC, and meteorological predictors with ensemble models can improve CO prediction. The findings provide a data-driven basis for air-quality monitoring in Shah Alam and may inform environmental management and related sustainability initiatives.

Future research could extend this work using models that explicitly capture temporal dependence, such as LSTM, GRU, or hybrid ARIMA-machine-learning approaches. These methods may better represent seasonality and longer-term pollutant dynamics. Further data enrichment is also recommended. Integrating real-time information on traffic flows, industrial-emission activity, and additional meteorological variables could provide a more dynamic representation of conditions associated with sudden changes in CO levels. Expanding the analysis to multiple monitoring stations across the Klang Valley would also allow the generalisability of the findings to be assessed at a broader regional scale.

Future studies could also investigate hybrid or ensemble systems that combine statistical and machine-learning approaches. Incorporating contextual variables such as traffic regulations, urban zoning, and public-transport activity may further improve the usefulness of the models for environmental decision-making and regional air-quality management.

Acknowledgement

The authors gratefully acknowledge the Department of Environment, Putrajaya, Malaysia, for providing the data and the Faculty of Computer and Mathematical Sciences for the support provided for this publication.

Funding

The authors received no specific funding for this work.

Author Contribution

Norshahida Shaadan conceptualised and designed the study. Farzana Zubaidi conducted the literature review, developed the methodology, acquired and processed the data, performed the analysis, and prepared the first draft of the manuscript. Septia Devi Prihastuti Yasmirullah contributed to proofreading and manuscript review. All authors reviewed and approved the final version.

Declaration of Generative AI in the Writing Process

During the preparation of this work, the authors used Grammarly and Gemini to improve the readability and language quality of the manuscript. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability/Supplementary Materials

Data sharing does not apply to this article as no new datasets were generated or analysed during the current study.

Ethics Statement

The authors declare that this research did not involve human participants or animal subjects. Ethical approval to conduct the research was obtained from the Ethics Committee of FSKM, UiTM before the study commenced.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Ahmat, H., Musa, N. S., Nazamid, N., & Zaharin, N. A. (2019). Classification of high and low level of PM_{10} concentrations in Klang and Shah Alam, Malaysia. *Malaysian Journal of Computing, 4*(2), 325–334**.**

Almubaidin, M. A., & Mohd Zahari, N. S. (2024). Machine learning predictions for carbon monoxide emissions. *Journal of Environmental Management, 300*, 113654.

Azli, E. S., Mat Asri, N. S., Ahmad Tajuddin, S. N., Shaadan, N., Muhamad Jamil, S. A., & Sapri, S. (2023). Modelling time series tropospheric ozone data and the precursors to

study effect and relationship in Petaling Jaya and Shah Alam Malaysia. *Malaysian Journal of Computing, 8*(2), 1560–1573.

Babaali, H., Chamanpira, R., & Dehghani, R. (2026). Estimating evaporation in diverse Iranian climates using hybrid metaheuristic algorithms and ANN. *International Journal of Environmental Science and Technology*, *23*(3), 188.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32**.**

Chen, H., & Liu, J. (2022). The impact of industrial facilities on local air quality in urban settings. *Environmental Pollution, 284,* 117361.

Chen, P., Xu, X., & Lu, H. (2020). Long-term variations in carbon monoxide at a background station in the Yangtze River Delta, China. *Atmospheric Chemistry and Physics, 20(24),* 15969–15983.

Choi, W., Ho, C.-H., & Kim, K.-Y. (2023). Critical contribution of moisture to the air quality deterioration in warm and humid weather. *Scientific Reports, 13*(1).

Estévez, P. A., Tesmer, M., Pérez, C. A., & Zurada, J. M. (2009). *Normalized mutual information feature selection*. IEEE Transactions on Neural Networks.

Graham, J. W. (2009). Missing data analysis. *Annual Review of Psychology, 60*, 549–576.

Halawa, E., Bindar, Y., Purqon, A., & Srigutomo, W. (2022). Prediction of carbon monoxide concentration with variation of support vector regression kernel parameter value. *Journal of Mathematical and Fundamental Sciences*, *54*(1).

Huang, W., Chen, J., & Wu, Y. (2024). Optimized neural networks for regression in complex datasets. *Applied Soft Computing*, 127, 109565.

Ibrahim, A., Abdelhamid, A. A., & Almetwally, E. M. (2024). *Predictive modeling of global educational outcomes: A comparative analysis using machine learning regression techniques*. *Journal of AI & ML*, *7*(2), 51–62.

Jalali, H., Keynia, F., Amirteimoury, F., & Heydari, A. (2024). A short-term air pollutant concentration forecasting method based on a hybrid neural network and metaheuristic optimisation algorithms. *Sustainability, 16*(11), 4829.

Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*, *20*(7), 6357-6373.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Latif, S. D., Almalayih, M., Yafouz, A., Ahmed, A. N., Zaini, N., Irwan, D., AlDahoul, N., Sherif, M., & El-Shafie, A. (2023). Prediction of atmospheric carbon monoxide concentration utilizing different machine learning algorithms: A case study in Kuala Lumpur, Malaysia. *Environmental Technology & Innovation, 32, 103387.*

Leguijt, G., Maasakkers, J. D., Denier van der Gon, H. A. C., Segers, A. J., Borsdorff, T., & Aben, I. (2025). Comparing Space-Based to Reported Carbon Monoxide Emission

Estimates for Europe's Iron and Steel Plants. Available at https://doi.org/10.5194/egusphere-egu24-10779

Mahamat, A. A., Boukar, M. M., Leklou, N., Celino, A., Obianyo, I. I., Bih, N. L., ... & Savastanos Jr, H. (2024). Decision tree regression vs. gradient boosting regressor models for the prediction of hygroscopic properties of Borassus fruit fiber. *Applied Sciences*, *14*(17), 7540.

Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air quality forecasting using machine learning: comparative analysis and ensemble strategies for enhanced prediction. *Water, Air, & Soil Pollution*, *236*(7), 464.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features*. Advances in Neural Information Processing Systems, 31.

Shaadan, N., Rusdi, M. S., Nik Mohd Azmi, N. N. S., Talib, S. F., & Wan Azmi, W. A. (2019). Time series model for carbon monoxide (CO) at several industrial sites in Peninsular Malaysia. *Malaysian Journal of Computing, 4*(1), 246–260.

Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R., Sandu, A. V., Deak, G., & Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. *Sustainability, 14*(18), 11403.

Wang, Z., Zhao, Z., & Wang, C. (2021). Random Forest analysis of factors affecting urban carbon emissions in cities within the Yangtze River Economic Belt. *PLOS ONE, 16(6).*

WHO. (2021). WHO global air quality guidelines aim to save millions of lives from Air Pollution. World Health Organization. Available at https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

Xiong, W., Huang, B., & Li, X. (2023). Mini-review on machine learning applications in renewable energy power generation forecasting, *Frontiers in Energy Research.* Available at https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2023.1218603

Yu, W., Xia, L., & Cao, Q. (2024). A machine learning algorithm to explore the drivers of carbon emissions in Chinese cities. *Scientific Reports*, *14*(1), 23609.

Zahirsyah, M. D., & Setiyaningsih, T. (2025). Decision Tree Regression Approach to Modeling Dengue, Tuberculosis, and Diarrhea Case Numbers. *Journal Technology Information and Data Analytic*, *2*(2), 126-130.

Zhang, X., Wang, T., & Lai, Z. (2024). A feature-weighted support vector regression machine based on Hilbert–Schmidt independence criterion least absolute shrinkage and selection operator. *Information, 15*(10), 639**.**

Zhou, Z., Qiu, C., & Zhang, Y. (2023). A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models*. Scientific Reports, 13*, 22420**.**