Thai Journal of **Math**ematics
Volume 22 Number 2 (2024)
Pages 441–452

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

PM 2.5 Prediction & Air Quality Classification Using Machine Learning

Pichitpong Soontornpipit[1], **Lertsak Lekawat**[2], **Chatchai Tritham**[1,2], **Chattabhorn Tritham**[3], **Pornanong Pongpaibool**[4], **Narachata Prasertsuk**[4] **and Wachirapong Jirakitpuwapat**[5,*]

[1] *Department of Biostatistics, Faculty of Public Health, Mahidol University 420/1 Ratchawithi RD., Ratchathewi District, Bangkok 10400, Thailand*
e-mail : pichitpong.soo@mahidol.ac.th (P. Soontornpipit)

[2] *College of Advanced Manufacturing Innovation, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok 10520, Thailand*
e-mail : chatchai.tri@mahidol.ac.th (C. Tritham); lertsak@ine.co.th (L. Lekawat)

[3] *Department of Computer Engineering, Faculty of Engineering, Thammasart University, 99 Moo 18 Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12121, Thailand*
e-mail : memodia@live.com (C. Tritham)

[4] *National Science and Technology Development Agency 111 Thailand Science Park (TSP), Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand*
e-mail : pornanong.pon@nstda.or.th (P. Pongpaibool); narachata.pra@nstda.or.th (N. Prasertsuk)

[5] *Faculty of Science, Energy and Environment, King Mongkut's University of Technology North Bangkok, Rayong 21120, Thailand*
e-mail : wachirapong.j@sciee.kmutnb.ac.th

Abstract Forecasting plays a vital role in air pollution alerts and the management of air quality. Studies and observations conducted in Thailand indicate a concerning rise in pollution levels, particularly in the concentration of PM2.5. Bangkok, in particular, has been flagged for its alarmingly high PM2.5 concentrations. By projecting the future PM2.5 concentrations in these urban areas, we can obtain valuable short-term predictive information regarding air quality. After conducting experiments using four different machine learning algorithms, it was found that the LSTM (Long Short-Term Memory) model provides the most accurate forecasts based on various statistical evaluation indicators. These indicators include a Root Mean Square Error (RMSE) of 2.74, Mean Absolute Error (MAE) of 1.97, R-squared value of 0.94, and Mean Absolute Percentage Error (MAPE) of 10.53. Then the classified air quality based on PM2.5 from the LSTM model gives the best performance indicators including accuracy = 0.9072, precision = 0.8466, negative predict value = 0.9403, sensitivity = 0.8144, specificity = 0.9381, and F1-score = 0.8169. The results show that the machine learning model can predict PM2.5 concentration, which is suitable for early warning of pollution and information provision for air quality management systems in Bangkok.

MSC: 62H30; 62J99; 62P12; 68R01
Keywords: air quality classification; forecast; LSTM; machine learning; PM2.5

Submission date: 06.06.2023 / Acceptance date: 27.06.2024

*Corresponding author.

Published by The Mathematical Association of Thailand.

1. Introduction

Indeed, air pollution poses a substantial threat to development globally, with Asia being one of the regions most affected [1, 2]. The adverse impacts of air pollution extend to various aspects of society and the environment. It affects not only human health but also ecosystems, economies, and the overall quality of life. It encompasses the presence of various contaminants, such as particulate matter, mist, odor, fumes, gases, vapor, or smoke, in both indoor and outdoor atmospheres, with adverse effects on living organisms in terms of quantity, characteristics, and duration [3]. Exposure to particulate matter, specifically PM2.5, has been associated with various health effects. These include respiratory diseases, such as asthma and chronic obstructive pulmonary disease (COPD), as well as cardiovascular problems. Particulate matter can enter the circulatory system, potentially leading to systemic inflammation and cardiovascular complications. Furthermore, emerging research suggests that PM2.5 particles may have detrimental effects on brain function and neurological health [4, 5]. In Asia, the average annual concentration of PM2.5 exceeds that of Europe, North America, and Oceania, ranging from 16 to 58 μg m^{-3}, surpassing Europe and North America [6]. The recent rise in PM2.5 levels in Asia can be primarily attributed to the rapid urbanization and economic development in the region, with Southeast Asia being a significant contributor. The high rate of urbanization in many Asian countries has led to increased industrialization, energy consumption, transportation, and construction activities, all of which contribute to the release of pollutants, including PM2.5, into the atmosphere. Additionally, the growing population and increasing energy demand in Southeast Asian economies have resulted in a rise in emissions from industrial production, biomass burning, and residential activities. These factors, combined with unfavorable meteorological conditions and regional air pollution transport patterns, have led to the elevated levels of PM2.5 observed in Asia [7]. Industrial production, the electricity industry, residential activities, and biomass burning are the primary sources of PM2.5 pollution in Southeast Asia [8]. These emissions are most concentrated in industrial and residential areas within large cities, including the megacities of Southeast Asia, China, and India, which have high population densities, are directly associated with elevated PM2.5 concentrations [9]. The deteriorating ambient air quality in these cities is a harsh reality. PM2.5 pollution in megacities poses a significant risk to the population, exceeding the air quality guidelines set by the World Health Organization (WHO) [10, 11]. To address the issue of air pollution, it is crucial to establish pollution assessment objectives that support legislation aimed at preventing air pollution in these cities. Predicting PM2.5 concentrations is an integral part of action plans to reduce and control polluting activities. The forecasted results provide valuable insights into future PM2.5 concentrations, enabling better response planning and the implementation of measures to mitigate emission increases. Additionally, PM2.5 forecasting serves as a direct benefit by informing and raising awareness among the population about pollution levels.

The air quality problem in Bangkok is progressively worsening, particularly concerning the escalating PM2.5 concentration. Based on data from January 2021 to December 2021, the mean PM2.5 concentration in Bangkok is 25.004191 μg m^{-3}, with a standard deviation of 16.601135. The continuous increase in PM2.5 levels and its adverse effects on human health highlight the urgent need for a solution in Bangkok. Consequently, it

becomes crucial to forecast the level of PM2.5 to effectively prevent or minimize the risk of exposure.

Numerous research studies have been conducted on air quality prediction. These studies have explored the application of machine learning models that incorporate meteorological and emission data [12–14]. For instance, research conducted in Taiwan utilized machine learning regression models to predict PM2.5 levels based on emissions data, and the predicted values closely matched the actual values [15]. These predictive results hold substantial significance, particularly in health studies, considering the significant impact of PM2.5 concentrations on human health [16].

This study aims to introduce a straightforward, efficient, and precise machine learning approach for predicting PM2.5 concentrations, focusing on the case study in Bangkok. The research findings will contribute novelty by enabling the prediction of PM2.5 pollution not only in Bangkok but also in Thailand as a whole. This outcome serves as a foundational step for subsequent projects centered around pollution forecasting, developing an application to alert the public, and suggesting potential measures to mitigate pollution in Bangkok. Considering the evaluation of prior studies conducted in Thailand, this research will introduce a novel application of machine learning for pollution prediction specifically tailored to the context of Bangkok.

2. Data and Methodology

PM2.5 pollution has been a persistent issue for several years. However, the concerning aspect of the current year is that the onset of the problem happened earlier than the regular cycle, and the critical conditions have persisted for a longer duration compared to previous years. This development has raised significant concern and emphasizes the necessity to increase awareness among the general public. Forecasting PM2.5 concentrations can serve as an early warning system for pollution, providing crucial information for effective air quality management systems. By utilizing forecasted PM2.5 data, authorities can take proactive measures to mitigate pollution and minimize its impact on public health.

2.1. Study Area and Data Set

In this study, the dataset used for analysis includes both meteorological data and PM2.5 concentrations. The meteorological data encompasses several variables, namely temperature, wind speed, relative humidity, and blood pressure. This dataset was collected from the Air Quality and Noise Management Division Environment Bureau Bangkok, as depicted in Figure 1. A visual representation of the distribution of a dataset, showing the median, quartiles, and any outliers or extreme values is shown in Figure 2. The square matrix displays the pairwise correlations between variables in a dataset as shown in Figure 3.

On the other hand, the p-value significance level is a statistical measure used to determine the significance of the observed correlation coefficient. It helps assess whether the correlation between two variables is statistically significant in Figure 4. For the machine learning experiment conducted in this study, the data used consists of hourly observations spanning a specific time period. The training data covers the time period from April 1st, 2020, to June 13th, 2022. This data is utilized for training the machine learning model, allowing it to learn patterns and relationships between the input variables (meteorological data and PM2.5 concentrations) during this time period. On the other hand, the testing

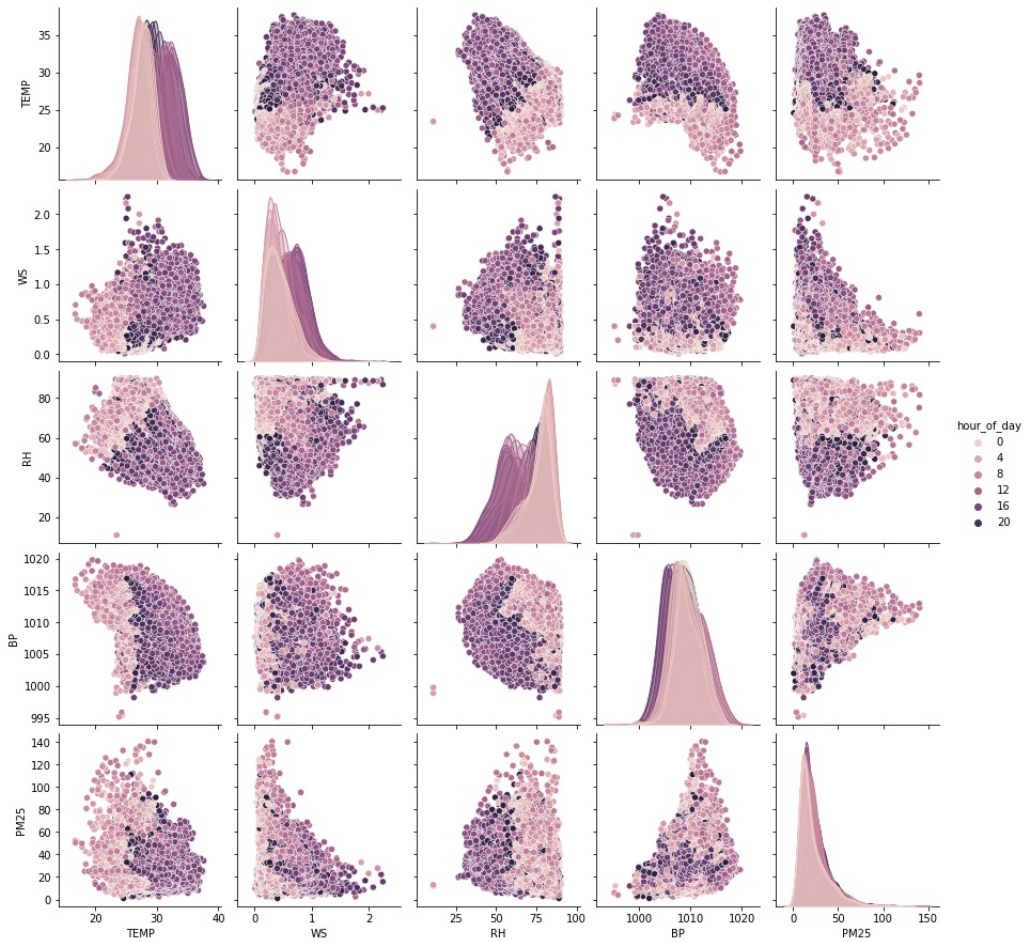


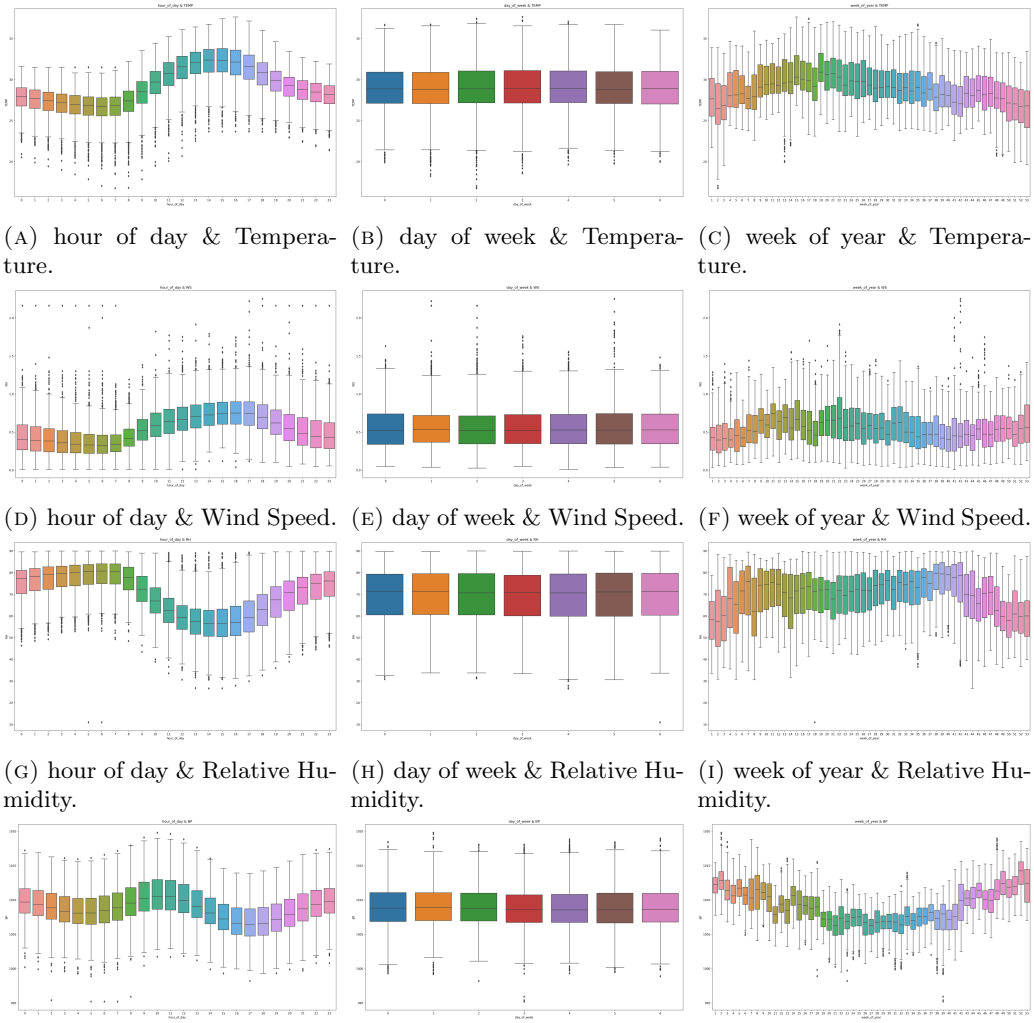
FIGURE 1. Data.

data is separate and distinct from the training data. It covers the time period from June 14th, 2022, to December 31st, 2022. The testing data is used to evaluate the performance and predictive capabilities of the trained machine learning model.

2.2. MACHINE LEARNING

To assess efficiency and determine the optimal predictive algorithm, various machine learning models are executed using different algorithms. The machine learning algorithms are implemented in Python [17] with the utilization of the Scikit-Learn library [18]. The algorithms employed include the Least Absolute Shrinkage and Selection Operator (LASSO) [19], Ridge [20], Extreme Gradient Boosting (XGBoost) [21], and Long Short-Term Memory (LSTM) [22]. It is worth noting that all these algorithms are regression models, meaning they provide precise predictive values as outcomes.

The machine learning model's input data consists of two types of information: meteorological data and PM2.5 concentration data. The meteorological data includes variables



(A) hour of day & Temperature. (B) day of week & Temperature. (C) week of year & Temperature.

(D) hour of day & Wind Speed. (E) day of week & Wind Speed. (F) week of year & Wind Speed.

(G) hour of day & Relative Humidity. (H) day of week & Relative Humidity. (I) week of year & Relative Humidity.

(J) hour of day & Blood Pressure. (K) day of week & Blood Pressure. (L) week of year & Blood Pressure.

FIGURE 2. Visual representation data.

such as temperature, wind speed, relative humidity, and blood pressure. These meteorological factors are known to influence air quality and can provide valuable insights for predicting PM2.5 concentrations. The dataset used for training the machine learning model spans a period of two years and nine months.

To ensure an effective evaluation of the model's performance, the dataset is split into two separate parts: a training period and a testing period. Approximately 80% of the data is allocated for the training period, allowing the model to learn and establish patterns from a substantial portion of the dataset. The remaining 20% of the data is reserved for the testing period, which serves as a means to evaluate the model's performance on unseen data.

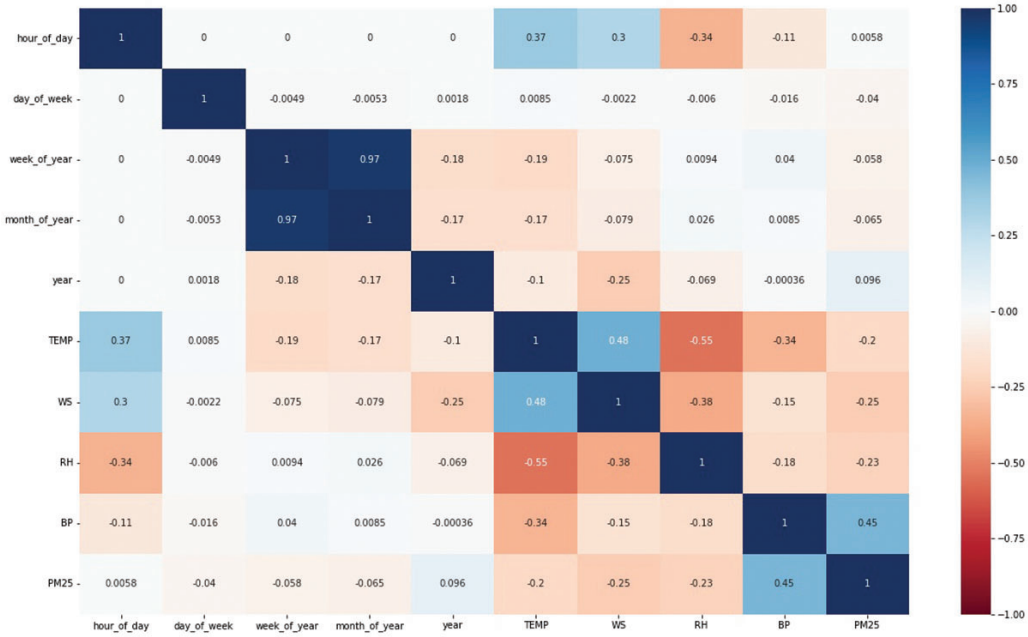


FIGURE 3. Correlation matrix.

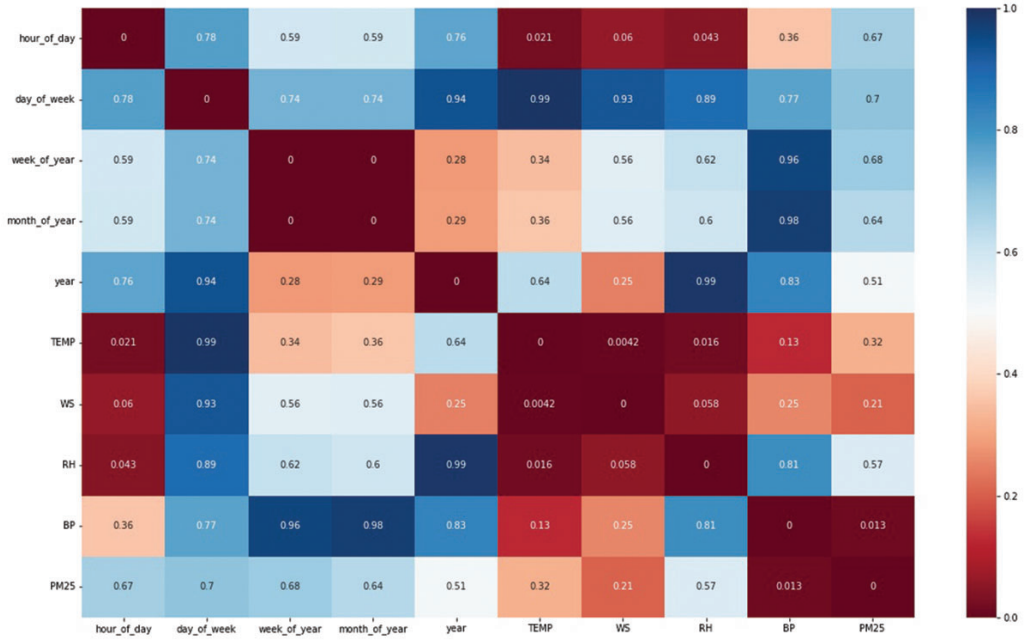


FIGURE 4. p values.

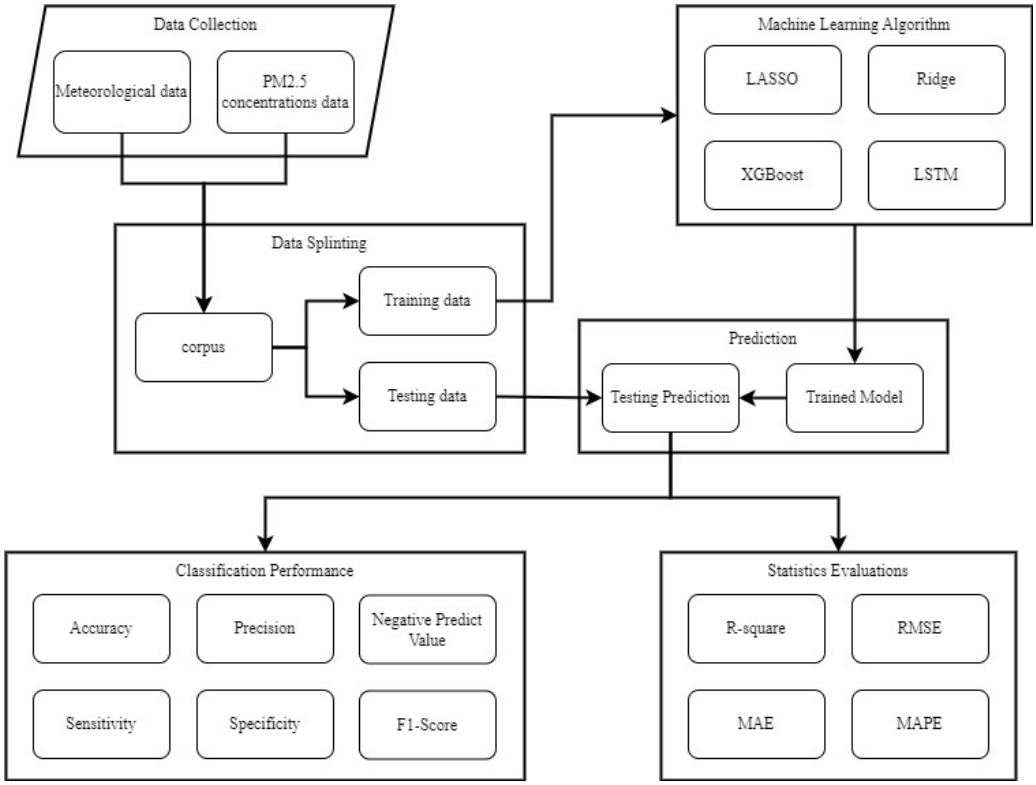


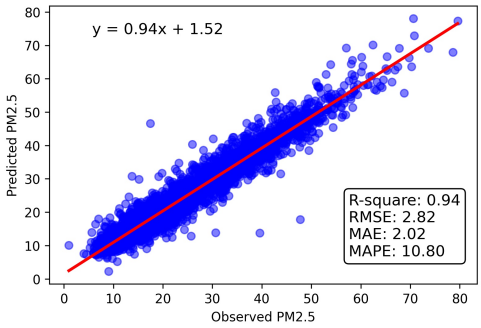
FIGURE 5. Machine learning flow diagram.

Starting with the prepossessing data, the procedure divides it into two sets: the training set and the test set. The machine learning model is then trained using the training set and all four considered algorithms. In this step, the model is fed training data, and it is left to discover the underlying patterns and correlations between the goal variable (PM2.5 concentrations) and the input variables (meteorological data and PM2.5 concentrations). Using the test set, the machine learning models' training efficiency is assessed after they have been trained using the training set. The test set indicates how effectively the model generalizes to new, unknown data because it contains data that the model did not observe during the training phase. To evaluate each algorithm's performance, the actual values of PM2.5 concentrations are compared with the predictions made by the model on the test set as shown in Figure 5.

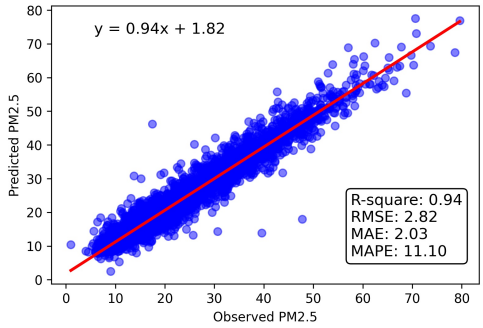
3. Result

3.1. Machine Learning Model Evaluation

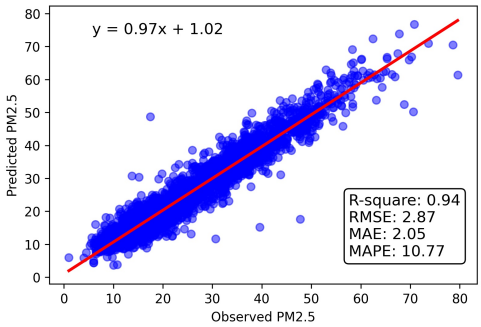
To evaluate the performance of the models, specific evaluation parameters are used. These parameters may include metrics such as the coefficient of determination (R-square), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics provide insights into how well the models are able to



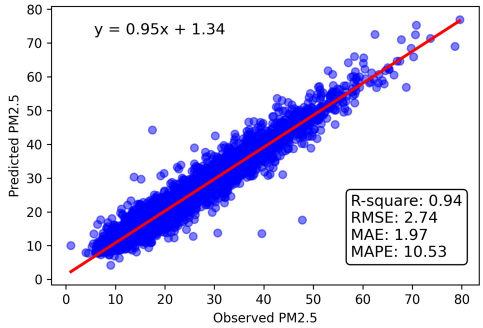
(A) LASSO.



(B) RIDGE.



(C) XGBOOST.



(D) LSTM.

FIGURE 6. Scatter and fitted plots of observed and predicted PM2.5 concentrations of different models.

predict PM2.5 concentrations and the extent of any errors or deviations from the actual values. The performance of the different models is shown in Figure 6. Additionally, the machine learning models in this study are capable of predicting PM2.5 concentrations for multiple days in the future, as depicted in Figure 7, which illustrates the forecast results for a seven-day period.

3.2. Air Quality Classification

To evaluate the predictive performance of the PM2.5 model, statistical indicators are used to assess the accuracy of specific forecast values. However, it's important to note that the errors may sometimes be significant, making it challenging to properly assess the model's performance. Additionally, since the observed data is based on hourly averages, the forecasting model's error can be higher than the actual values. To overcome these challenges, a confusion matrix can be employed to report the results and evaluate the performance of the classification model. The confusion matrix allows for the observation of the relationships between the model's outputs and the true values. In this study, the confusion matrix is optimized using the health impacts of PM2.5 [23]. The PM2.5 breakpoint scale consists of seven categories, and the forecast results are classified based on this scale, as shown in Table 1. The observed data is then compared to the forecasted

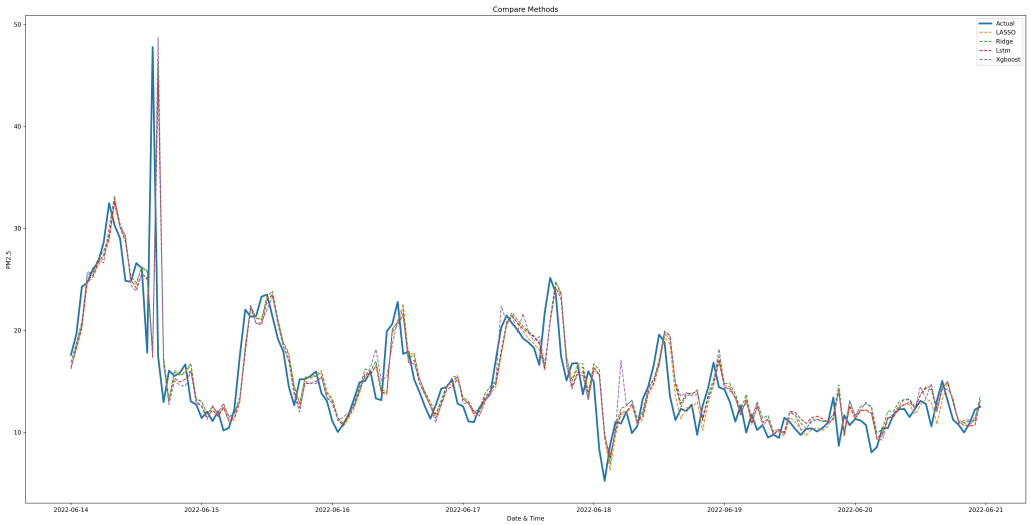
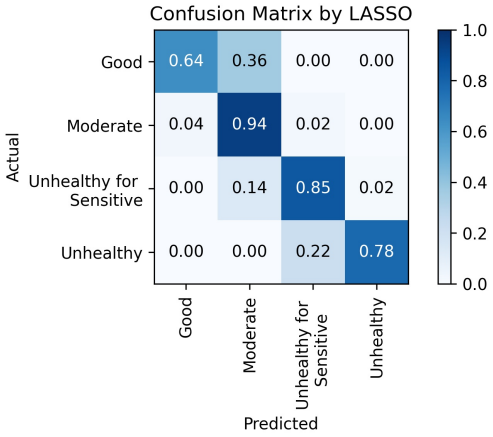


FIGURE 7. Comparison of the methods.

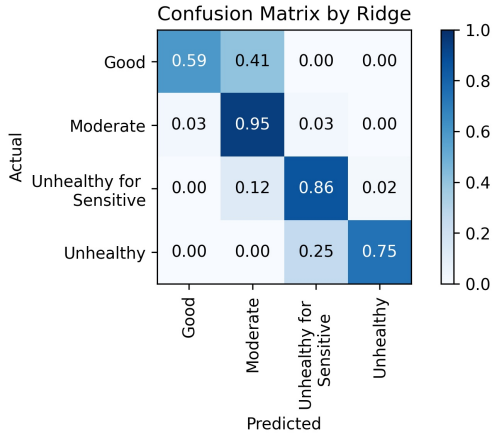
results using the confusion matrix, as shown in Figure 8. To evaluate the classification model, various performance indices are used. These include accuracy, precision, negative predictive value, sensitivity, specificity, and F1-score. Accuracy measures the proportion of correct predictions made by the model out of the total number of predictions. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The negative predictive value represents the proportion of true negative cases among those predicted as negative by the model. Sensitivity measures the ratio of correctly predicted positive observations to all observations in the actual class. Specificity represents the proportion of true negative cases correctly identified by the model out of the total number of actual negative cases. The F1-score is the weighted average of precision and recall, and it is particularly useful when dealing with imbalanced class distributions. The macro-average approach computes the metric independently for each class and then takes the average, treating all classes equally.

TABLE 1. Specific breakpoints for PM2.5 concentrations to determine the corresponding AQI levels.

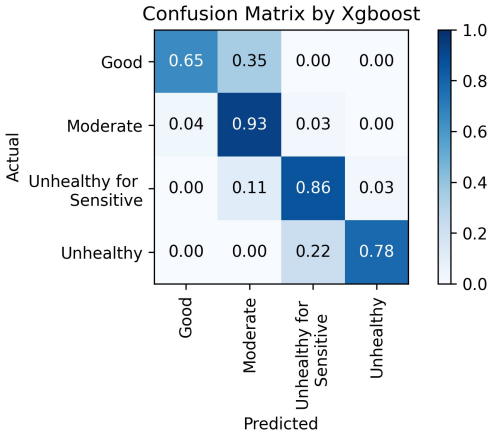
Air Quality Levels	Lower Category PM2.5	Upper Category PM2.5
Good	0	12
Moderate	12.1	35.4
Unhealthy for sensitive	35.5	55.4
Unhealthy	55.5	150.4
Very unhealthy	150.5	250.4
Hazardous	250.5	350.4



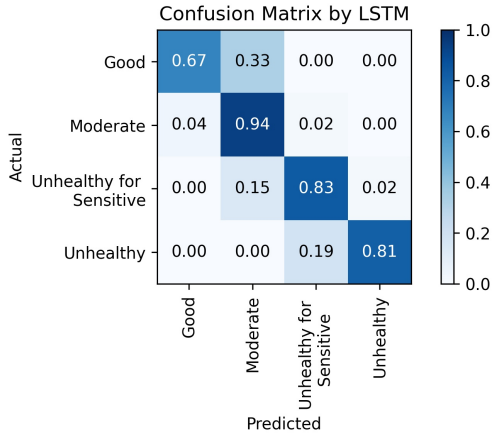
(A) LASSO.



(B) RIDGE.



(C) XGBOOST.



(D) LSTM.

FIGURE 8. models's confusion matrix.

TABLE 2. Performance of classification.

Method	Performance					
	Accuracy	Precision	Negative Predict Value	Sensitivity	Specificity	F1-score
LASSO	0.9007	0.8394	0.9367	0.8014	0.9338	0.8031
Ridge	0.8932	0.8326	0.9328	0.7865	0.9288	0.7867
Xgboost	0.9031	0.8389	0.9379	0.8061	0.9354	0.8074
LSTM	0.9072	0.8466	0.9403	0.8144	0.9381	0.8169

4. CONCLUSION

This study aims to develop a machine learning model that is effective, efficient, and accessible for predicting future PM2.5 concentrations. By leveraging the power of machine

learning algorithms and utilizing a comprehensive dataset that includes meteorological data and historical PM2.5 concentrations, the model aims to provide accurate and reliable predictions. By utilizing a machine learning algorithm and a sufficiently large input dataset, a stable and high-performing predictive model can be developed. When combined with forecasted meteorological data, this machine learning model can provide short- to medium-term predictions of PM2.5 concentrations on an hourly basis. Although LASSO and Ridge are often utilized for regression issues, by constructing lagged features, they can be modified for time series forecasting. They might not be able to capture intricate patterns in time series data, though, due to their linear structure. For time series data, XGBoost is more potent than linear models because it can handle non-linear correlations and interactions between lagged components. LSTM is among the models studied. LSTM has the highest performance metrics because it is specifically made for sequential data and is particularly good at capturing temporal patterns and long-term dependencies. LSTM exhibits strong performance with evaluation indicators including an R-squared value of 0.94, RMSE of 2.74, MAE of 1.97, MAPE of 10.53, and an accuracy of 90.72 based on the confusion matrix. Moreover, the future PM2.5 concentration prediction model aligns well with meteorological data, yielding results similar to those obtained from observed meteorological data.

Acknowledgements

Authors thank for data support by the Air Quality and Noise Management Division Environment Bureau Bangkok. This research was funded by the National Security and Dual-Use Technology Center: NSD from National Science and Technology Development Agency, Thailand.

References

[1] World Bank, Institute for Health Metrics and Evaluation and World Bank, The Cost of Air Pollution: Strengthening the Economic Case for Action, World Bank Group, IHME, 2016

[2] E. Marsden, G. Bathan, B. Tsevegjav, M.C.R. Velez, Making Urban Asia's Air Cleaner (ADBBriefs), Asian Development Bank, Manila, Philippines, 2019.

[3] H.E. Hesketh, Introduction to Air Pollution, in: L.K. Wang, N.C. Pereira, (Eds.), Air and Noise Pollution Control, Humana Press, Totowa, NJ (1979), 3–39.

[4] D.P. Croft, W. Zhang, S. Lin, S.W. Thurston, P.K. Hopke, M. Masiol, S. Squizzato, E. van Wijngaarden, M.J. Utell, D.Q. Rich, The association between respiratory infection and air pollution in the setting of air quality policy and economic change, Ann. Am. Thorac. 16 (2019) 321–330.

[5] X. Zhang, J. Kang, H. Chen, M. Yao, J. Wang, PM2.5 meets blood: In vivo damages and immune defense, Aerosol Air Qual. Res. 18 (2018) 456–470.

[6] M. Crippa, G. Janssens-Maenhout, D. Guizzardi, R. Van Dingenen, F. Dentener, Contribution and uncertainty of sectorial and regional emissions to regional and global PM2.5 health impacts, Atmos. Chem. Phys. 19 (2019) 5165–5186.

[7] D. Yang, C. Ye, X. Wang, D. Lu, J. Xu, H. Yang, Global distribution and evolvement of urbanization and PM2.5 (1998–2015), Atmos. Environ. 182 (2018) 171–178.

[8] T. Amnuaylojaroen, J. Inkom, R. Janta, V. Surapipith, Long range transport of southeast Asian PM2.5 pollution to northern Thailand during high biomass burning episodes, Sustainability 12 (2020) 10049.

[9] L. Zhang, J.P. Wilson, B. MacDonald, W. Zhang, T. Yu, The changing PM2.5 dynamics of global megacities based on long-term remotely sensed observations, Environ. Int. 142 (2020) 105862.

[10] M. Krzyzanowski, J.S. Apte, S.P. Bonjour, M. Brauer, A.J. Cohen, A.M. Prüss-Ustun, Air pollution in the megacities, Curr. Environ. Health Rep. 1 (2014) 185–191.

[11] M.E. Marlier, A.S. Jina, P.L. Kinney, R.S. DeFries, Extreme air pollution in global megacities, Curr. Clim. Change Rep. 2 (2016) 15–27.

[12] J. Du, F. Qiao, L. Yu, Temporal characteristics and forecasting of PM2.5 concentration based on historical data in Houston, USA. Resour. Conserv. Recycl. 147 (2019) 145–156.

[13] B. Zhai, J. Chen, Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. Sci. Total Environ. 635 (2018) 644–658.

[14] J. Zhang, W. Ding, Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong, IJERPH 14 (2017) 114.

[15] Doreswamy, K.S. Harishkumar, K.M. Yogesh, I. Gad, Forecasting air pollution particulate matter (PM2.5) using machine learning regression models, Procedia Comput. Sci. 171 (2020) 2057–2066.

[16] D.J. Lary, T. Lary, B. Sattler, Using machine learning to estimate global PM2.5 for environmental health studies, Environ. Health Insights 9s1 (2015) EHI.S15664.

[17] G. Van Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[18] F. Pedregosa et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[19] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. B 58 (1996) 267–288.

[20] D.E. Hilt, D.W. Seegrist, Ridge, a computer program for calculating ridge regression estimates (1977) https://doi.org/10.5962/bhl.title.68934.

[21] M. Chen, Q. Liu, S. Chen, Y. Liu, C.H. Zhang, R. Liu, XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system, IEEE Access 7 (2019) 13149–13158.

[22] S. Hochreiter; J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[23] U.S. EPA, Technical assistance document for the reporting of daily air quality: The Air Quality Index (AQI), September 2018 [edition], U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division, Research Triangle Park, NC, USA, 2018.