Comparative Modeling of Air Pollutants in Malaysian Cities Using Correlation, Causality, and Neural Network Approaches

Norazrin Ramli [1,2,*], Zulkifli A. Rais [1], Norazian M. Noor [1,2], Ahmad Z. Ul-Saufie [3], Hazrul A. Hamid [4], and Mohd Khairul N. Mahmad [5]

[1] Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, Arau 02600, Perlis, Malaysia
[2] Sustainable Environment Research Group (SERG) Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), Universiti Malaysia Perlis, Arau 02600, Perlis, Malaysia
[3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Shah Alam 40450, Selangor, Malaysia
[4] School of Distance Education, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia
[5] Mining and Energy Resources Academy (MERA), Jalan Kuala Ketil, Parit Panjang, 09100 Baling, Kedah
Email: norazrin@unimap.edu.my (N.R.); zulkiflirais@studentmail.unimap.edu.my (Z.A.R.); norazian@unimap.edu.my (N.M.N.); ahmadzia101@uitm.edu.my (A.Z.U.); hazrul@usm.my (H.A.H.); nizam.mahmad@gmail.com (M.K.N.M.)
*Corresponding author

Manuscript received November 17, 2025; revised February 10, 2026; accepted March 6, 2026; published August 19, 2026

***Abstract*—Air quality prediction plays a crucial role in addressing urban environmental challenges and supporting sustainable policy interventions. This study presents a comparative modeling analysis of five major air pollutants (PM_{10}, SO_2, NO_2, O_3, and CO) across three Malaysian cities: Nilai, Larkin, and Pasir Gudang. Observed pollutant concentrations were modeled using three approaches: Pearson correlation, Granger causality, and Artificial Neural Networks (ANN), with performance evaluated using Root Mean Square Error (RMSE), Mean Error (ME), Normalized Absolute Error (NAE), and the coefficient of determination (R^2). The results revealed that Multiple Linear Regression (MLR), guided by Pearson correlation-based variable selection, consistently provided the highest accuracy ($R^2 > 0.95$) across all cities, highlighting the predominance of linear pollutant-meteorological relationships. ANN exhibited moderate-to-strong performance ($R^2 = 0.53$–0.74), particularly in Pasir Gudang, where industrial complexity introduced nonlinear emission behaviors. Granger causality contributed supplementary insights into temporal dependencies but yielded lower predictive accuracy. Comparative outcomes demonstrated clear spatial variability, with Nilai and Larkin favoring linear modeling due to stable emission patterns, while Pasir Gudang benefited from ANN's nonlinear adaptability. This study underscores the importance of combining correlation-driven linear modeling with data-intelligent nonlinear approaches to improve air-quality forecasting reliability in Malaysia. The findings provide valuable guidance for data-driven environmental policy and the design of hybrid predictive frameworks for sustainable urban air management in Southeast Asia.**

***Keywords*—air pollution modeling, air quality prediction, granger causality, Pearson correlation, artificial neural network**

I. Introduction

Air pollution continues to be a significant issue in Malaysia, driven by rapid urbanization, industrial expansion, and increased vehicle emissions. The World Health Organization (WHO) has continuously recognized particulate matter (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO) as principal pollutants associated with respiratory diseases, diminished visibility, and climate change [1]. Urban industrial hubs like Larkin and Pasir Gudang sometimes exhibit Air Pollutant Index (API) values surpassing prescribed thresholds, especially during arid seasons when air dispersion is limited [2].

Precise forecasting of air pollution levels is crucial for air quality management, policy development, and early warning systems. Traditional deterministic models like Gaussian plume, AMS/EPA Regulatory Model (AERMOD), and California Puff Model (CALPUFF) necessitate comprehensive emission inventories and meteorological data, which are frequently inadequate in Malaysian environments [3]. As a result, statistical and machine-learning methodologies have been prominent for their capacity to discern correlations between contaminants and forecast future concentrations utilizing observational data.

Pearson correlation, Granger causality, and Artificial Neural Networks (ANN) provide complementary analytical advantages among data-driven methodologies. The Pearson correlation evaluates linear associations between pollutant concentrations and climatic variables [4], whereas Granger causality offers temporal and directional analysis to ascertain if one time series can predict another [5]. Conversely, ANN models excel in capturing nonlinear and dynamic interactions, frequently surpassing classic statistical models in predictive accuracy [6]. Nevertheless, limited research has methodically contrasted these three methodologies within the Malaysian context across several locations exhibiting diverse industrial and urban attributes. Prior studies have predominantly concentrated on individual pollutants or specific locations, exemplified by PM_{10} modeling in Kuala Lumpur [7] and ozone forecasts in Penang [8].

This study seeks to address that deficiency by assessing the prediction efficacy of correlation-based, causality-based, and neural network models in three Malaysian cities: Nilai, Larkin, and Pasir Gudang. The aim is to evaluate model resilience across diverse urban, industrial, and climatic environments while offering practical recommendations for environmental management. Unlike previous Malaysian air-quality studies that typically focus on single cities, limited pollutants, or a single modeling approach, this study presents a systematic multi-city comparison of correlation-based, causality-based, and full-parameter input-selection strategies for both Linear (Multiple Linear Regression (MLR)) and Nonlinear (ANN) models using consistent datasets and evaluation metrics. This study specifically assesses and contrasts the prediction efficacy of Pearson correlation, Granger causality, Multiple Linear Regression (MLR), and ANN models for concentrations of PM_{10}, SO_2, NO_2, O_3, and CO. By explicitly benchmarking Pearson correlation and Granger causality as variable-selection mechanisms across three contrasting urban industrial environments, this work clarifies their respective predictive and interpretive roles. It examines how the spatial diversity of model performance among these cities can inform data-driven air quality control policies in Malaysia.

II. Literature Review

The air quality of Malaysia is influenced by urbanization, industrialization, and transboundary haze incidents from adjacent areas [9]. Industrial centres like Pasir Gudang frequently exhibit elevated levels of SO_2 and PM_{10}, whereas congested traffic routes in Larkin and Nilai substantially contribute to CO and NO_2 emissions [10]. The Malaysian Department of Environment (DOE) consistently monitors ambient air via its countrywide network, delivering real-time data for PM_{10}, SO_2, NO_2, O_3, and CO [11]. Nevertheless, deficiencies in monitoring coverage and absent data require the implementation of modeling techniques to guarantee accurate predictions and discern inter-pollutant relationships.

The Pearson correlation is one of the most straightforward ways to quantify linear relationships between contaminants and meteorological data. Recent research in Kuala Lumpur and Penang has demonstrated moderate-to-high relationships (r = 0.6–0.8) between carbon monoxide and temperature, as well as between particulate matter PM_{10} and wind speed [12]. Nonetheless, correlation does not imply causation and is constrained in elucidating temporal dynamics [13]. Granger causality analysis provides temporal insights by assessing whether historical values of one variable might forecast another. A 2023 study by Istiana *et al.* [14] in Aerosol and Air Quality Research revealed substantial Granger correlations between meteorological parameters and $PM_{2.5}$ in Jakarta, underscoring its regional significance. Likewise, Ping *et al.* [15] utilized Granger causality on Malaysian pollutant data, demonstrating that PM_{10} frequently Granger-causes SO_2 in industrial areas. Artificial Neural Networks (ANNs) have demonstrated significant efficacy in air-quality forecasting owing to their capacity to discern nonlinear correlations within extensive datasets. Ul-Saufie *et al.* [16] and Hameed *et al.* [17] highlighted the improved predictive capabilities of ANNs in comparison to linear regression models. Recent advancements, including deep learning and transfer learning frameworks [18], significantly improve the adaptability of artificial neural networks across various monitoring stations.

Hybrid methodologies that combine statistical pre-processing with artificial neural network topologies are emerging. Ul-Saufie *et al.* [16] employed feature selection to enhance PM_{10} prediction in Malaysia, whereas Sousa *et al.* [19] showed that the integration of correlation-based variable screening with ANN training markedly increased ozone forecasting precision. Comparative modeling research examining numerous contaminants and cities is scarce in Malaysia. This study integrates interpretive (correlation and causality) and predictive (Artificial Neural Network) frameworks utilizing actual monitoring data from three urban industrial areas, facilitating both scientific comprehension and practical implementation in air quality management.

III. Materials and Methods

This study was designed to compare the predictive accuracy and interpretative capability of several modeling approaches for air-pollutant forecasting in Malaysian urban environments. The methodological framework integrates data acquisition, pre-processing, variable selection, and model development using both statistical and machine-learning techniques. The overall workflow involved five sequential stages:

(1) Acquisition and compilation of pollutant and meteorological data
(2) Data quality screening and normalization
(3) Implementation of Pearson correlation and granger causality analyses to examine inter-pollutant and temporal relationships
(4) Construction and training of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) models
(5) Model evaluation using standardized statistical performance indicators.

This structured framework ensured consistency and reproducibility across all study locations. The integration of statistical and machine-learning methods provided both interpretability and predictive accuracy—where Pearson and Granger analyses identified key influencing factors, and MLR and ANN models quantified linear and nonlinear dependencies between pollutant concentrations and meteorological drivers.

A. Study Area

The research includes three cities in Malaysia: Nilai (Negeri Sembilan), Larkin (Johor Bahru), and Pasir Gudang (Johor Bahru). Each city embodies a unique mix of urbanization degree, industrial engagement, and climatic conditions, originating from mixed residential-industrial areas. Daily average concentrations of PM_{10}, SO_2, NO_2, O_3, and CO for the years 2017–2021 were sourced from the Department of Environment (DOE) Malaysia. Meteorological characteristics such as air temperature, relative humidity, and wind speed were included to depict meteorological variability. Data completeness was above 95%, while absent values (<5%) were estimated by linear interpolation. All variables were standardized to a 0–1 scale prior to model training to guarantee numerical stability and comparability among predictors. The summary of daily observations after quality control for the years 2017–2021 is presented in Table 1.

Table 1. Summary of daily observations after quality control (2017–2021)

City	Initial days	Valid days used	Missing days (%)	Missing-data treatment
Nilai	1825	1641	10%	Rows with missing values removed
Larkin	1825	1641	10%	Rows with missing values removed
Pasir Gudang	1825	1641	10%	Rows with missing values removed

B. Modeling Framework

Four modeling approaches were developed and compared:

(1) Pearson Correlation (PC): Used to assess linear dependencies among pollutants and meteorological parameters.
(2) Granger Causality (GC): Applied to determine temporal dependencies and lagged influences.
(3) Multiple Linear Regression (MLR): Served as a baseline statistical model based on linear relationships.
(4) Artificial Neural Network (ANN): A feedforward Multilayer Perceptron (MLP) trained through backpropagation to capture nonlinear interactions.

Granger causality analysis was applied to identify temporal and directional dependencies between air pollutants and meteorological parameters. The analysis was conducted using a bivariate time-series framework to examine whether historical values of one variable significantly improved the prediction of another. Prior to analysis, all time series were tested for stationarity using the Augmented Dickey–Fuller (ADF) test. Non-stationary series were differenced once to achieve stationarity before further testing. Optimal lag length was determined using the Akaike Information Criterion (AIC), with lag structures limited to short-term dependencies to reflect atmospheric response times. Statistical significance was evaluated at the 5% confidence level. The results of the Granger causality analysis were primarily used to support interpretive understanding of pollutant–meteorological interactions and as a variable-selection mechanism for subsequent predictive modeling. This implementation is consistent with established environmental applications of Granger causality in Malaysia and comparable tropical regions.

Artificial Neural Network (ANN) models were developed using a feedforward Multilayer Perceptron (MLP) architecture to capture nonlinear relationships between air pollutant concentrations and meteorological parameters. The ANN structure consisted of an input layer representing selected pollutant and meteorological variables, three hidden layers, and a single output neuron corresponding to the predicted pollutant concentration. Sigmoid activation functions were applied to all hidden layers, while a linear activation function was used in the output layer to support continuous-valued regression.

Model training was performed using the backpropagation algorithm with a Mean Squared Error (MSE) loss function and an adaptive learning-rate optimizer. To minimize overfitting, early stopping was implemented based on validation loss convergence. The dataset was partitioned using a time-ordered split, with 70% of the earliest observations used for training and 30% of the most recent observations reserved for validation. Hyperparameter configurations were determined through preliminary trial experiments to ensure model stability and generalization. This ANN implementation follows a validated modeling framework previously applied to Malaysian air-quality datasets, ensuring methodological robustness and reproducibility.

In this study, k-fold cross-validation is employed, a widely recognized method where the dataset is divided into k subsets or "folds". The model is trained k times, each time using a different fold as the validation set and the remaining $k-1$ folds as the training set. The average performance across all folds is then calculated, providing an unbiased estimate of the model's accuracy. This approach reduces the risk of overfitting by ensuring that each data point is used for both training and validation, which enhances the model's reliability.

The MLR models were refined by a stepwise regression approach in SPSS to reduce multicollinearity and preserve only statistically relevant factors. Artificial Neural Network configurations in RapidMiner were established with three hidden layers, sigmoid activation functions, and an adaptive learning rate optimizer. To preserve temporal dependency and prevent data leakage, all datasets were divided using a time-ordered split. The earliest 70% of observations were used for model training, while the remaining 30% were reserved for independent validation. Variable selection, including Pearson correlation filtering and Granger causality screening, was performed exclusively on the training dataset. Model performance metrics were computed using validation data only, ensuring unbiased evaluation of predictive accuracy. Data preprocessing, correlation analysis, and regression modeling were performed using IBM SPSS Statistics 29, while RapidMiner Studio 10.3 was employed for the training, validation, and testing of the ANN model. These platforms were chosen for their dependability and integrated analytical functionalities, facilitating the smooth execution of both statistical and machine-learning tasks.

C. Model Evaluation Metric

Model performance was evaluated using four statistical indices: Root Mean Square Error (RMSE), Mean Error (ME), Normalized Absolute Error (NAE), and the Coefficient of Determination (R^2), as expressed by the following equations in Table 2.

Table 2. Performance indicators

Performance Indicators	Equation	Criteria
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$	The best model has the values closer to zero or the smallest values.
Mean Error	$ME = {}^{1}/_{n}\sum_{i=1}^{n}\lvert P_i - O_i\rvert$	
Normalized Absolute Error	$NAE = \left(\frac{MAE}{\bar{O}}\right)$	
Coefficient of Determination	$R^2 = \left({}^{1}/_{n}\frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{S_{pred} \cdot S_{obs}}\right)^2$	The best model has the values closer to 1.

n is the number of observations data; O_i is the observed data; $\bar{O}$ is the mean of the observed data; P_i is the predicted data; $\bar{P}$ is the mean of the predicted data; S_{obs} is the standard deviation of observed data, and S_{pred} is the standard deviation of predicted data.

Lower values of RMSE, ME, and NAE indicate higher model accuracy, whereas R^2 values closer to 1 represent better model fit. An NAE value approaching 0 denotes minimal deviation between observed and predicted concentrations, reflecting the most accurate model performance. All error metrics were cross-validated to ensure internal consistency between absolute, normalized, and variance-based indicators.

D. Variable Selection Techniques

To identify the most influential predictors for PM_{10} forecasting, three complementary variable-selection strategies were applied: All-Parameter inclusion (PM_{10}, SO_2, NO_2, O_3, CO, relative humidity, wind speed, and temperature), Pearson correlation filtering, and Granger causality-based selection. These methods ensured that linear, nonlinear, and temporal relationships were effectively represented in the modeling process.

The All-Parameter approach incorporated all available meteorological and pollutant variables into the models without exclusion. This inclusive method prevents the omission of potentially valuable information, allowing models, particularly ANNs, to explore the full spectrum of predictor–response relationships [20]. However, this approach may also introduce redundant predictors, increase statistical noise, and cause overfitting in ANNs or multicollinearity in MLR models.

The Pearson correlation method isolated predictors with the strongest linear relationships to PM_{10} by computing correlation coefficients between each variable and the target pollutant [21]. This filtering process reduced dimensionality, minimized redundancy, and improved computational efficiency. However, Pearson correlation cannot capture time-lagged effects, which are common in atmospheric systems [22], such as delayed PM_{10} responses to changes in humidity or wind speed.

The Granger causality analysis identified predictors with significant temporal precedence affecting PM_{10} concentrations. Unlike correlation, Granger causality determines whether past values of a variable can improve the forecast of another, making it particularly useful for time-series environmental data [21]. For example, elevated temperature or precursor gas concentrations on a given day may lead to higher PM_{10} levels on subsequent days due to secondary particle formation. Despite its strengths, Granger causality assumes linear relationships and may be influenced by confounding variables, requiring careful interpretation.

By integrating these three selection strategies, this study systematically evaluated how variable filtering affects both linear (MLR) and nonlinear (ANN) model performance. The All-Parameter configuration established a broad baseline; Pearson correlation offered a streamlined, linear-focused subset; and Granger causality incorporated the temporal dimension. Collectively, these approaches form a robust methodological framework for assessing the sensitivity, adaptability, and predictive efficiency of PM_{10} forecasting models under various predictor configurations [23].

IV. Result and Discussion

This section discusses and analyzes the findings of the comparative modeling assessment performed in the three chosen Malaysian cities: Nilai, Larkin, and Pasir Gudang. The results are analyzed regarding the predicted accuracy, robustness, and interpretability of each model. Statistical indices such as RMSE, ME, NAE, and R^2 were employed to assess model performance, where diminished RMSE, ME, and NAE values, alongside elevated R^2 values, signify enhanced predictive efficacy. The analysis delineates the strengths and weaknesses of each modelling technique (Pearson correlation, Granger causality, MLR, and ANN) in predicting the concentrations of PM_{10}, SO_2, NO_2, O_3, and CO under varying meteorological and emission conditions.

Granger causality analysis revealed consistent temporal relationships between selected meteorological variables and pollutant concentrations across the three study locations. Wind speed frequently exhibited unidirectional causality with PM_{10}, indicating its dominant role in pollutant dispersion processes. In several cases, NO_2 was identified as a Granger-causal precursor to O_3, reflecting established photochemical formation mechanisms. These causal linkages provided interpretive insight into pollutant dynamics and informed variable selection for predictive modeling, although Granger causality-based models generally demonstrated lower predictive accuracy compared to correlation-based approaches. Table 3 shows the summary of significant Granger-causal relationships between meteorological variables, gaseous pollutants, and PM_{10}.

Table 3. Significant Granger-causal relationships between meteorological variables, gaseous pollutants, and PM_{10} ($p < 0.05$) were found

Summary	Nilai	Larkin	Pasir Gudang
Predictor variable	SO_2	NO_2	SO_2
Targeted variable	PM_{10}	PM_{10}	PM_{10}
Optimal lag (days)	1	1	1
F-statistic	17.3825	7.4099	8.5181
p-value	<0.05	<0.05	<0.05
Causality type	Unidirectional	Unidirectional	Bidirectional

A. Performance of ANN and MLR for Nilai, Larkin, and Pasir Gudang

1) Nilai

Table 4 indicates the performance of the ANN and MLR models in forecasting PM_{10}, SO_2, NO_2, O_3, and CO concentrations in Nilai, assessed using RMSE, MAE, NAE, and R^2 metrics.

- Overall performance: Multiple Linear Regression (MLR) consistently outperformed ANN across all pollutants and variable-selection strategies, particularly when Pearson correlation was applied.
- PM_{10}: MLR achieved very high accuracy under Pearson correlation ($R^2 = 0.9918$, RMSE = 4.51), indicating that PM_{10} variability in Nilai is governed by stable, linearly related climatic and emission factors. ANN performance was substantially lower and sensitive to redundant inputs.
- SO_2: Multiple linear regression demonstrated robust predictive capability (R^2 = 0.9890–0.9936), reflecting predictable linear behavior likely driven by industrial and combustion sources. ANN showed negligible predictive skill ($R^2 \approx 0.01$), suggesting limited nonlinear dynamics.
- NO_2: Multiple linear regression achieved strong accuracy across both All Parameter and Pearson configurations (R^2 = 0.9566–0.9688), consistent with traffic-driven linear emission patterns. ANN performed moderately with All Parameters but degraded significantly when causality-based inputs were prioritized.
- O_3: MLR performed exceptionally well under Pearson correlation ($R^2 = 0.9936$, RMSE = 0.0013), indicating ozone formation dominated by stable photochemical processes. ANN struggled due to numerical instability and narrow concentration ranges.

- CO: MLR consistently delivered superior performance, particularly with Pearson correlation (R^2 = 0.9936, RMSE = 0.0405), confirming strong linearity with vehicular emissions. ANN showed moderate accuracy but declined under Granger-based inputs.
- Variable selection effects: Pearson correlation improved model performance by removing irrelevant predictors and emphasizing linear relationships. Granger causality provided limited predictive benefit for short-term datasets, while the All-Parameter approach introduced redundancy that primarily affected ANN performance.

Nilai exhibits predominantly linear air-quality behavior, making MLR the most suitable predictive model (R^2 > 0.98 for all major pollutants). ANN requires careful variable filtering to achieve stable results, aligning with findings from prior studies in moderately urbanized environments. The results correspond with earlier research by Ul-Saufie *et al.* [16] and Shams *et al*. [24], indicating that MLR is most effective in moderately urbanized regions with stable pollution patterns, whereas ANN is advantageous mostly in highly dynamic, nonlinear environments.

Table 4. Performance indicators for the ANN and MLR model of PM_{10}, SO_2, NO_2, O_3, and CO concentration at Nilai

Pollutant	Model	Variables Selection	RMSE	ME	NAE	R^2
PM_{10}	ANN	All Parameters	9.9230	4.9033	0.7140	0.7131
		Pearson Correlation	10.7870	2.4177	0.7650	0.0056
		Granger Causality	15.3340	1.8210	0.9470	0.0062
	MLR	All Parameters	4.8488	2.9745	0.0829	0.9819
		Pearson Correlation	4.5113	2.7423	0.0764	0.9918
		Granger Causality	19.6722	11.5471	0.3219	0.2253
SO_2	ANN	All Parameters	0.0010	0.0000	1.0570	0.0127
	MLR	All Parameters	0.0005	0.0005	0.4357	0.9890
		Pearson Correlation	0.0002	0.0002	0.1170	0.9936
NO_2	ANN	All Parameters	0.0030	0.0002	0.6410	0.5542
		Pearson Correlation	0.0030	0.0003	0.7740	0.0408
		Granger Causality	0.0030	0.0015	0.8330	0.0689
	MLR	All Parameters	0.0135	0.0134	1.0615	0.9566
		Pearson Correlation	0.0012	0.0009	0.0749	0.9688
		Granger Causality	0.0042	0.0034	0.2712	0.3180
CO	ANN	All Parameters	0.1010	0.0181	0.6660	0.6223
		Granger Causality	0.1150	0.0212	0.7680	0.0003
	MLR	All Parameters	0.0611	0.0457	0.0756	0.9820
		Pearson Correlation	0.0405	0.0278	0.0460	0.9936
		Granger Causality	0.1297	0.0965	0.1598	0.5854
O_3	ANN	All Parameters	0.0040	0.0020	1.0030	0.3353
		Granger Causality	0.0050	0.0017	0.9370	0.0043
	MLR	All Parameters	0.0111	0.0109	1.1510	0.8198
		Pearson Correlation	0.0013	0.0011	0.4203	0.9936
		Granger Causality	0.0048	0.0040	0.1200	0.0227

2) Larkin

Table 5 explores the performance metrics of the model for forecasting PM_{10}, SO_2, NO_2, O_3, and CO concentrations at Larkin utilizing ANN and MLR across three variable-selection methodologies. The performance was assessed using RMSE, MAE, NAE, and R^2 metrics.

- Overall performance: Multiple Linear Regression (MLR) consistently outperformed Artificial Neural Networks (ANN), particularly under the Pearson correlation framework, indicating that air quality variability in Larkin is largely governed by linear climatic and emission-related factors.
- PM_{10}: MLR achieved excellent predictive accuracy with Pearson correlation (R^2 = 0.9907, RMSE = 3.29), confirming strong linear control by meteorological dispersion factors such as wind speed and temperature. ANN captured limited nonlinear behavior but showed reduced stability when redundant inputs were included.
- SO_2: MLR demonstrated exceptionally strong performance (R^2 = 0.9938, RMSE = 0.0005), reflecting stable, linearly driven emissions from industrial and combustion sources. ANN performance was weak (R^2 = 0.1595), indicating minimal nonlinear dynamics for SO_2 in Larkin.
- NO_2: MLR showed consistently high accuracy across all configurations (R^2 = 0.9730–0.9804), highlighting dominant linear relationships with traffic emissions and temperature. ANN performed satisfactorily only under the All-Parameter setup and degraded substantially under Granger causality inputs.
- O_3: MLR yielded superior results under Pearson correlation (R^2 = 0.9265, RMSE = 0.0123), suggesting ozone levels are primarily regulated by linear photochemical processes. Artificial Neural Networks demonstrated moderate capability but failed to fully capture complex precursor–meteorology interactions.
- CO: MLR consistently delivered high prediction accuracy (R^2 = 0.9854–0.9936), confirming strong linear dependence on vehicular emissions and dispersion conditions. ANN showed moderate accuracy but was sensitive to noise and lag-based inputs.
- Variable selection effects: Pearson correlation produced the best results by filtering irrelevant variables, while the All-Parameter configuration yielded balanced but slightly weaker performance. Granger causality generally reduced predictive accuracy, especially for ANN, due to weak short-term lag effects.

Air-quality behavior in Larkin is predominantly linear and meteorology-dependent, making MLR the most suitable predictive model. ANN provided reasonable performance for PM_{10} and O_3 but exhibited reduced stability in high-dimensional input settings, consistent with findings in other urban environments with stable emission patterns.

In conclusion, MLR was the most effective model for all pollutants at Larkin, with average R^2 values surpassing 0.97 for NO_2, SO_2, and CO. Artificial Neural Networks demonstrated satisfactory performance for PM_{10} and O_3, although they exhibited reduced stability under high-dimensional input settings. The results indicate that air-quality behavior in Larkin is mostly influenced by linear, meteorology-dependent interactions rather than significant nonlinear or delayed dynamics. The results align with those shown by Ul-Saufie *et al.* [16] and Shams *et al.* [24], who identified MLR as especially beneficial in urban areas characterized by stable traffic and industrial emission patterns.

Table 5. Performance indicators for the ANN and MLR model of PM_{10}, SO_2, NO_2, O_3, and CO concentration at Larkin

Pollutant	Model	Variable Selection	RMSE	ME	NAE	R^2
PM_{10}	ANN	All Parameters	6.0100	0.1551	0.6160	0.6360
		Pearson Correlation	7.6350	0.4815	0.7570	0.4160
		Granger Causality	7.0070	1.8182	0.7310	0.5390
	MLR	All Parameters	5.1803	2.5090	0.0862	0.8478
		Pearson Correlation	3.2882	2.0230	0.0695	0.9907
		Granger Causality	11.0455	6.1137	0.2099	0.2836
SO_2	ANN	All Parameters	0.0010	0.0036	0.9170	0.1595
	MLR	All Parameters	0.0004	0.0003	0.1686	0.9403
		Pearson Correlation	0.0005	0.0004	0.2200	0.9938
		Granger Causality	0.0009	0.0007	0.4017	0.0105
NO_2	ANN	All Parameters	0.0020	0.0048	0.5270	0.7086
		Pearson Correlation	0.0030	0.0090	0.7570	0.4323
		Granger Causality	0.0040	0.0093	0.9730	0.0464
	MLR	All Parameters	0.0031	0.0029	0.2489	0.9730
		Pearson Correlation	0.0013	0.0010	0.0869	0.9804
		Granger Causality	0.0045	0.0036	0.3045	0.0120
CO	ANN	All Parameters	0.1680	0.7139	0.6520	0.5617
		Granger Causality	0.2030	0.5863	0.8050	0.3526
	MLR	All Parameters	0.0556	0.0470	0.0741	0.9854
		Pearson Correlation	0.0366	0.1786	0.0479	0.9936
		Granger Causality	0.2159	0.0304	0.2815	0.3872
O_3	ANN	All Parameters	0.0040	0.0034	0.6480	0.5612
		Pearson Correlation	0.0050	0.0052	0.8230	0.2708
		Granger Causality	0.0040	0.0064	0.6410	0.5607
	MLR	All Parameters	0.0117	0.0113	0.7435	0.8796
		Pearson Correlation	0.0123	0.0119	0.7843	0.9265
		Granger Causality	0.0344	0.0340	2.2440	0.3862

3) Pasir Gudang

Table 6 displays the performance assessment of ANN and MLR models for forecasting PM_{10}, SO_2, NO_2, O_3, and CO concentrations in Pasir Gudang, utilizing All Parameter, Pearson correlation, and Granger causality selection methodologies.

- Overall performance: Multiple Linear Regression (MLR) consistently outperformed Artificial Neural Networks (ANN) for all pollutants, confirming that air quality variability in Pasir Gudang is predominantly governed by linear relationships between industrial emissions and climatic factors.
- PM_{10}: MLR achieved high predictive accuracy under both All Parameter and Pearson correlation configurations (R^2 = 0.9716–0.9767, RMSE = 2.71–2.85), indicating strong linear dependence on wind speed, humidity, and industrial dust emissions. ANN demonstrated moderate performance but exhibited signs of overfitting under reduced input sets.
- SO_2: MLR performed exceptionally well (R^2 = 0.9890–0.9936, RMSE = 0.0002–0.0005), reflecting stable and predictable industrial emission patterns. ANN showed weak and inconsistent performance (R^2 = 0.27–0.48), indicating limited nonlinear dynamics for SO_2.
- NO_2: MLR achieved near-perfect accuracy across all configurations (R^2 = 0.9925–0.9994), highlighting strong linear associations with industrial combustion and vehicular emissions. ANN captured some nonlinear variability under All Parameter inputs but lost robustness when predictor sets were reduced.
- O_3: MLR demonstrated outstanding predictive capability under Pearson correlation (R^2 = 1.0000, RMSE = 0.0097), indicating ozone formation dominated by stable photochemical processes. ANN showed moderate performance but suffered reduced generalization under simplified input structures.
- CO: MLR consistently produced the highest accuracy (R^2 = 1.0000, RMSE = 0.0461), confirming the linear and predictable nature of carbon monoxide emissions. ANN exhibited modest predictive skill and was sensitive to noise and lag-based inputs.
- Variable selection effects: Pearson correlation yielded the strongest performance by prioritizing linearly influential predictors, while the All-Parameter configuration enhanced ANN learning at the cost of increased noise. Granger causality generally reduced predictive accuracy due to weak lagged dependencies in the dataset.

Air-quality behavior in Pasir Gudang is primarily linear and industrially driven, making MLR the most reliable modeling approach. ANN identified limited nonlinear behavior for PM_{10} and NO_2 but lacked stability under variable-reduction and temporal causality settings, consistent with findings in other industrialized environments. Although ANN demonstrated efficacy in identifying some nonlinear patterns for PM_{10} and NO_2, it exhibited instability when exposed to varying reduction or temporal causality assessments. The findings align with those of Shams *et al.* [24] and Abdullah *et al.* [25], indicating that MLR is more effective in industrialized contexts with pronounced deterministic emission patterns, whereas ANN excels in dynamic or mixed urban environments.

Table 6. Performance indicators for the ANN and MLR models of PM_{10}, SO_2, NO_2, O_3, and CO concentrations at Pasir Gudang

Pollutant	Model	Variable Selection	RMSE	ME	NAE	R^2
PM_{10}	ANN	All Parameters	6.4420	1.5602	0.5860	0.6315
		Pearson Correlation	8.2590	1.8780	0.7560	0.3884
	MLR	All Parameters	2.8539	1.9841	0.0742	0.9716
		Pearson Correlation	2.7074	1.8787	0.0716	0.9767
SO_2	ANN	All Parameters	0.0010	0.0004	1.0250	0.4790
		Pearson Correlation	0.0010	0.0003	0.9400	0.2797
		Granger Causality	0.0010	0.0000	0.8270	0.2681
	MLR	All Parameters	0.0005	0.0005	0.4357	0.9890
		Pearson Correlation	0.0002	0.0002	0.1170	0.9936
NO_2	ANN	All Parameters	0.0030	0.0009	0.5370	0.7415
		Pearson Correlation	0.0050	0.0018	0.8820	0.4169
		Granger Causality	0.0040	0.0004	0.6660	0.5781
	MLR	All Parameters	0.0059	0.0058	0.5102	0.9925
		Pearson Correlation	0.0015	0.0012	0.1052	0.9994
		Granger Causality	0.0053	0.0041	0.3649	0.2973
CO	ANN	All Parameters	0.1160	0.0006	0.6760	0.5392
		Granger Causality	0.1280	0.0249	0.7520	0.4557
	MLR	All Parameters	0.0585	0.0489	0.0743	0.9864
		Pearson Correlation	0.0461	0.0373	0.0566	1.0000
		Granger Causality	0.1329	0.1070	0.1624	0.6255
O_3	ANN	All Parameters	0.0050	0.0008	0.6490	0.5294
		Pearson Correlation	0.0060	0.0001	0.9240	0.1244
		Granger Causality	0.0050	0.0007	0.8020	0.4013
	MLR	All Parameters	0.0102	0.0098	0.6760	0.9504
		Pearson Correlation	0.0097	0.0095	0.6573	1.0000
		Granger Causality	0.0289	0.0282	1.9575	0.4500

B. Comparative Synthesis of Model Performance across Nilai, Larkin, and Pasir Gudang

The comparative assessment of ANN and MLR performance in Nilai, Larkin, and Pasir Gudang demonstrates consistent geographical patterns that highlight the impact of local emission sources, urban shape, and meteorological variability on model performance (Table 7).

Overall, the comparative results indicate that air-quality behavior across all three Malaysian study sites is largely linear in nature, supporting the superior performance of MLR models, while ANN provides complementary insight into limited nonlinear dynamics under specific conditions.

In Nilai, MLR demonstrated exceptional efficacy for all pollutants, especially PM_{10}, SO_2, and NO_2, underscoring the preeminence of linear correlations between emission sources and air dispersion mechanisms. The mild industrial activity and consistent meteorological conditions in Nilai facilitated linear modeling techniques, enabling multiple linear regression to surpass artificial neural networks in all configurations. The robust R^2 values (0.9819–0.9936) in the Pearson correlation indicate that fluctuations in pollutants were influenced by stable traffic emissions and consistent weather conditions. Although ANN can identify nonlinear oscillations, it shows heightened sensitivity to noise, especially under the Granger causality framework.

In Larkin, MLR consistently showed superior performance, with R^2 values surpassing 0.97 for NO_2, SO_2, and CO, signifying robust linear relationships between pollutants and climatic factors. The city's traffic-centric emissions, coupled with limited industrial fluctuations, enabled MLR to accurately estimate air pollution dynamics. ANN had considerable predictive capability for PM_{10} and O_3, attaining R^2 values over 0.60, indicating mild nonlinearities at peak hours or under fluctuating meteorological conditions. A variable-selection study indicated that Pearson correlation enhanced the performance of both models by eliminating extraneous inputs, whereas Granger causality offered no predictive improvement owing to the negligible impact of lagged pollutant linkages in this urban context.

Table 7. Comparative summary of MLR and ANN performance across study sites

Summary	Nilai	Larkin	Pasir Gudang
Dominant model	MLR	MLR	MLR
Best variable selection	Pearson Correlation	Pearson Correlation	Pearson Correlation
Typical R^2 range (MLR)	>0.98 (all pollutants)	0.97–0.99	0.97–1.00
Pollutants with the highest accuracy	PM_{10}, SO_2, NO_2, O_3, CO	PM_{10}, SO_2, NO_2, CO	CO, O_3, NO_2, SO_2
ANN behavior	Moderate for PM_{10} & CO; unstable under reduced inputs	Reasonable for PM_{10} & O_3; sensitive to noise and lag-based inputs	Captured limited nonlinearities for PM_{10} & NO_2; reduced robustness
Dominant air quality characteristics	Predominantly linear, traffic- and climate-driven pollution	Linear, meteorology-dependent urban emission patterns	Strongly linear, industrially dominated emission behavior

In Pasir Gudang, an industrialized coastal city, the dynamics of pollutants were intricate due to the interplay of industrial emissions, marine air currents, and atmospheric dispersion. Although MLR exhibited remarkable accuracy ($R^2 > 0.97$ for all pollutants), ANN outperformed the other two cities, especially for PM_{10} and NO_2, achieving R^2 values between 0.63 and 0.74. Although MLR achieved higher overall accuracy in Pasir Gudang, the moderate ANN performance for PM_{10} and NO_2 suggests the presence of secondary nonlinear dynamics superimposed on a dominant linear structure. These nonlinear components likely reflect episodic industrial releases and coastal meteorological variability, which ANN partially captured despite overall lower predictive consistency. This enhancement signifies that the ANN was more adept at catching nonlinear variations induced by sporadic industrial effluents and fluctuating meteorological conditions. Nevertheless, the comprehensive model comparison validated that linear patterns remained predominant, as evidenced by the superior performance of MLR in both All Parameters and Pearson correlation configurations.

In all three locations, Pearson correlation proved to be the most efficient variable-selection method, consistently enhancing the accuracy of both ANN and MLR by emphasizing meteorological and pollutant predictors with robust linear correlations. The All-Parameter strategy provided extensive coverage but sometimes included superfluous inputs that impaired the ANN's generalization, whereas the Granger causality method typically underachieved, indicating the restricted temporal lag effect in short-term pollutant datasets.

The combined data indicate that Malaysia's urban and industrial areas display primarily linear connections between pollutants and climatic factors, rendering multiple linear regression a useful and interpretable method for air quality prediction. In regions characterized by intricate industrial processes and fluctuating emissions, such as Pasir Gudang. Artificial Neural Networks (ANNs) are essential in identifying nuanced nonlinear dynamics that linear models could miss. It is advisable to integrate correlation-based feature selection with hybrid linear–nonlinear modeling frameworks for future nationwide air-quality evaluations, since this approach ensures a balance of interpretability, computational efficiency, and predictive robustness across various environmental scenarios.

V. Conclusion

This study evaluated the predictive performance of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) models for five major air pollutants (PM_{10}, SO_2, NO_2, O_3, and CO) across three Malaysian cities using Pearson correlation, Granger causality, and all-parameter variable-selection strategies to balance interpretability, computational efficiency, and robustness given the relatively short temporal coverage of the dataset.

For the cities and five-year period analyzed, MLR consistently demonstrated superior predictive accuracy, particularly when combined with Pearson correlation-based variable selection, indicating the dominance of linear pollutant–meteorological relationships under relatively stable emission conditions.

ANN models exhibited moderate predictive capability, especially in Pasir Gudang, where industrial activity and coastal meteorology introduced additional nonlinear variability. These results suggest that while linear relationships dominate overall, secondary nonlinear dynamics may still contribute to pollutant behavior in complex urban–industrial environments. Granger causality analysis provided valuable interpretive insights into temporal dependencies but offered limited improvement in short-term predictive accuracy.

Overall, the findings indicate that, within the scope of the models, locations, and temporal coverage examined, correlation-driven linear modeling remains an effective and interpretable approach for urban air-quality prediction in Malaysia, while ANN serves as a complementary tool for capturing nonlinear effects where emission patterns are more dynamic. While these models remain widely used in air-quality applications, more advanced machine-learning techniques such as attention-based architectures, recurrent neural networks (e.g., long short-term memory), Generative Adversarial Network (GAN)-augmented learning for data imbalance, tree-based ensemble methods, and transformer-based tabular models have recently demonstrated strong performance in closely related fault detection and diagnosis (AFDD) studies. The exclusion of these advanced approaches represents a limitation of the present work.

Future research will extend the proposed framework by integrating such methods alongside the existing variable-selection strategies, enabling direct comparison between traditional regression, shallow neural networks, and modern deep-learning architectures. Incorporating these techniques is expected to improve robustness under complex nonlinear dynamics, enhance cross-site transferability, and better capture temporal dependencies in Malaysian air-quality systems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, N.R.; methodology, Z.A.R.; software, N.R. and Z.A.R.; validation, N.M.N. and A.Z.U.; formal analysis, Z.A.R.; investigation, N.R. and Z.A.R.; resources, N.M.N.; data curation, A.Z.U. and H.A.H.; writing—original draft preparation, N.R.; writing—review and editing, Z.A.R. and N.R.; project administration, M.K.N.M.; funding acquisition, N.R. and N.M.N. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Ministry of Higher Education (MoHE) Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2023/TK08/UNIMAP/02/24).

Acknowledgment

We extended our thanks to the Department of Environment (DOE) Malaysia for granting access to the comprehensive and detailed datasets essential to the analysis and findings.

References

[1] World Health Organization, *Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease*, Geneva, 2021.

[2] Department of Environment Malaysia, *Clean Air Action Plan (CAAP) 2021–2030*, Putrajaya, 2021.

[3] Y. Huang *et al.*, "Investigating Southeast Asian biomass burning by the WRF-CMAQ two-way coupled model: Emission and direct aerosol radiative effects," *Atmospheric Environment*, vol. 294, 119521, 2023.

[4] P. Chaloulakou, A. Kassomenos, and N. Spyrellis, "Measurement of PM_{10} and $PM_{2.5}$ in Athens," *Atmospheric Environment*, vol. 37, pp. 649–660, 2003.

[5] A. Z. Ul-Saufie, M. Norlela, and N. Rahim, "Modeling of PM_{10} with regression and Gaussian dispersion," *Science of the Total Environment*, vol. 458–460, pp. 16–27, 2013.

[6] M. Gardner and S. Dorling, "Neural network modelling of urban NOx," *Atmospheric Environment*, vol. 33, pp. 709–719, 1999.

[7] A. Z. Ul-Saufie, A. S. Yahya, N. A. Ramli, N. Rosaida, and H. A. Hamid, "Future daily PM10 concentration prediction by combining regression models and feedforward backpropagation models with Principal Component Analysis (PCA)," *Atmospheric Environment*, vol. 77, pp. 621–637, 2013.

[8] M. Muhammad, A. Z. Ul-Saufie, and S. M. Deni, "Three days ahead prediction of daily 12 hour Ozone (O_3) concentrations for urban area in Malaysia," *Journal of Environmental Science and Technology*, vol. 8, no. 3, pp. 102–112, 2015.

[9] ASEAN Secretariat, *Roadmap on ASEAN Cooperation towards Transboundary Haze Pollution Control with Means of Implementation*. Jakarta, Indonesia: ASEAN Secretariat, Feb. 2021.

[10] R. Afroz, S. Hasan, and A. Ibrahim, "Review of air pollution and health impacts in Malaysia," *Environmental Research*, vol. 92, pp. 71–77, 2023.

[11] Department of Environment Malaysia, Annual Air Quality Report, 2023.

[12] S. N. M. Tarmizi *et al*., "Ten years seasonal and spatial Particulate Matter (PM10) Concentration at Selected Region in Klang Valley, Malaysia", *International Journal of Sciences: Basic and Applied Research* (IJSBAR), vol. 49, no. 2, pp. 154–165, Feb. 2020,

[13] Y. P. Rybarczyk, N. H. Dave, T. I. Tapia-Flores, and R. Zalakeviciute, "Inferring causal interplay between air pollution and meteorology," *Frontiers in Big Data*, vol. 8, 1710462, 2025.

[14] T. Istiana, R. Ghani, and N. Idris, "Causality analysis of air quality and meteorological parameters," *Aerosol and Air Quality Research*, vol. 23, 2023.

[15] W. Y. Ping, Z. Abd Rais, N. Ramli, N. Mohamed Noor, A.Z. Ul-Saufie, H.A. Hamid and M.K.N. Mahmad, "Granger causality analysis of air pollutants and meteorological parameters," *Environ. Earth Sci. Proc.*, vol. 33, 6, 2025.

[16] A. Z. Ul-Saufie, N. A. Norazian, and H. Zainal, "Improving air-pollution prediction using wrapper feature selection," *Sustainability*, vol. 14, 11403, 2022.

[17] S. Hameed, A. Latif, and F. Rashid, "Deep-learning multimodal urban air-quality prediction," *Journal of Environmental Management*, vol. 340, 2023.

[18] I. Jairi, S. Ben-Othman, L. Canivet, and H. Zgaya-Biau, "Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants," *Environmental Technology & Innovation*, vol. 36, 103793, 2024. 10.1016/j.eti.2024.103793

[19] S. I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, and M. C. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling & Software*, vol. 22, no. 1, pp. 97–103, 2007.

[20] A. Tella, A. L. Balogun, and I. Faye, "Spatio-temporal modelling of the influence of climatic variables and seasonal variation on PM_{10} in Malaysia using Multivariate Regression (MVR) and GIS," *Geomatics, Natural Hazards and Risk*, vol. 12, no. 1, pp. 443–468, 2021.

[21] N. M. Ariff, M. A. A. Bakar, and H. Y. Lim, "Prediction of PM_{10} Concentration in Malaysia Using K-Means Clustering and LSTM Hybrid Model," *Atmosphere*, vol. 14, no. 5, 853, 2023.

[22] A. Bashardoost, M. S. Mesgari, and M. Karimi, "Enhancing spatial modeling and risk mapping of six air pollutants using synthetic data integration with convolutional neural networks," *Frontiers in Environmental Science*, vol. 12, 1399339, 2024.

[23] İ. Aksangür, B. Eren, and C. Erden, "Evaluation of data preprocessing and feature selection process for prediction of hourly PM_{10} concentration using long short-term memory models," *Environmental Pollution*, vol. 311, 119973, 2022.

[24] S. R. Shams, S. Kalantary, A. Jahani, S. M. P. Shams, B. Kalantari, D. Singh, M. Moeinnadini, and Y. Choi, "Assessing the effectiveness of Artificial Neural Networks (ANN) and Multiple Linear Regressions (MLR) in forecasting AQI and PM10 and evaluating health impacts through AirQ+ (case study: Tehran)," *Environ. Pollut.,* vol. 338, 2023.

[25] S. Abdullah, N. N. L. M. Napi, A. N. Ahmed, W. N. W. Mansor, A. A. Mansor, M. Ismail, A. M. Abdullah, and Z. T. A. Ramly. "Development of multiple linear regression for Particulate Matter (PM10) Forecasting during episodic transboundary haze event in Malaysia," *Atmosphere*, vol. 11, 289, 2020.