The 3rd ASEAN School of Business Network International Conference 2026
www.aseansbn.org August 6-7, 2026 Pathum Thani, Thailand

Machine Learning-Based PM2.5 Prediction Using Meteorological Data for Environmental Decision Support

Kanok-on Hoondee [1], Jıraphat Lapwattanaworakul [2*]
[1]King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[2]Rangsit University, Pathum Thani, Thailand

*Corresponding Author: jiraphat.l@rsu.ac.th

ABSTRACT

Fine particulate matter (PM2.5) remains a critical environmental and public health concern. Accurate prediction of PM2.5 concentrations is essential for air quality monitoring, pollution mitigation, and evidence-based decision-making. Although machine learning has demonstrated promising performance in air quality prediction, many existing studies emphasize predictive accuracy while providing limited comparative evaluation and insufficient interpretation of the variables contributing to model predictions.

This study develops a comparative machine learning framework for predicting daily PM2.5 concentrations using meteorological variables and historical PM2.5 observations. Four models—Linear Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Stacking Ensemble—were evaluated using MAE, RMSE, MAPE, and R^2. Explainable Artificial Intelligence was incorporated through SHapley Additive exPlanations (SHAP) to interpret the contributions of meteorological, temporal, and historical PM2.5 features.

The results show that XGBoost achieved the best overall predictive performance, with the lowest RMSE of 6.9639 and the highest R^2 of 0.7917, while the Stacking Ensemble obtained the lowest MAE and MAPE. SHAP analysis indicated that historical PM2.5 features were the strongest contributors to prediction, while relative humidity was the most influential meteorological variable, followed by temperature and wind speed. These findings demonstrate the potential of combining comparative machine learning and explainable AI to support interpretable PM2.5 forecasting and environmental decision-making.

Keywords: Machine Learning, PM2.5 Prediction, Meteorological Data, Explainable Artificial Intelligence, Environmental Decision Support

DOI: https://doi.org/10.64458/asbnic.v3.144

INTRODUCTION

Meteorological Variables Affecting PM2.5

Meteorological conditions significantly influence the formation, accumulation, transport, and dispersion of fine particulate matter (PM2.5). Previous studies have identified temperature, relative humidity, wind speed, rainfall, atmospheric pressure, and wind direction as relevant predictors of PM2.5 concentrations

across different atmospheric and geographical settings (Sirithian & Thanatrakolsri, 2022; Aman et al., 2025; Peng et al., 2024).

Relative humidity is frequently associated with PM2.5 variability because moisture conditions can affect aerosol growth and particle accumulation. Temperature may influence atmospheric stability and chemical transformation processes, while wind speed and wind direction affect pollutant transport and dispersion. Rainfall can reduce airborne particulate concentrations through wet deposition, whereas atmospheric pressure may reflect broader weather conditions associated with pollutant accumulation. However, the magnitude and direction of these relationships can vary according to season, geographical characteristics, and local emission (Sirithian & Thanatrakolsri, 2022; Aman et al., 2025; Wood, 2022)

Accordingly, meteorological variables are increasingly integrated with historical PM2.5 observations in machine learning models to capture both environmental influences and temporal dependence. Previous studies indicate that combining these data sources can improve forecasting performance by representing short-term pollutant persistence, seasonal variation, and changing weather conditions (Bhatta et al., 2026; Peng et al., 2024; Thapa et al., 2026). This evidence provides the basis for incorporating meteorological variables together with lagged and rolling PM2.5 features in the predictive framework developed in this study.

Machine Learning Approaches for PM2.5 Prediction

Machine learning has been widely applied to PM2.5 prediction because of its ability to model complex and potentially nonlinear relationships between pollution concentrations and environmental variables. Commonly used approaches include Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Linear Regression (LR). Depending on the characteristics of the dataset and prediction task, these models have demonstrated promising performance in air quality forecasting and often provide greater flexibility than conventional statistical methods (Breiman, 2001; Chen & Guestrin, 2016; Peng et al., 2024)

Several recent studies have demonstrated the effectiveness of machine learning for PM2.5 prediction. Bhatta et al., (2026) compared multiple machine learning models for forecasting PM2.5 concentrations in Bangkok and demonstrated the potential of data-driven approaches for air quality prediction and trend analysis. Rosca et al., (2025) reviewed and evaluated various data-driven methods for urban air quality prediction, highlighting the strong performance of tree-based ensemble models. In Thailand, Thongrod et al., (2022) compared several machine learning algorithms for predicting PM2.5 and PM10 concentrations in Chiang Mai and confirmed the applicability of machine learning under local environmental conditions.

Stacking Ensemble has also emerged as an effective strategy for combining the complementary strengths of multiple predictive models. In a stacking architecture, predictions generated by base learners are integrated through a meta-learner to improve robustness and predictive performance. Previous studies indicate that stacking-based models can outperform individual algorithms when the component models capture different aspects of the underlying data patterns (Ravindiran et al., 2025). This evidence motivates the inclusion of a Stacking Ensemble model in the present study alongside Linear Regression, Random Forest, and XGBoost.

Explainable Artificial Intelligence for PM2.5 Prediction

Although complex machine learning models can achieve strong predictive performance, their limited transparency may restrict their practical use in environmental monitoring and decision-making. Explainable Artificial Intelligence (XAI) techniques have therefore become increasingly important for identifying how predictor variables contribute to model outputs.

Among the available XAI methods, SHapley Additive exPlanations (SHAP) is widely used to interpret machine learning predictions by estimating the contribution of each feature to individual and overall model outputs (Lundberg & Lee, 2017) Recent studies have applied SHAP to PM2.5 prediction to

examine the influence of meteorological variables, historical pollutant observations, and emission-related factors (Aman et al., 2025; Rosca et al., 2025).

However, SHAP values should be interpreted as measures of contribution to model predictions rather than evidence of causal relationships. Nevertheless, the integration of SHAP with machine learning provides a practical approach for improving model transparency and supporting evidence-based environmental interpretation. Accordingly, the present study applies SHAP to the best-performing prediction model to identify the most influential historical, temporal, and meteorological features.

Research Gap

Despite substantial advances in machine learning-based PM2.5 prediction, several important research gaps remain. Although previous studies have reported promising predictive performance, comparatively fewer have systematically evaluated multiple machine learning models while simultaneously investigating the contribution of meteorological variables to model interpretation. Furthermore, the integration of Explainable Artificial Intelligence (XAI) techniques into comparative machine learning studies remains relatively limited, reducing model transparency and practical interpretability. In addition, many existing studies primarily emphasize predictive performance while providing limited discussion of how prediction results can support environmental monitoring and evidence-based decision-making. These limitations highlight the need for a comparative and interpretable machine learning framework that integrates predictive performance evaluation with meteorological factor analysis to support practical environmental decision-making.

Our Study

Based on the identified research gaps, this study develops a comparative machine learning framework for PM2.5 prediction using meteorological data. The proposed framework systematically evaluates multiple machine learning models and investigates the relative contribution of historical PM2.5, temporal, and meteorological variables to improve model interpretability. Unlike previous studies that primarily emphasize predictive accuracy or focus on individual modeling techniques, this study integrates comparative model evaluation with explainable model interpretation to establish an environmental decision-support framework for evidence-based environmental monitoring and pollution management.

METHOD

This study developed a machine learning framework to predict daily PM2.5 concentrations using meteorological variables and historical PM2.5 observations. The proposed framework consisted of six main stages: data collection, data preprocessing, feature engineering, machine learning model development, model evaluation, and model interpretation using Explainable Artificial Intelligence (XAI).

Study Area and Data Collection

Daily PM2.5 concentration data from 2023 to 2025 were collected from official governmental air-quality monitoring systems, including Air4Thai, AirBKK, and the Chiang Mai Air Quality Monitoring Center. Meteorological variables, including temperature, relative humidity, wind speed, and rainfall, were collected from the corresponding observation records. The initial raw air quality dataset contained 115,080 records and four variables before data selection and preprocessing. Subsequently, the air quality and meteorological datasets were integrated using the observation date to construct a unified dataset for machine learning model development.

The selected meteorological variables were chosen based on previous studies demonstrating their associations with PM2.5 formation, transport, and dispersion. Historical PM2.5 observations were

additionally incorporated to capture temporal persistence, while temporal variables were included to represent seasonal and short-term temporal patterns. This combination provides complementary information for improving prediction performance while supporting subsequent model interpretation.

Data Preprocessing

The collected datasets were cleaned and prepared prior to model development. Date formats from different data sources were standardized to ensure consistency across datasets. Duplicate records were removed, and the air quality and meteorological datasets were merged into a unified analytical dataset based on the observation date. Records without PM2.5 measurements were excluded because PM2.5 served as the prediction target and could not be reliably imputed. In contrast, missing values in the meteorological variables were estimated using linear interpolation, followed by forward-fill and backward-fill procedures to address leading or trailing missing observations that could not be interpolated directly. These preprocessing strategies preserved temporal continuity while minimizing unnecessary data loss. The resulting dataset provided a consistent and reliable foundation for subsequent feature engineering and machine learning model development.

Feature Engineering

To improve the ability of the models to capture temporal dependencies and recurring pollution patterns, additional features were engineered from historical PM2.5 observations and calendar information. Lag features (PM25_Lag1, PM25_Lag3, and PM25_Lag7) were generated to incorporate PM2.5 concentrations from one, three, and seven previous days, respectively, thereby capturing short-term temporal dependence. Rolling mean features with 3-day, 7-day, and 30-day windows were generated to represent short-term fluctuations and long-term pollution trends while reducing the influence of day-to-day variability. Furthermore, calendar-based variables, including Month, Day, Weekday, Quarter, and Season, were extracted from the observation date to represent recurring seasonal, monthly, and weekly patterns that may influence PM2.5 concentrations. Observations with undefined values introduced during lag and rolling window calculations were removed to ensure complete feature vectors for model training. The final analytical dataset consisted of 1,992 observations and was subsequently used for model development and evaluation.

Machine Learning Model Development

Four machine learning algorithms were implemented and comparatively evaluated for PM2.5 prediction:

- Linear Regression (LR)
- Random Forest (RF)
- Extreme Gradient Boosting (XGBoost)
- Stacking Ensemble

Linear Regression was employed as the baseline model because of its simplicity and interpretability. Random Forest, introduced by Breiman (2001), was selected for its capability to model nonlinear relationships through an ensemble of decision trees. Extreme Gradient Boosting (XGBoost), proposed by Chen and Guestrin (2016), was adopted because of its high predictive performance, regularization mechanism, and boosting strategy. A Stacking Ensemble model was developed by combining Random Forest and XGBoost as base learners with a Linear Regression meta-learner to improve prediction accuracy.

The final analytical dataset (n = 1,992) was divided into training and testing sets using an 80:20 chronological split, preserving the temporal order of the observations. Accordingly, 1,593 observations (80%) were used for model training, while 399 observations (20%) were reserved for model testing and

performance evaluation. This chronological data partition was adopted to prevent data leakage and to better simulate real-world PM2.5 forecasting scenarios.

Model performance was evaluated using four commonly adopted regression metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2).

Explainable Artificial Intelligence

To improve model interpretability, SHapley Additive exPlanations (SHAP), proposed by (Lundberg & Lee, 2017), were applied to the best-performing prediction model (XGBoost). SHAP analysis was performed to quantify the contribution of each predictor variable to the model predictions. The analysis included both engineered features (e.g., lag variables and rolling mean features) and meteorological variables, including temperature, relative humidity, wind speed, and rainfall, to investigate their contributions to PM2.5 prediction.

To provide comprehensive model interpretation, SHAP summary (beeswarm) plots, SHAP bar plots, and SHAP dependence plots were generated to visualize the overall importance of predictor variables and the effects of individual meteorological variables on PM2.5 predictions. In addition, XGBoost feature importance and permutation importance analyses were conducted to complement the SHAP interpretation and to assess the consistency of feature importance rankings across different interpretability methods.

RESULTS

This section presents the experimental results of the four machine learning models developed for PM2.5 prediction: Linear Regression (LR), Random Forest (RF), XGBoost (XGB), and Stacking Ensemble (SE). The models were evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). Furthermore, SHAP analysis was employed to identify the most influential variables affecting PM2.5 prediction.

Model Performance Comparison

Table 1. Performance Comparison of Machine Learning Models

Model	MAE	RMSE	MAPE (%)	R^2
Linear Regression	4.6960	8.5149	18.27	0.6885
Random Forest	4.5260	7.2488	16.90	0.7743
XGBoost	4.5165	6.9639	16.49	0.7917
Stacking Ensemble	4.4696	7.0832	16.38	0.7845

As shown in Table 1, all ensemble-based machine learning models outperformed the baseline Linear Regression model. Random Forest, XGBoost, and Stacking Ensemble achieved lower prediction errors and higher R^2 values than Linear Regression. Among the evaluated models, XGBoost achieved the lowest RMSE (6.9639) and the highest R^2 (0.7917), indicating the best overall predictive performance. Although Stacking Ensemble obtained the lowest MAE and MAPE values, its RMSE and R^2 were slightly lower than those of XGBoost. Therefore, XGBoost was selected for further interpretation using SHAP analysis.

SHAP Analysis

To improve model interpretability, SHAP analysis was applied to the XGBoost model. Figure 1 presents the SHAP summary plot illustrating the contribution of each feature to PM2.5 prediction. The results indicate that temporal features derived from historical PM2.5 observations, including Rolling Mean (3 days), PM25_Lag1, Rolling Mean (7 days), and Rolling Mean (30 days), contributed most significantly to

the model predictions. Among the meteorological variables, relative humidity exhibited the highest contribution, followed by temperature and wind speed, whereas rainfall showed the lowest contribution. These findings indicate that historical PM2.5 observations are the primary predictors of future PM2.5 concentrations, while meteorological variables further enhance the predictive performance of the model.

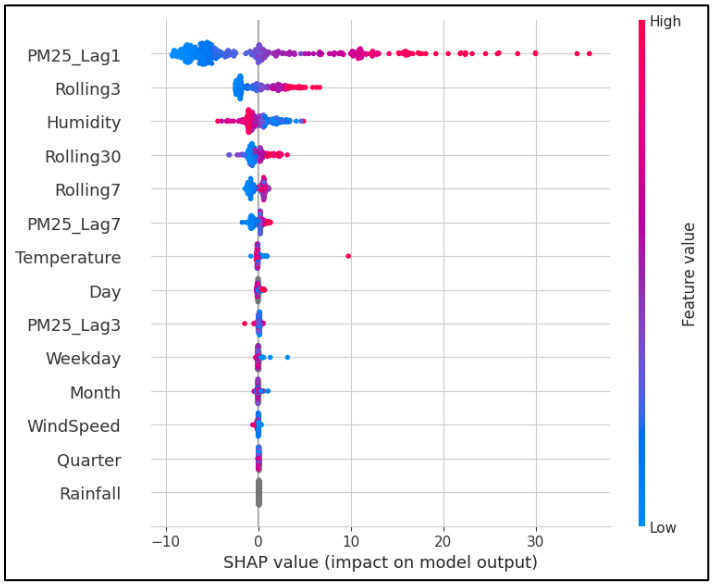


Figure 1. SHAP Summary Plot

Discussion

The findings demonstrate that machine learning techniques can effectively predict daily PM2.5 concentrations using meteorological variables and historical PM2.5 observations. Among the evaluated models, XGBoost achieved the best overall predictive performance, with the lowest RMSE and the highest R^2 value. This suggests that XGBoost is more effective at capturing the complex nonlinear relationships among historical PM2.5 observations, meteorological variables, and temporal features than the other evaluated models. These findings support the first and second research objectives of developing and comparing machine learning models for PM2.5 prediction.

The SHAP analysis provided additional insights into the prediction process by identifying the most influential variables. Historical PM2.5 features, particularly PM25_Lag1 and Rolling Mean (3 days), contributed most significantly to model prediction, indicating that recent PM2.5 conditions are strong predictors of subsequent PM2.5 concentration. Among the meteorological variables, Relative Humidity was identified as the most influential factor, followed by Temperature and Wind Speed, whereas Rainfall showed the lowest contribution. The relatively low contribution of rainfall may be associated with the limited number of rainfall events recorded during the study period, although this relationship may vary across different climatic conditions.

These findings are generally consistent with previous studies reporting the effectiveness of ensemble machine learning models for PM2.5 prediction (Bhatta et al., 2026; Ravindiran et al., 2025) Furthermore, the present study extends existing research by integrating comparative model evaluation with Explainable Artificial Intelligence (SHAP), enabling both accurate prediction and transparent interpretation of feature contributions. This combination provides additional value for environmental monitoring and evidence-based decision-making.

CONCLUSION

This study successfully developed and evaluated four machine learning models for predicting daily PM2.5 concentrations using meteorological variables and historical PM2.5 observations. The

experimental results demonstrated that all evaluated models were capable of predicting PM2.5 concentrations, with ensemble-based models outperforming the baseline Linear Regression model. Among the evaluated algorithms, XGBoost achieved the best overall predictive performance, obtaining the lowest RMSE and the highest coefficient of determination (R^2), indicating its effectiveness in modeling the complex relationships among historical PM2.5 observations, meteorological variables, and temporal features.

Furthermore, SHAP analysis improved model interpretability by identifying the variables that contributed most significantly to model prediction. Historical PM2.5 features, particularly PM25_Lag1 and Rolling Mean (3 days), were identified as the most influential features contributing to model predictors, while Relative Humidity was the most influential meteorological variable, followed by Temperature and Wind Speed. These findings demonstrate the value of combining historical PM2.5 observations with meteorological variables for accurate and interpretable PM2.5 prediction.

Overall, the proposed framework successfully achieved all research objectives by developing effective machine learning models, comparing the predictive performance of multiple algorithms, and identifying the key variables influencing PM2.5 prediction through Explainable Artificial Intelligence. The findings highlight the potential of integrating machine learning with Explainable Artificial Intelligence to support air quality monitoring, pollution warning systems, and evidence-based environmental decision-making. The proposed framework also provides a practical foundation for developing explainable machine learning applications that support data-driven environmental management and future smart-city air quality monitoring systems.

Recommendations

Future research should utilize more complete and higher-quality datasets with fewer missing values to improve model robustness and prediction accuracy. In addition, incorporating longer observation periods, additional monitoring stations, and more environmental variables, such as atmospheric pressure, wind direction, and satellite-derived data, may further enhance model performance. Future studies may also compare advanced deep learning models and additional Explainable Artificial Intelligence (XAI) techniques to improve both prediction accuracy and model interpretability. Furthermore, the proposed framework could be integrated into real-time air quality monitoring and early warning systems to support environmental management and public health decision-making.

ACKNOWLEDGEMENTS OR NOTES

The authors would like to express their sincere gratitude to the Centre of Excellence in AI and Supercomputing, Rangsit University, for providing academic support and research facilities throughout this study. The authors would also like to express their deepest appreciation to Jiraphat Lapwattanaworakul, Ph.D., for his valuable guidance, constructive comments, and continuous support throughout this research. The authors also acknowledge Air4Thai, AirBKK, and DPM Alert, managed by the Department of Disaster Prevention and Mitigation (DDPM), for providing publicly accessible air quality data used in this research. Finally, the authors would like to thank all individuals who contributed valuable suggestions and support during the completion of this study.

REFERENCES

Aman, N., Panyametheekul, S., Pawarmart, I., Xian, D., Gao, L., Tian, L., Manomaiphiboon, K., & Wang, Y. (2025). Machine learning-based quantification and separation of emissions and meteorological effects on PM2.5 in Greater Bangkok. Scientific Reports, 15(1), 14775. https://doi.org/10.1038/s41598-025-99094-6

Bhatta, J., Acharya, S. R., & Yang, K. M. (2026). Machine learning-enhanced air quality forecasting and trend analysis: A five-year comprehensive assessment of PM2.5 concentrations in Bangkok, Thailand. Environmental Challenges, 22, 101442. https://doi.org/10.1016/j.envc.2026.101442

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

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Peng, Z., Zhang, B., Wang, D., Niu, X., Sun, J., Xu, H., Cao, J., & Shen, Z. (2024). Application of machine learning in atmospheric pollution research: A state-of-art review. Science of The Total Environment, 910, 168588. https://doi.org/10.1016/j.scitotenv.2023.168588

Ravindiran, G., Karthick, K., Rajamanickam, S., Datta, D., Das, B., Shyamala, G., Hayder, G., & Maria, A. (2025). Ensemble stacking of machine learning models for air quality prediction for Hyderabad city in India. iScience, 28(2), 111894. https://doi.org/10.1016/j.isci.2025.111894

Rosca, C.-M., Carbureanu, M., & Stancu, A. (2025). Data-Driven Approaches for Predicting and Forecasting Air Quality in Urban Areas. Applied Sciences, 15(8), 4390. https://doi.org/10.3390/app15084390

Sirithian, D., & Thanatrakolsri, P. (2022). Relationships between Meteorological and Particulate Matter Concentrations (PM2.5 and PM10) during the Haze Period in Urban and Rural Areas, Northern Thailand. Air, Soil and Water Research, 15, 11786221221117264. https://doi.org/10.1177/11786221221117264

Thapa, I., Devkota, B., Lamichhane, B. R., Devkota, B. P., Dhakal, R., & Horanont, T. (2026). Applying Deep Learning and Machine Learning Algorithms to Estimate PM$_{2.5}$ Concentration Using Satellite Data and Meteorological Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 19, 2387–2404. https://doi.org/10.1109/JSTARS.2025.3620408

Thongrod, T., Lim, A., Ingviya, T., & Owusu, B. A. (2022). Prediction of PM2.5 and PM10 in Chiang Mai Province: A Comparison of Machine Learning Models. 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), 1–4. https://doi.org/10.1109/ITC-CSCC55581.2022.9894884

Wood, D. A. (2022). Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining. Sustainability Analytics and Modeling, 2, 100002. https://doi.org/10.1016/j.samod.2021.100002