AJVAH
e-ISSN: 2735-1165
DOI: https://doi.org/10.53797/ajvah.v7i1.1.2026

Assessing Meteorological Drivers of PM2.5 in a Tropical Coastal Industrial Zone: A Comparative Study of Linear and Interpretable Machine Learning Models in Perai, Penang

Hongzhi Lu[1], & Hongxue Lu[2] *

[1]School of Industrial Technology, Universiti Sains Malaysia,
Gelugor, 11800, Malaysia

[2]University of Malaya,
Kuala Lumpur, 50603, Malaysia

*Corresponding author email: elena.hongxue@gmail.com

Received 19 March 2026; Accepted 20 May 2026; Available online 27 May 2026

Abstract: Fine particulate matter (PM2.5) prediction in tropical coastal industrial zones is complicated by continuous industrial emissions, localized precipitation, sea-breeze circulation, and humidity-related measurement effects. This study developed a comparative and interpretable framework for daily PM2.5 estimation in the Perai Heavy Industrial Zone, Penang, using 97 concurrent observations from 2025-2026. Four meteorological predictors--temperature, wind speed, pressure, and precipitation--were evaluated with Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). MLR achieved the strongest baseline predictive performance, indicating that simple linear structure can remain competitive when sample size is limited and temporal leakage is controlled. XGBoost was retained for interpretation because it captured non-linear local interactions more effectively than the other ensemble alternative. SHapley Additive exPlanations (SHAP) identified temperature and precipitation as the dominant drivers. The positive precipitation-PM2.5 relationship suggests that hygroscopic aerosol growth and optical sensor response may partly offset the expected wet-scavenging effect in this setting. The findings show that localized, interpretable modelling can support air-quality warning, sensor calibration, and meteorology-sensitive emission management in tropical industrial regions.

Keywords: PM2.5 Prediction; XGBoost; Multiple Linear Regression; SHAP; Aerosol Hygroscopic Growth; Tropical Micro-climate.

1. Introduction

Air pollution, particularly fine particulate matter (PM2.5), remains a major public health and environmental concern because long-term exposure is associated with respiratory and cardiovascular risks and premature mortality (Cohen, et al., 2017; Di, et al., 2017; World Health Organization, 2021). Although air-quality modelling has advanced substantially, short-term PM2.5 prediction remains less settled in tropical coastal industrial zones, where emissions, humidity, precipitation, and sea-breeze circulation interact at local scales (Li, et al., 2026).

Perai, Penang, provides a relevant setting for this problem. The area combines electronics and semiconductor manufacturing with heavy industry, chemical processing, metal-related activities, and port logistics. This mixed industrial profile creates a persistent local emission background, while Malaysia's maritime tropical climate adds rainfall, humidity, and seasonal circulation effects that differ from those in many temperate air-quality studies (Latif, et al., 2014; Tangang, et al., 2012). Regional evidence also shows that particulate composition and carbonaceous aerosol variability in Southeast Asia can be influenced by industrial activity, biomass burning, and moisture-related processes (Amil, et al., 2016; Kalita, et al., 2020; Pani, et al., 2021).

Given the lack of localized and interpretable models for PM2.5 prediction in tropical industrial zones, this study aims to evaluate whether relatively simple statistical models and more flexible machine-learning models can reliably estimate short-term PM2.5 variability in Perai. This gap is important because recent Malaysian and Southeast Asian

*Corresponding author: author@organisation.edu.co

studies have demonstrated the value of machine learning for PM2.5 estimation, but fewer studies connect local industrial meteorology with explainable model interpretation at site scale (Aman, et al., 2025; Thaifa, et al., 2025; Zaman, et al., 2021). Broader studies of PM2.5-O3 response, satellite-based estimation, and regional trend reconstruction further indicate that pollutant behaviour can vary substantially by source mix and monitoring design (Jung, et al., 2018; N. Zhang et al., 2022; Zhao, et al., 2025).

Accordingly, this study compares Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) using recent local observations, and then applies SHapley Additive exPlanations (SHAP) to interpret the most informative non-linear model. The objective is not only to identify the most accurate model, but also to clarify how meteorological variables contribute to PM2.5 prediction in a tropical coastal industrial environment.

1. Literature Review

1.1 Meteorological Drivers of PM2.5 in Tropical and Coastal Settings

Studies in temperate and subtropical regions often link high PM2.5 episodes to synoptic-scale stagnation, low-pressure systems, regional emissions, or pandemic-related activity changes (Ma, et al., 2021; Ning, et al., 2018). In Southeast Asia, however, PM2.5 variability is also shaped by tropical rainfall, humidity, biomass-burning transport, and marine-influenced circulation (Amil, et al., 2016; Kalita, et al., 2020; Pani, et al., 2021). Long-term climate and aerosol datasets reinforce the point that local interpretation is necessary because regional circulation and monitoring coverage can shape observed pollution patterns (Tangang, et al., 2012; Zhong, et al., 2022).

The literature also indicates that moisture can complicate both physical PM2.5 behaviour and measurement. Brown carbon and secondary aerosol processes have been observed in Southeast Asian environments, while humidity can affect the apparent mass recorded by optical instruments (Amil, et al., 2016; Pani, et al., 2021). Satellite-based and ground-station validation studies likewise show that uneven monitoring distribution and spatial representativeness can influence PM2.5 estimates, making localized validation especially important for coastal industrial settings (Jung, et al., 2018; Li, et al., 2020; Zhong, et al., 2022).

1.2 Machine-Learning Models for Air-Quality Prediction

Machine-learning models have been widely applied to air-quality forecasting because they can represent non-linear interactions among emissions, meteorology, land use, and temporal variables (Castelli, et al., 2020; Lei, et al., 2022; Zaman, et al., 2021). Recent studies have extended this work through deep learning, hybrid architectures, and high-resolution urban prediction frameworks for PM2.5, PM10, and ozone (He, et al., 2025; K. Zhang, et al., 2023). Related building-scale and indoor-environment studies also show that PM2.5 dynamics may depend on setting-specific predictors that cannot be assumed from outdoor regional models alone.

Comparative work suggests that no single algorithm consistently dominates across air-quality settings. SVR has been used successfully in some urban forecasting studies, but its performance is sensitive to kernel choice, scaling, and hyperparameter tuning (Arampongsanuwat & Meesad, 2011; Liu, et al., 2017). Feature selection and hyperparameter optimization can improve model stability, especially in small or high-dimensional datasets (Akiba, et al., 2019; Cai, et al., 2018). Tree-based ensembles such as RF and XGBoost are often competitive because they capture non-linear interactions, but their variable-importance measures require careful interpretation when predictors are correlated or unevenly sampled (Breiman, 2001; Chen & Guestrin, 2016; Strobl, et al., 2007).

1.3 Explainable AI and SHAP in Environmental Modelling

SHAP has become a common tool for interpreting machine-learning outputs because it estimates the marginal contribution of each predictor to model predictions (Lundberg & Lee, 2017). Recent environmental studies have used SHAP with XGBoost and related models to examine air quality, ozone, landslide susceptibility, and spatial pollutant mapping, demonstrating its value for translating complex model behaviour into ranked and directional feature effects (He et al., 2025; Hu et al., 2022; Li et al., 2026; Song et al., 2023; K. Zhang et al., 2023).

At the same time, SHAP results should be interpreted as model-based evidence rather than direct causal proof. Feature importance can be affected by sampling, predictor correlation, spatial coverage, and the physical representativeness of the dataset (Li et al., 2020; Strobl et al., 2007). This study therefore uses SHAP as an explanatory complement to predictive comparison, not as a substitute for atmospheric process measurement or source-apportionment analysis.

2. Materials and Methods

2.1 Study Area and Data Acquisition

This study focuses on the Perai Heavy Industrial Zone, located in Penang, Malaysia (approximate coordinates: 5.38° N, 100.38° E). As one of the most critical industrial hubs in Southeast Asia, Perai's industrial portfolio includes significant heavy manufacturing, chemical processing, and metal smelting, ensuring a persistent, non-seasonal emission baseline of

primary pollutants. Climatologically, Perai features a typical tropical rainforest climate (Köppen Af) characterized by high humidity, uniform temperatures, and frequent precipitation (Tangang et al., 2012).

Geographically, Perai is situated on the mainland coastline of the Malacca Strait, directly facing Penang Island. This specific topology creates a narrow marine corridor (the Penang Strait) that profoundly influences local micrometeorology. This geographical configuration not only dictates the land-sea breeze diurnal cycle but also creates a "topographical funneling effect," which can selectively channel or trap atmospheric pollutants depending on prevailing monsoonal wind vectors. Raw meteorological data (average ambient temperature in °C, wind speed in km/h, atmospheric pressure in hPa, and precipitation in mm) and regional PM2.5 concentration data () were initially retrieved spanning from January 2025 to April 2026.

Air quality monitoring stations in tropical industrial zones can experience sensor downtime or transmission failures during extreme weather events, including monsoonal rainfall and lightning. To improve data integrity and reduce artificial imputation bias, an internal intersection was applied to the datasets, and days with missing PM2.5 sensor records were excluded. This filtering process produced 97 concurrent daily observations for which meteorological and air-quality measurements were available.

2.2 Data Preprocessing

To improve model robustness and convergence stability, missing values in precipitation were imputed as zero, reflecting the absence of rainfall on those specific days. Because tropical rainfall is episodic rather than continuous, this treatment is meteorologically defensible for daily records while avoiding unnecessary interpolation. After imputation, all input variables were standardized before model fitting, a step especially relevant for SVR and for feature-selection workflows in compact datasets (Cai et al., 2018).

To reduce the risk of data leakage and temporal overfitting in the 97-sample dataset, temporal identifiers such as month and day were excluded from the feature space. The models were therefore trained on four meteorological predictors only: temperature, wind speed, atmospheric pressure, and precipitation. Standardization was applied to support distance-based algorithms such as SVR and to improve numerical comparability across predictors.

The standardized dataset was then chronologically split into a training set (80%) and an independent testing set (20%). To strengthen validation despite the small sample size, the training phase incorporated a k-fold cross-validation strategy before the final holdout evaluation. This design follows the logic of recent optimization and computational-intelligence studies, where tuning choices are evaluated within the training process before final testing is performed on unseen data (Akiba et al., 2019).

2.3 Machine Learning Algorithms and SHAP Theoretical Framework

The modelling framework compared MLR and SVR with two tree-based ensemble methods, RF and XGBoost. MLR served as an interpretable linear baseline, while SVR tested a kernel-based non-linear approach that has been used in particulate-matter prediction (Arampongsanuwat & Meesad, 2011; Liu et al., 2017). RF and XGBoost were included to capture potential interactions and threshold effects among the meteorological predictors (Breiman, 2001; Chen & Guestrin, 2016).

SHAP was used to interpret the XGBoost model because tree-based ensembles can identify non-linear structure but are less transparent than linear models. In this study, SHAP provides a consistent way to compare the contribution of each meteorological predictor to daily PM2.5 estimates, while avoiding overstatement of conventional impurity-based importance measures (Lundberg & Lee, 2017; Strobl et al., 2007).

Operationally, SHAP estimates the marginal contribution of each meteorological feature to the predicted PM2.5 concentration across possible feature combinations. Local explanations were aggregated to identify global patterns, and directional SHAP values were then used to infer whether higher values of a meteorological feature tended to raise or lower predicted PM2.5 within the XGBoost model. This approach aligns with recent environmental applications that use explainability to support model diagnosis rather than to claim direct causality (He et al., 2025; Hu et al., 2022).

3. Results and Discussion

3.1 Predictive Performance: Simplicity vs. Complexity

Before interpreting individual model behaviour, Table 1 summarizes predictive performance on the independent testing dataset. Higher R-squared values and lower RMSE and MAE values indicate stronger predictive accuracy, allowing the models to be compared on both explained variance and absolute prediction error.

Table 1: Performance evaluation of machine learning models

Algorithm	R-squared (R2)	RMSE (µg/m3)	MAE (µg/m3)
Multiple Linear Regression (MLR)	0.5924	12.8401	10.1493

Support Vector Regression (SVR)	-0.0066	20.1775	15.0265
Random Forest Regressor (RF)	0.5226	13.8963	10.7533
XGBoost Regressor	0.5573	13.3817	9.4826

The results indicate that the traditional MLR model achieved the highest predictive accuracy in this dataset, with the strongest R-squared value and the lowest RMSE. This outcome suggests that, under a small-sample setting with carefully restricted predictors, the dominant daily meteorological signal may be sufficiently linear for short-term prediction. Similar caution is noted in comparative air-quality studies, where more complex models do not automatically outperform simpler baselines when data volume, predictor coverage, or local representativeness is limited (Castelli et al., 2020; Lee et al., 2020; Lei et al., 2022).

The SVR model underperformed relative to the other approaches, suggesting difficulty in mapping the feature space under the current sample size and distribution. This result is consistent with the known sensitivity of RBF-kernel SVR to scaling, penalty selection, and kernel parameters, even though SVR has performed well in other particulate-matter forecasting contexts (Arampongsanuwat & Meesad, 2011; Liu et al., 2017).

Among the non-linear models, XGBoost provided the strongest ensemble performance and outperformed RF on MAE. Although it did not surpass MLR in overall predictive accuracy, XGBoost was retained for interpretability analysis because its structure can reveal threshold-like and interaction patterns that may be relevant for local management (Chen & Guestrin, 2016; Song et al., 2023). This balanced use of XGBoost avoids treating the most complex model as automatically superior.

3.2 Global Feature Importance via SHAP

After the predictive comparison, SHAP was applied to the XGBoost model to examine which meteorological variables contributed most to its predictions. Figure 1 should be read as a global ranking of feature contribution, rather than as direct evidence of causality.

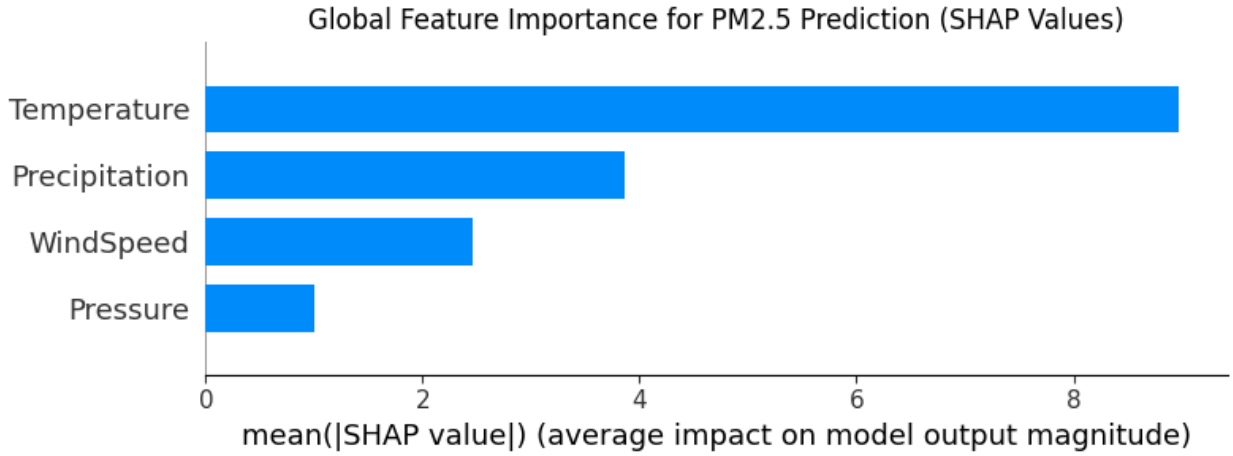

Figure 1: SHAP summary plot detailing global feature importance for $PM_{2.5}$ prediction derived from the XGBoost model

Ambient Temperature emerged as the most influential predictor in the SHAP ranking. This pattern is consistent with studies showing that temperature can influence secondary aerosol formation, photochemical activity, and non-linear PM2.5-O3 responses, particularly where industrial precursors and meteorology interact (Li et al., 2026; Song et al., 2023; Zhao et al., 2025). In Perai, higher surface temperatures may also strengthen local sea-breeze development, altering pollutant transport along the Penang Strait.

Atmospheric Pressure showed the lowest overall contribution to short-term PM2.5 variability in this model. One plausible explanation is that pressure gradients near the equator are often weaker than those in mid-latitude systems, reducing the short-term predictive contrast of pressure compared with temperature, rainfall, and wind speed. This differs from mid-latitude studies where pressure systems can strongly modulate pollution accumulation (Ning et al., 2018), reinforcing the need for locally calibrated interpretation in Malaysia's maritime climate (Tangang et al., 2012).

3.3 Directional Impacts of Meteorological Drivers

Global feature ranking does not show whether high or low values of a predictor increase modelled PM2.5. For this reason, Figure 2 is used to examine the direction and spread of SHAP values across the four meteorological variables.

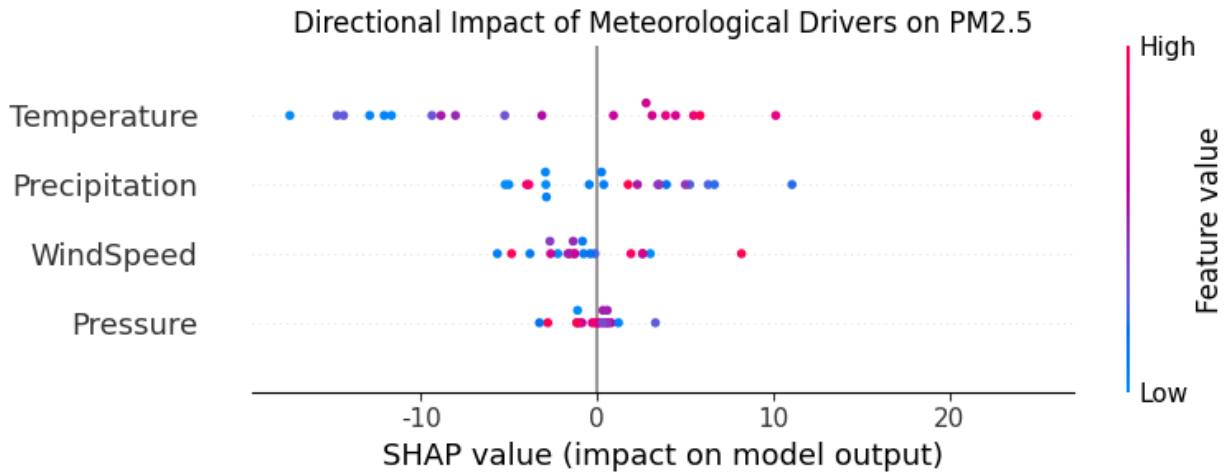

Figure 2: SHAP beeswarm plot illustrating the directional impact and distribution of meteorological factors on $PM_{2.5}$ predictions.

Precipitation showed a positive association with PM2.5 in the XGBoost-SHAP output. This result differs from the common wet-scavenging expectation, in which rainfall removes particles from the atmosphere. In the observed period, high precipitation values were more likely to coincide with high-humidity conditions, local emissions, and reduced dispersion than with simple washout, a pattern that is plausible in tropical aerosol environments (Aman et al., 2025; Amil et al., 2016; Kalita et al., 2020; Pani et al., 2021).

This positive association may be interpreted through the combined effects of hygroscopic aerosol growth and measurement response. Under humid tropical conditions, water-soluble fine particles can absorb moisture and increase in size, while optical PM2.5 monitors may report elevated apparent mass when relative humidity is high (Amil et al., 2016; Pani et al., 2021). The result should therefore be read as a local model signal that requires follow-up humidity and aerosol-composition measurements.

The coastal geography of Perai provides an additional explanation. Sea-breeze circulation along the Penang Strait can transport marine moisture inland during daytime heating. When this humid boundary layer interacts with continuous industrial emissions, it may temporarily enhance measured PM2.5 or slow pollutant removal. This interpretation is consistent with the broader need to separate emission-driven pollutant loading from meteorology-driven measurement and transport effects in coastal industrial areas (Latif et al., 2014; N. Zhang et al., 2022).

Wind Speed generally showed an inverse relationship with PM2.5 accumulation. Higher wind speeds tended to correspond to negative SHAP values, suggesting a dilution or dispersion effect when winds transported locally generated industrial emissions away from the monitoring location. Comparable machine-learning studies have also found that meteorological dispersion variables can become influential predictors when local emissions and monitoring position are explicitly represented (He et al., 2025; K. Zhang et al., 2023).

3.4 Policy Implications for Regional Air Quality Management

The findings offer practical implications for regional air-quality management. First, local warning systems could benefit from integrating short-term meteorological forecasts, especially temperature, precipitation, and wind conditions, rather than relying only on static emission indicators. Regional control studies suggest that PM2.5 and ozone mitigation need to account for non-linear pollutant responses, spatial heterogeneity, and temporal changes in activity patterns (Ma et al., 2021; N. Zhang et al., 2022; Zhao et al., 2025).

For example, periods combining high ambient temperature, high humidity, and weak dispersion could be used to trigger closer monitoring or temporary preventive measures in high-emission industrial activities. Such strategies would need to be validated against longer local datasets and source information, but they are consistent with explainable modelling studies that use SHAP to identify conditions under which pollution-prone areas or high-concentration episodes become more likely (Song et al., 2023).

Second, the results highlight the need to distinguish actual pollutant loading from humidity-related measurement effects. Heated inlets, calibration checks, and localized humidity-correction algorithms could improve the reliability of optical PM2.5 monitoring in Perai. This recommendation is consistent with station-distribution and validation studies showing that monitoring design can affect PM2.5 estimates and model transferability (Jung, et al., 2018; Li, et al., 2020; Zhong, et al., 2022).

4. Conclusion

This study evaluated a comparative and interpretable framework for forecasting daily PM2.5 concentrations in the Perai Heavy Industrial Zone, Penang. The results show that MLR provided the strongest predictive accuracy in a small,

carefully filtered dataset, while XGBoost offered useful explanatory value through SHAP. Temperature and precipitation were the most influential meteorological drivers in the XGBoost interpretation, whereas wind speed contributed mainly through dispersion and atmospheric pressure had limited short-term influence.

The main contribution is the integration of model comparison with interpretable feature analysis in a localized tropical industrial setting. The findings suggest that simple models may remain competitive when datasets are small and predictor sets are focused, while explainable machine learning can still clarify threshold-like meteorological effects relevant to local monitoring and warning systems. For Perai, the results point to the practical importance of combining emission control with meteorological awareness, especially during warm, humid, and weakly dispersed conditions.

Limitations and Future Work: The study remains constrained by the 97-day observational record and by the absence of direct relative humidity, aerosol-composition, and source-apportionment measurements. Future work should extend the monitoring period across monsoonal transitions, include humidity and industrial activity indicators, validate the model with additional monitoring stations, and test whether local humidity correction improves PM2.5 prediction. These steps would strengthen model transferability and support more operational air-quality management in tropical coastal industrial zones.

Data Availability Statement:
The original dataset, data cleaning scripts, and processed files used in this study are available in the project repository listed in the Appendix. A Zenodo DOI can be added once the archive record is finalized.

Acknowledgement

The authors express their gratitude to Universiti Sains Malaysia and Universiti Malaya for Support Given.

Appendix: Code Repository and Data

The Python source code, data cleaning scripts, and processed datasets used in this study have been uploaded to the zenodo open-source community and can be accessed via the following link:

LU, H. (2026). Assessing Meteorological Drivers of PM2.5 in a Tropical Coastal Industrial Zone: A Comparative Study of Linear and Interpretable Machine Learning Models in Perai, Penang [Data set]. Zenodo. https://doi.org/10.5281/zenodo.20129553

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623-2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701

Aman, N., Panyametheekul, S., Pawarmart, I., Sudhibrabha, S., & Manomaiphiboon, K. (2025). A visibility-based historical PM2.5 estimation for four decades (1981-2022) using machine learning in Thailand: Trends, meteorological normalization, and influencing factors using SHAP analysis. Aerosol and Air Quality Research, 25, Article 4. https://doi.org/10.1007/s44408-025-00007-z

Amil, N., Latif, M. T., Khan, M. F., & Mohamad, M. (2016). Seasonal variability of PM2.5 composition and sources in the Klang Valley urban-industrial environment. Atmospheric Chemistry and Physics, 16(8), 5357-5381. https://doi.org/10.5194/acp-16-5357-2016

Arampongsanuwat, S., & Meesad, P. (2011). Prediction of PM10 using support vector regression. In International Proceedings of Computer Science and Information Technology (Vol. 6, pp. 120-124). IACSIT Press. https://hero.epa.gov/reference/4244373

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020, Article 8049504. https://doi.org/10.1155/2020/8049504

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., ... Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air

pollution: An analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907-1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., & Schwartz, J. D. (2017). Air pollution and mortality in the Medicare population. New England Journal of Medicine, 376(26), 2513-2522. https://doi.org/10.1056/NEJMoa1702747

He, Z., Guo, Q., Zhang, Z., Feng, G., Qiao, S., & Wang, Z. (2025). Forecasting daily ambient PM2.5 concentrations in Qingdao City using deep learning and hybrid interpretable models and analysis of driving factors using SHAP. Toxics, 14(1), 44. https://doi.org/10.3390/toxics14010044

Hu, X., Zhang, J., Xue, W., Zhou, L., Che, Y., & Han, T. (2022). Estimation of the near-surface ozone concentration with full spatiotemporal coverage across the Beijing-Tianjin-Hebei region based on extreme gradient boosting combined with a WRF-Chem model. Atmosphere, 13(4), 632. https://doi.org/10.3390/atmos13040632

Jung, C.-R., Hwang, B.-F., & Chen, W.-T. (2018). Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2.5 concentrations in Taiwan from 2005 to 2015. Environmental Pollution, 237, 1000-1010. https://doi.org/10.1016/j.envpol.2017.11.016

Kalita, G., Kunchala, R. K., Fadnavis, S., & Kaskaoutis, D. G. (2020). Long term variability of carbonaceous aerosols over Southeast Asia via reanalysis: Association with changes in vegetation cover and biomass burning. Atmospheric Research, 245, 105064. https://doi.org/10.1016/j.atmosres.2020.105064

Lei, T. M. T., Siu, S. W. I., Monjardino, J., Mendes, L., & Ferreira, F. (2022). Using machine learning methods to forecast air quality: A case study in Macao. Atmosphere, 13(9), 1412. https://doi.org/10.3390/atmos13091412

Li, D., Liu, M., Han, H., & Wang, J. (2026). Nonlinear impacts of air pollutants and meteorological factors on PM2.5: An interpretable GT-iFormer model with SHAP analysis. Atmosphere, 17(3), 266. https://doi.org/10.3390/atmos17030266

Li, T., Shen, H., Zeng, C., & Yuan, Q. (2020). A validation approach considering the uneven distribution of ground stations for satellite-based PM2.5 estimation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1312-1321. https://doi.org/10.1109/JSTARS.2020.2977668

Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., & Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. PLOS ONE, 12(7), e0179763. https://doi.org/10.1371/journal.pone.0179763

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://papers.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., He, M. Z., Li, S., Shi, W., & Li, T. (2021). Random forest model based fine scale spatiotemporal O3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. Environmental Pollution, 276, 116635. https://doi.org/10.1016/j.envpol.2021.116635

Ning, G., Wang, S., Yim, S. H. L., Li, J., Hu, Y., Shang, Z., Wang, J., & Wang, J. (2018). Impact of low-pressure systems on winter heavy air pollution in the northwest Sichuan Basin, China. Atmospheric Chemistry and Physics, 18(18), 13601-13615. https://doi.org/10.5194/acp-18-13601-2018

Pani, S. K., Lin, N.-H., Griffith, S. M., Chantara, S., Lee, C.-T., Thepnuan, D., & Tsai, Y. I. (2021). Brown carbon light absorption over an urban environment in northern peninsular Southeast Asia. Environmental Pollution, 276, 116735. https://doi.org/10.1016/j.envpol.2021.116735

Song, Y., Zhang, C., Jin, X., Zhao, X., Huang, W., Sun, X., Yang, Z., & Wang, S. (2023). Spatial prediction of PM2.5 concentration using hyper-parameter optimization XGBoost model in China. Environmental Technology & Innovation, 32, 103272. https://doi.org/10.1016/j.eti.2023.103272

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8, 25. https://doi.org/10.1186/1471-2105-8-25

Tangang, F. T., Juneng, L., Salimun, E., Sei, K. M., Le, L. J., & Muhamad, H. (2012). Climate change and variability over Malaysia: Gaps in science and research information. Sains Malaysiana, 41(11), 1355-1366. https://www.ukm.my/jsm/english_journals/vol41num11_2012/vol41num11_2012pg1355-1366.html

Thaifa, H., Muhammad, M., ul-Saufie, A. Z., Abd Hadi, N. A., Sulong, N. A., & Prasasti, C. I. (2025). Enhancing short-term PM2.5 prediction in Shah Alam using wrapper feature selection and machine learning techniques. Israa University Journal for Applied Science, 8(2), 1-32. https://doi.org/10.52865/sfwd4660

World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://www.who.int/publications/i/item/9789240034228

Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating PM2.5 concentrations across Malaysia. Applied Sciences, 11(16), 7326. https://doi.org/10.3390/app11167326

Zhang, K., Yang, X., Cao, H., Thé, J., Tan, Z., & Yu, H. (2023). Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial-temporal attention and residual learning. Environment International, 171, 107691. https://doi.org/10.1016/j.envint.2022.107691

Zhang, N., Guan, Y., Jiang, Y., Zhang, X., Ding, D., & Wang, S. (2022). Regional demarcation of synergistic control for PM2.5 and ozone pollution in China based on long-term and massive data mining. Science of the Total Environment, 838, 155975. https://doi.org/10.1016/j.scitotenv.2022.155975

Zhao, N., Zhang, H., & Wang, G. (2025). Revealing the nonlinear responses of PM2.5 and O3 to VOC and NOx emissions from various sources in Shandong, China. Journal of Hazardous Materials, 489, 137655. https://doi.org/10.1016/j.jhazmat.2025.137655