The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W6-2025
3rd International Workshop on Evaluation and BENCHmarking of Sensors, Systems and
GEOspatial Data in Photogrammetry and Remote Sensing (GEOBENCH), 20–21 November 2025, Wroclaw, Poland

Evaluating Machine Learning Methods for $PM_{2.5}$ Estimation of GEMS Satellite AOD Data

Nadine Grace Caido[1], James Roy Lesidan[2], Maria Cecilia D. Galvez[1], Edgar A. Vallar[1]

[1] Department of Physics, College of Science, De La Salle University, Manila, Philippines
(nadine_grace_caido_a, maria.cecilia.galvez, edgar.vallar)@dlsu.edu.ph
[2] Department of Physics, College of Arts and Sciences, Visayas State University, Leyte, Philippines
jamesroy.lesidan@vsu.edu.ph

Keywords: air quality, GEMS satellite, $PM_{2.5}$, aerosol optical depth, machine learning.

Abstract

Rapid urbanization and industrialization affected the air quality in the Philippines. Fine particulate matter ($PM_{2.5}$) are of particular concern due to their health, environmental, as well as climate effects. Due to the lack of active and available air quality monitoring in the Philippines, air quality monitoring and mitigation cannot be performed. Satellite air quality data can be utilized to provide extensive spatial and temporal coverage. In this study, aerosol optical depth (AOD) data from the Geostationary Environment Monitoring Spectrometer (GEMS) onboard the GEO-COMPSAT-2B satellite was used to estimate $PM_{2.5}$ and compared with data from a ground monitoring station in Manila, Philippines along with meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5). Random forest (RD), support vector machine (SVM), and eXtreme Gradient Boosting (XGBoost) were evaluated for their accuracy in predicting ground-level $PM_{2.5}$. SVM achieved the highest accuracy (R^2: 0.998) followed by RF (R^2: 0.997), and then XGBoost (R^2: 0.673). SHAP analysis showed that wind speed has the highest contribution in predicting $PM_{2.5}$ This study shows that satellite air quality data can be used for ground-level $PM_{2.5}$ estimation.

1. Introduction

Most of the world's population are living in highly urbanized areas with unhealthy concentrations of fine particulate matter ($PM_{2.5}$). Asia's growing industry contributed anthropogenic emissions that are greater than those of Europe and America (Kumar et al., 2018). A study analyzed the air quality of the 50 most polluted cities in the world, which includes Manila, Philippines, and noted the lack of $PM_{2.5}$ monitoring in the Philippines (Rodríguez-Urrego and Rodríguez-Urrego, 2020). $PM_{2.5}$ greatly affects the health of humans, animals, and the environment. $PM_{2.5}$ pollutants have a large surface area that adsorbs toxic substances that if inhaled directly into the lungs and the bloodstream will affect other organs. A review by Tantengco and Guinto (2022) stated that air pollution negatively impacts the health of the Filipinos and the country's economy. $PM_{2.5}$ exceeds the WHO guidelines, especially during the dry season which can reach up to $58.4\ \mu g/m^3$ in traffic sites in Metro Manila. These effects emphasize the need for a comprehensive air quality monitoring system in highly urbanized areas that are accessible and retrievable by the local government.

With limited continuous air quality monitoring stations, a satellite remote sensing technology such as the Geostationary Environment Monitoring Spectrometer (GEMS) can offer free air quality data with an extensive spatial and temporal coverage. GEMS was launched in 2020 from the GEO-COMPSAT-2B satellite and developed by South Korea. It is collocated at 128.2° E over the equator and captures air quality data of South Korea and nearby Asian countries, including the Philippines. Hourly data are available at 3.5 km $\times$ 7 km resolution. To effectively improve its capability, implementing correlation between the GEMS aerosol optical depth (AOD) and ground monitoring $PM_{2.5}$ data should be studied and analyzed properly. In Manila, where proper ground monitoring stations are limited, validation of GEMS data will strengthen the capability and expand the framework for a comprehensive, reliable and accurate air quality monitoring system. The main objective of this study is to estimate $PM_{2.5}$ concentrations from GEMS aerosol optical depth (AOD) using machine learning models that are validated against a ground monitoring station. The specific objectives are the following:

1. Evaluate the trend of $PM_{2.5}$ in Manila, Philippines.
2. Determine the contribution of different meteorological features to the prediction of $PM_{2.5}$ by the machine learning models.
3. Compare which machine learning algorithm (Random Forest (RF), Support Vector Machine (SVM), and XGBoost) is most suitable for predicting $PM_{2.5}$ from GEMS satellite data.

2. Materials and Methods

2.1 Study Area

De La Salle University (DLSU) is located in Malate, Manila along Taft Avenue in the Philippines. It is part of the University Belt (U-Belt), a sub-district in Manila where a large concentration of universities is located, and heavy vehicular activities are observed. Therefore, high values of air pollutants are measured (Kecorius et al., 2017). The ground air quality monitoring station was placed in front of the Henry Sy (14.56° E, 120.99° N) building facing Taft Avenue (Figure 1).

2.2 Data Sources

To study the changes in air quality, the data were collected from two different sources (a) ground-based monitoring data, and (b) GEMS satellite air quality data. Daily average concentrations of $PM_{2.5}$ are obtained from the Air Quality Monitoring Station (AQMS) at Henry Sy inside De La Salle University campus. Similarly, daily average concentrations of

Figure 1. Study site. The red dot represents the ground monitoring site in De La Salle University.

GEMS AOD data were also utilized. Pixel values extracted from the location of AQMS $PM_{2.5}$ are collocated with that of GEMS AOD. Meteorological parameters including temperature, surface pressure, total column rainwater, relative humidity, and wind speed are retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) with $0.25° \times 0.25°$ horizontal resolution at hourly temporal resolutions from 2020 to 2023.

2.3 Feature Selection

The following features were used based on their relevance in predicting $PM_{2.5}$:

- aod: Average of all AOD values at 0.47 µm.
- $temp$: Ground temperature.
- ws: Wind speed.
- RH: Relative humidity.
- $tcrw$: Total column rainwater.
- sp: Surface pressure.
- aqi: Air Quality Index.

2.4 Machine Learning Models

Three machine learning models were compared based on their predictive capabilities of $PM_{2.5}$: Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost). The RF model uses an ensemble of decision trees, where each tree is generated by bootstrap sampling from the training dataset (Amiri and Zare Shahne, 2025). The general framework for $PM_{2.5}$ estimation in RF with N samples is:

$$f(x) = \sum_{z=1}^{z} C_z I(x \in R_z) \tag{1}$$

$$\hat{c}_z = mean(y_i \mid x \in R_z \tag{2}$$

where $f(x)$ = regression tree function
x_i, y_i = sample from region Z
R_Z = region

SVM uses supervised machine learning and finds an N-dimensional hyperplane with the maximum margin to classify data (Amiri and Zare Shahne, 2025). The linear kernel function is shown as:

$$Linear\ Kernel = k(X_i, X_j) = x_j^T x_j \tag{3}$$

where X_i, X_j = independent random vectors

The XGBoost model is a decision tree-based Gradient Boosting framework that combines weak models to create a stronger model with a built-in parallel processing for faster training of models with large datasets (Amiri and Zare Shahne, 2025). The objective function equation of XGBoost is given as:

$$Ob^{(t)} = \sum_{j=1}^{T} \left[G_j W_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T \tag{4}$$

where $Ob^{(t)}$ = objective function at iteration t
T = total leaf nodes in tree
G_j, H_j = cumulative sum of first- and second-order partial derivatives of the samples in leaf node j
λ, γ = constants
W_j = score value of j-th leaf node
w_j = weight value of j-th leaf node

The XGBoost model was coupled with the SHapley Additive exPlanation (SHAP) method to attribute $PM_{2.5}$ prediction with each feature and understand their importance as predictor variables. SHAP is an exploratory model framework based on game theory and local interpretability (Lundberg and Lee, 2017). Shapley values were calculated for each feature variable where their impact in $PM_{2.5}$ prediction was quantified. The images were split into a training set, a testing set, and a validation set with a ratio of 80:10:10.

2.5 Model Validation

R^2, RMSE, and MAE were used to evaluate the performance of each algorithm between the predicted results of $PM_{2.5}$ concentrations from RF, SVM, and XGBoost and the ground measurements of $PM_{2.5}$. The dataset are divided into 10 equal parts, so-called 10-fold cross-validation. Nine parts are used as training data and the remaining part is used as validation data. The predicted values are compared with the observed values where each validation result has its corresponding accuracy, and the average of all the ten validations is used to evaluate the accuracy of each algorithm. The formula for each accuracy statistical value is:

$$R^2 = 1 - \frac{\sum_{n=1}^{n}(O_i - P_i)^2}{\sum_{n=1}^{n}(O_i - \bar{O}_i)^2} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{n}(O_i - P_i)^2}{n}} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{n=1}^{n} |(O_i - P_i) \tag{7}$$

where O_i = observed value
P_i = predicted value
$\bar{O}_i$ = average of observed values
n = number of predicted values

3. Results

3.1 Model Validation

This paper evaluated different machine learning algorithms to estimate $PM_{2.5}$ concentrations in Manila, Philippines from 2020 to 2023 using AOD, $PM_{2.5}$ air quality index (AQI), and meteorological data. The daily average AOD pixel closest to the ground monitoring station was compared to the daily average $PM_{2.5}$ with outlier values not included. Figure 2 shows the correlation of each variable in relation to each other. $PM_{2.5}$ values have a moderate correlation with relative humidity and a moderate negative correlation with wind. Meanwhile, AOD values have a negative moderate correlation with surface pressure and a moderate correlation with rain.

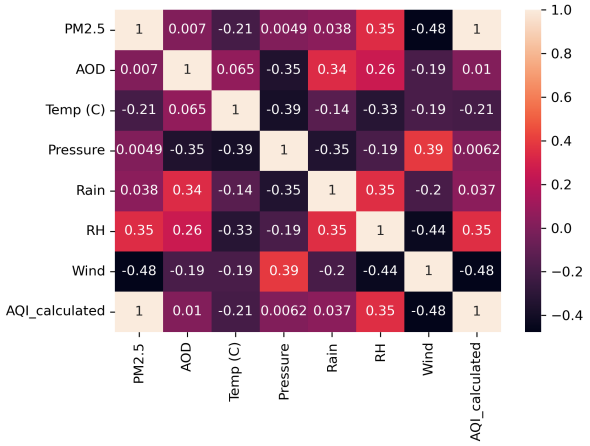

Figure 2. Correlation of variables in relation to each other.

Figure 3, 4, and 5 show the fitting results of RF, SVM, and XGBoost. RF and SVM achieved a very strong correlation between the predicted $PM_{2.5}$ from GEMS and ground $PM_{2.5}$, while XGBoost achieved a strong correlation. Table 1 shows the R^2, RMSE, and MAE values for each model. SVM achieved the highest accuracy value for $PM_{2.5}$ (0.998) prediction followed by RF (0.997) and XGBoost (0.673).

Model	R^2	RMSE	MAE
RF	0.997	0.223	0.050
SVM	0.998	0.180	0.032
XGBoost	0.673	2.23	1.70

Table 1. Accuracy index for RF, SVM, and XGBoost models.

Both SVM and RF achieved very high accuracy values, which suggests that they can accurately predict $PM_{2.5}$ using satellite data. SVM also achieved the lowest RMSE (0.180) and MAE (0.032). The nature of data used may explain the values in the accuracy of each model. Each feature has a distinct boundary, which could explain the slightly higher accuracy of SVM than RF. And since the number of data is not extremely large, this difference is very minimal. So for small datasets, the datatype and computation time will be the key factors in choosing between RF and SVM. In terms of XGBoost, its computational time is fast compared to RF and SVM. It is more efficient for very large and complex datasets. Unfortunately, it also requires parameter tuning for optimal performance, requiring more computational resources. This could be the reason why its accuracy is low compared to RF and SVM, because it needs several adjustments in its hyperparameters.

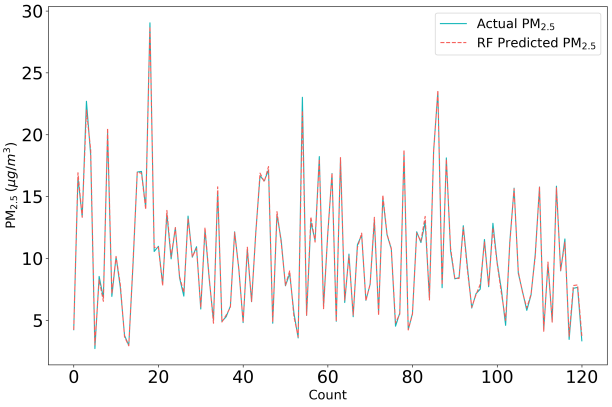

Figure 3. Model fitting performance for RF.

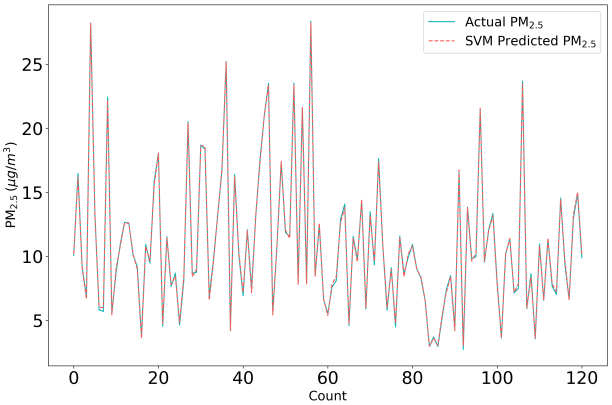

Figure 4. Model fitting performance for SVM.

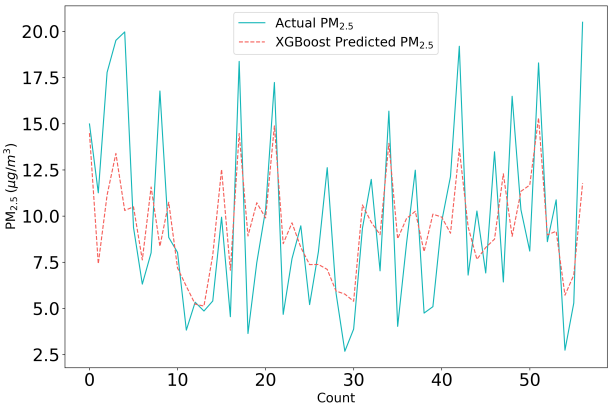

Figure 5. Model fitting performance for XGBoost.

3.2 Feature Importance Analysis

Overall, RF and SVM achieved good performances, meaning that the models were able to accurately predict $PM_{2.5}$ from satellite data. Figure 6 shows the contribution of each feature for predicting $PM_{2.5}$, which agrees with the SHAP values in

Figure 7. For all models, wind speed has the highest contribution in the prediction, followed by temperature, surface pressure, relative humidity, AOD, and rain. As shown earlier in Figure 2, $PM_{2.5}$ has the highest negative correlation value with wind (0.48). Similarly, it is identified as the most important predictor variable among all the features used. Wind has a horizontal and vertical effect on the transport of $PM_{2.5}$, as well as the speed of concentration and diffusion of pollutants. As the wind speed increases, the concentration of $PM_{2.5}$ decreases due to dispersion (Li et al., 2017).

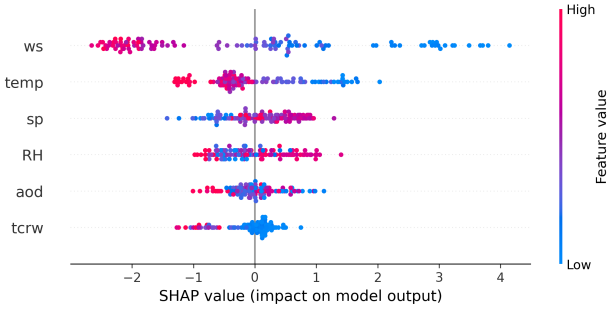

Figure 6. Summary plot for the SHAP analysis for predicting $PM_{2.5}$.

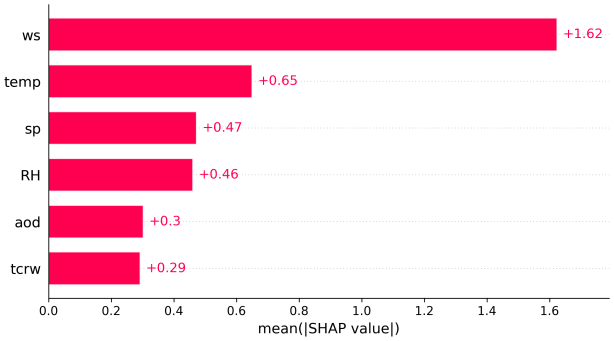

Figure 7. Mean SHAP values of features for predicting $PM_{2.5}$.

4. Conclusion

This study evaluated random forest (RF), support vector machine (SVM), and eXtreme Gradient Boosting (XGBoost) to predict $PM_{2.5}$ in Manila, Philippines. The predictor variables include aerosol optical depth (AOD), temperature, surface pressure, relative humidity, rainfall, and air quality index. Ground monitoring $PM_{2.5}$ data was obtained from 2020 to 2023. Overall, the SVM model performed with the highest accuracy for $PM_{2.5}$ prediction, followed by RF, then XGBoost. SHAP analysis showed the important variables affecting $PM_{2.5}$, which was identified as wind speed. The authors suggest exploring other machine learning models on a bigger dataset for a more robust study. This study highlights the importance of satellite data for continuous spatiotemporal analysis of a large area which can be achieved easier and faster compared to ground monitoring stations, especially in places where resources and manpower are insufficient.

References

Amiri, Z., Zare Shahne, M., 2025. Modeling PM2.5 concentration in tehran using satellite-based Aerosol optical depth (AOD) and machine learning: Assessing input contributions and prediction accuracy. *Remote Sensing Applications: Society and Environment*, 38, 101549.

Kecorius, S., Madueño, L., Vallar, E., Alas, H., Betito, G., Birmili, W., Cambaliza, M. O., Catipay, G., Gonzaga-Cayetano, M., Galvez, M. C., Lorenzo, G., Müller, T., Simpas, J. B., Tamayo, E. G., Wiedensohler, A., 2017. Aerosol particle mixing state, refractory particle number size distributions and emission factors in a polluted urban environment: Case study of Metro Manila, Philippines. *Atmospheric Environment*, 170, 169-183.

Kumar, P., Patton, A. P., Durant, J. L., Frey, H. C., 2018. A review of factors impacting exposure to PM2.5, ultrafine particles and black carbon in Asian transport microenvironments. *Atmospheric Environment*.

Li, X., Feng, Y. J., Liang, H. Y., 2017. The Impact of Meteorological Factors on PM2.5 Variations in Hong Kong. *IOP Conference Series: Earth and Environmental Science*, 78(1), 012003.

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

Rodríguez-Urrego, D., Rodríguez-Urrego, L., 2020. Air quality during the COVID-19: PM2.5 analysis in the 50 most polluted capital cities in the world. *Environmental Pollution*, 266, 115042.

Tantengco, O. A. G., Guinto, R. R., 2022. Tackling air pollution in the Philippines. *The Lancet Planetary Health*, 6, e300.