Glovento Journal of Integrated Studies (GJIS) | ISSN: 3117-3314
Volume 1 (2025) | Article 6

A Hybrid Deep Learning Model for Forecasting PM2.5 Concentrations in Northern Thailand from Satellite Images

[1]Chutinun Potavijit, [2]Parichart Pattarapanitchai, [3,*]Chalermrat Nontapa

[1,2,3]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
Email: [1]chutinun_pota@cmu.ac.th, [2]parichart.p@cmu.ac.th, [3]chalermrat.n@cmu.ac.th,
*Corresponding Author: Chalermrat Nontapa, (Email: chalermrat.n@cmu.ac.th)

Abstract- Air pollution is a significant environmental issue with extensive impacts, particularly concerning particulate matter smaller than 2.5 microns (PM2.5), which poses serious public health risks,especially respiratory diseases such as various diseases, ischemic heart disease, strokes, chronic obstructive pulmonary disease, tracheal, bronchus, lung cancer, and even increased premature death rates. Northern Thailand is one of the areas with the most severe PM2.5 problems, especially during the summer (February to May), primarily due to the large amount of agricultural field burning and forest fires by ethnic groups after the harvest season.This research proposes a hybrid model of Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) for PM2.5 concentration forecasting using satellite images of four environmental variables: aerosol optical depth, temperature, precipitation, and ozone. These variables are important factors in the occurrence of PM2.5. The efficiency of the CNN-LSTM model was assessed by comparing performance with classification deep learning models (CNN, LSTM), Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), and Multiple Linear Regression (MLR). The findings indicate that The CNN-LSTM model achieves higher accuracy than the other models, achieving an R2 of 98.38%, MAPE of 2.47%, and significantly lower RMSE (3.0672 µg/m3) and MAE (0.8560 µg/m3). In conclusion, this research highlights the important implications of supporting government policy formulation and public preparedness to address the PM2.5 problem, which varies in severity across seasons.

Keywords: Air Pollution, PM2.5, Deep Learning, Satellite Images, Air quality forecasting, Northern Thailand.

I. Introduction

In recent years, PM2.5 has emerged as an air pollution issue that requires immediate attention because PM2.5 has widespread impacts, including significant health risks, particularly respiratory diseases. Additionally, PM2.5 has considerable economic effects by increasing the social costs associated with goods affected by PM2.5 pollution.The occurrence of PM2.5 involves both direct and indirect processes. Direct factors contributing to PM2.5 include pollutants released directly from industrial sources or human activities, such as industrial factories, power generation, transportation, and wildfires. Indirect factors refer to complex chemical reactions in the atmosphere influenced by environmental variables, including air pollutants and meteorological factors such as aerosol optical depth (AOD), precipitation, ozone (O_3), and temperature.

In Thailand, the PM2.5 problem is particularly severe in the northern region.The most common causes include forest fires during the summer (February to May), the burning of fields, and agricultural forests by local ethnic groups after the harvest season.These activities result in high levels of PM2.5, which significantly affect the health of residents.For example, Chiang Mai, one of the northern provinces, has been ranked among the most polluted cities in the world. During the dry season, the air pollution levels in Chiang Mai often exceed World Health Organization (WHO) standards by up to 20 times(World Health Organization, 2021). This period, often referred to as the "smog season," is caused by agricultural burning and forest fires set to prepare land for new plantations.

Currently, air quality monitoring has seen numerous innovations, such as air quality monitoring stations, laser scattering methods, portable sensors, portable air quality monitors, data from applications and websites, and satellite imagery. However, using satellite imagery effectively for monitoring air quality over large areas and providing continuous, verifiable data remains a challenge

(UNEP, 2022). This is because satellite image analysis is still complex and requires advanced technology for analyzing such data.One promising approach is the application of deep learning models, especially Convolutional Neural Networks (CNNs), which are specifically designed for analyzing image data or extracting deep features from large and complex datasets with high efficiency. Additionally, Long Short-Term Memory (LSTM) models possess unique capabilities to retain and learn sequential data relationships, such as time-series data, and capture complex temporal relationships, such as variations in PM2.5 levels associated with meteorological factors and other variables. Previous studies (Ahmed et al., 2022; Kristiani et al., 2022) have demonstrated that CNNs and LSTM models are highly effective in accurately predicting PM2.5 concentrations. Therefore, developing and optimizing these deep learning classification models can enhance the accuracy of PM2.5 forecasts across various regions, particularly in areas lacking air quality monitoring stations or other high-cost air quality measurement tools.

The hybrid model is a highly promising approach to enhancing the precision of PM2.5 forecasting by combining the strengths of different models. For instance, Zhang et al. (2021) and Liu et al. (2022) used CNN to extraction spatial features from satellite imagery, allowing them to account for variations in PM2.5 levels due to spatial differences. Additionally, LSTM networks were used to capture the temporal relationships in the data, learning the trends and variations in PM2.5 concentrations over time. To increase the accuracy of the model, Liu et al. (2022) incorporated both ground station data and satellite imagery (multi-source data) in training the model, enabling the model to learn from both ground-based and satellite-based data simultaneously. However, there are some limitations that affect the accuracy of the predictions. First, the ability to capture complex spatial and temporal relationships with high uncertainty remains a challenge, particularly in areas with rapid changes, such as regions with heavy traffic or sudden weather changes. Second, the limited number of ground monitoring stations that do not cover all areas may result in errors in regions without monitoring data.

This research aims to develop a new hybrid deep learning model combining CNN and LSTM to enhance the performance of classification models in forecasting PM2.5 concentrations in northern Thailand. The model utilizes satellite imagery data of four environmental variables: aerosol optical depth, temperature, precipitation, and ozone. The proposed model's performance will be compared to that of CNN, LSTM, SARIMAX, and MLR models. This approach addresses the limitations of current air quality monitoring methods and improves forecasting accuracy.

II. Methodology

A. CNNs

CNN's area category of Artificial Neural Networks (ANNs) developed by LeCun et al. (1989). In the architecture of CNN, specific layers are added, including convolutional layers that extract significant characteristics from the data, and pooling layers that decrease the data dimensions obtained from convolutional layers to reduce the number of parameters and unnecessary computations. These layers work in conjunction with fully connected layers present in the hidden layers of ANNs to enhance the processing efficiency of complex data structures such as images or videos, which are commonly used in applications like medical imaging and biometric user identity authentication.The important layers in CNNs are the convolution layers and pooling layers.

The convolution operation is defined by the following equation:

$$C(m,n) = (I * F)(m,n) = \sum_i \sum_j F(i,j) \cdot I(m-i, n-j) \quad (1)$$

where I is the input image, *F* is the kernel (filter), and $C(m,n)$ is the output feature map in the position of (m,n).

Max Pooling is widely utilized, and its function can be expressed with the following formula:

$$P(m,n,z) = \max_{(i,j) \in R_{m,n}} C(i,j,z) \quad (2)$$

where $P(m,n,z)$ is the result of the pooling operator at position (m,n) in k -th feature map, and $C(i,j,z)$ is the feature value at position (i,j) inside the pooling region $R_{i,j}$ in z-th feature map.

$$c_t = w_i \cdot x_i + e_i \quad (3)$$

where c_t is the output vector, x_i are the feature maps after layer i and x_0 is the input vector, w_t is the weight vector, and e_i is the bias vector.

A. LSTM

LSTM is a category of recurrent neural network (RNN) proposed to analyze sequential data and capture long-term citation. Unlike traditional RNN, which is unable to summarize long-term dependencies because of the vanishing gradient issue, the LSTM model has a unique architecture that enables them to update information regarding long-term citation (Li, Hua, & Wu, 2020). The structure of an LSTM model consists of multiple cells, each containing three important components: forget gate, input gate, and output gate. Generally, the steps of an LSTM model unit at time t can be carried out as follows:

The forget gate, which decides the amount of past information to retained or discarded.

$$f_t = \sigma(w_t \cdot [p_{t-1}, x_t] + e_f] \quad (4)$$

Where the values obtained from the Sigmoid function range between 0 and 1. w_t is the weight matrix, p_{t-1} represents the output of the prior cell state, x_t is the input to the cell state at time t, and e_f is the bias term.

The Input Gate handles the reception of new information and recording into each node. The first part involves checking whether the update of the cell state needs to be updated or not.

$$i_t = \sigma(w_i \cdot [p_{t-1}, x_t] + e_i) \quad (5)$$

The second part generates $\tilde{C}_t$ (Candidate Values) in the state when the input gate assign whether to update the cell state, using the Tanh function.

$$\tilde{C}_t = \tanh(w_c \cdot [p_{t-1}, x_t] + e_i) \quad (6)$$

The Output Gate is responsible for preparing the processed information by using the Sigmoid function (σ) to select which information will be sent out. The final step is then passed through the Tanh function before being output.

$$o_t = \sigma(w_o \cdot [p_{t-1}, x_t] + e_o) \quad (7)$$

$$p_t = o_t \cdot \tanh(\tilde{c}_t) \quad (8)$$

Where, o_t is the Output Gate value, $\tilde{c}_t$ is the present cell state, and $\tanh()$ is the Tanh function

B. Proposed Method

This research proposes a CNN-LSTM model, a hybrid model that combines CNN and LSTM networks, to forecast PM2.5 concentrations in Northern Thailand for satellite imageries. First, four environmental variables (temperature, AOD, precipitation, and O_3) are input into the CNN. Then use CNN to analyze and extracts spatial features from these datasets to generate monthly PM2.5 concentration outputs. Next, the estimation errors from the CNN are analyzed using the LSTM, which is capable of handling sequential data to identify patterns in the errors. Finally, the PM2.5

Glovento Journal of Integrated Studies (GJIS) | ISSN: 3117-3314
Volume 1 (2025) | Article 6

concentrations obtained from the CNN are combined with the errors analyzed by the LSTM to enhance prediction accuracy. Can be expressed as an equation like this:

$$y_t = c_t + h_t + \varepsilon_t \quad (9)$$

where y_t is the PM2.5 concentration (monthly), c_t is the PM2.5 concentration at time t (monthly) as forecast by CNN, h_t is the error estimated by LSTM, and ε_t is the random bias. The CNN-LSTM model chart is as follows:

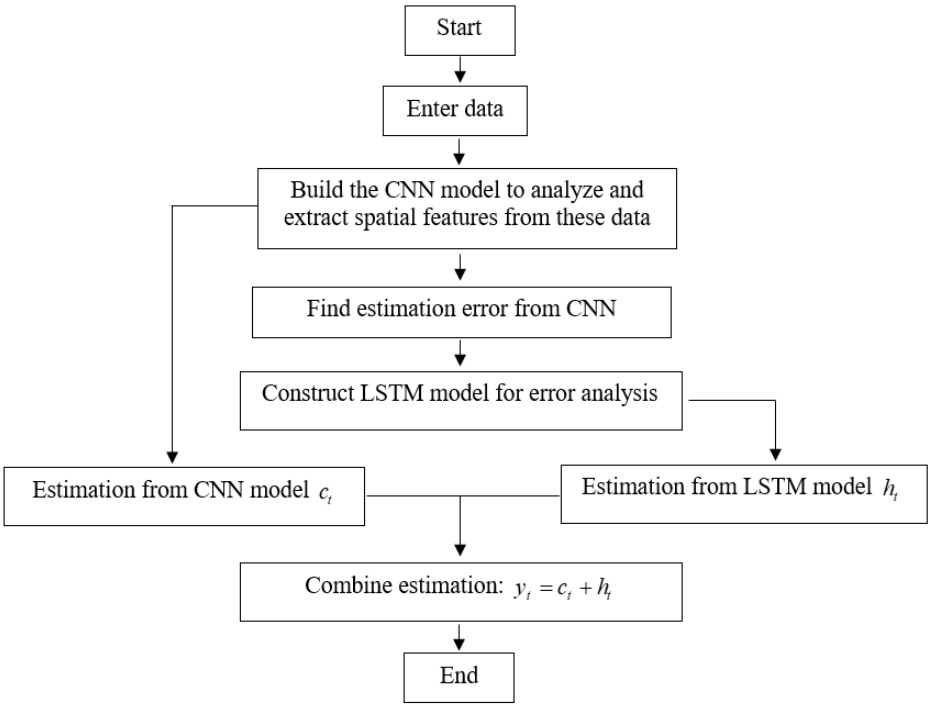

Figure 1: Architecture of CNN-LSTM model

D. Data Collection

The data used in this research includes satellite imagery of environmental variables affecting the occurrence of PM2.5, including AOD, temperature, precipitation, and O3, is collected from Google Earth Engine. These data have been processed and stored in high-resolution GeoTIFF files, with each image having dimensions of 448 by 372 pixels.. Additionally, PM2.5 concentration data collected from ground monitoring stations is also included. These datasets are gathered monthly, from January 2016 to December 2023, covering the research area in Northern Thailand, which includes Chiang Mai, Chiang Rai, Lamphun, Lampang, Phrae, Nan, Phayao, Mae Hong Son, and Uttaradit.

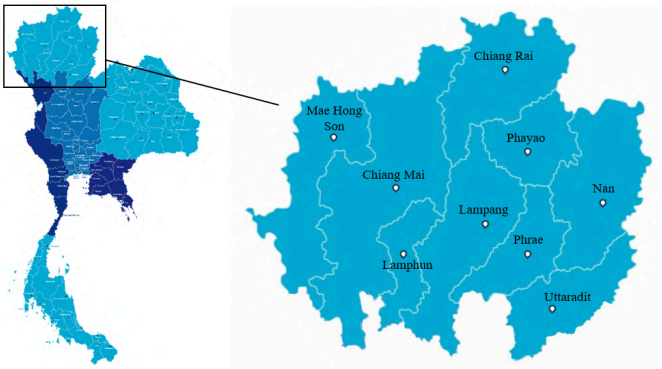

Figure 2: Research area

The ozone data is collected from the Total Ozone Mapping Spectrometer and the Ozone Monitoring Instrument aboard the Aura satellite (NASA, 2023). The AOD data is gathered from the MODIS Terra and MODIS Aqua satellites using the MAIAC algorithm (NASA, 2023). Temperature and precipitation data are sourced from TerraClimate, a dataset that provides monthly climate predictions using data from several sources (University of Idaho, 2023).
PM2.5 concentrations data collected from the Air4Thai website (Pollution Control Department, n.d.) provides monthly concentration measurements of PM2.5 from ground monitoring stations across Northern Thailand. These data are collected from various sensors that provide reliable measurements for monitoring air quality in the country.

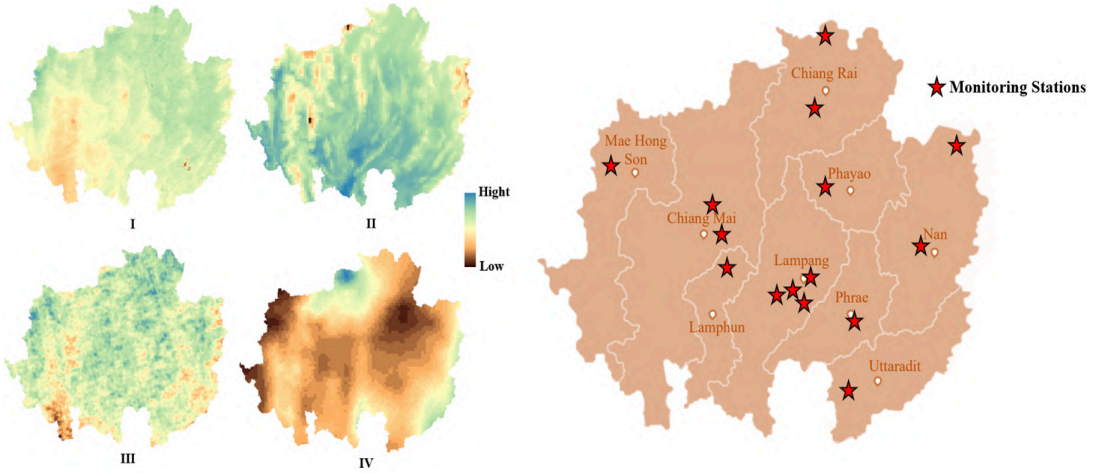

Figure 3: Satellite Images of Environmental Variables for March 2021 and Map of Monitoring Stations in Northern Thailand: (I) AOD, (II) Temperature, (III) Ozone, (IV) Precipitation

E. Data Preprocessing

The PM2.5 concentration data from the Air4Thai website is cleaned by removing missing data. The data is then organized by month and year, creating new features for each monitoring station monthly for each year.

The satellite imagery, stored in GeoTIFF format with consistent dimensions of 448x372 pixels, is processed using the 'rasterio' library. After organizing the time-series data, missing values in the satellite images are removed to maintain data integrity. Subsequently, values are extracted from each satellite image variable based on the latitude and longitude coordinates of the air quality monitoring stations that PM2.5 dataset. These extracted values are then used to create new features, enabling accurate linkage between the two data sources.
In this study, the dataset is split into two segments: 70% is used for training the model, while 30% is allocated for testing, model development, and performance evaluation.

Furthermore, before applying the MLR and SARIMAX models for performance comparison, we conducted an analysis of the Variance Inflation Factor (VIF) to evaluate multicollinearity between the independent variables (AOD, Ozone, Temperature, and Precipitation). Only variables with a VIF of less than 5 were included in the models to avoid multicollinearity issues.

Table 1: VIF for Independent Variables

Variables	VIF
AOD	1.9204
Ozone	2.4748
Temperature	2.6789
Precipitation	1.9496

Additionally, stationarity of the data was examined using the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

Table 2: Stationarity Test Results

Test	Test Statistic	p-value	Alternative Hypothesis
ADF Test	-4.3528	0.01	stationary
KPSS Test	0.14064	0.10	stationary

As shown in Table 1, the VIF values for all variables were below the acceptable threshold of 5, indicating no significant multicollinearity issues. Furthermore, Table 2 highlights that all variables satisfy the stationarity criteria under both ADF and KPSS tests.

F. Model Evaluation Metric

- Mean Absolute Percentage Error (MAPE) is a metric that assesses a model's accuracy by computing the average absolute percentage errors.

$$MAPE(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (8)$$

- Mean Absolute Error (MAE) is a metric that measures a model's accuracy by computing the average absolute error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (9)$$

- Root Mean Square Error (RMSE) is a metric that assesses a model's accuracy by taking the square root of the average squared errors.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (10)$$

- Coefficient of Determination (R^2) measures how well independent variables explain dependent variable variance.

$$R^2(\%) = \left[1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}\right] \times 100 \quad (4)$$

III. Results and Discussion

The evaluation results of the CNN-LSTM model, compared with CNN, LSTM, MLR, and SARIMAX models for monthly PM2.5 concentration forecasting in Northern Thailand, are shown in Table 3. These results are based on satellite imagery data of four environmental variables (AOD, Ozone, Temperature, and Precipitation) collected from January 2016 to December 2023. The experiments were conducted using Python.

Table 3: Model performance for monthly PM2.5 forecasting in Northern Thailand using satellite data

Models	MAE	MAPE (%)	RMSE	R^2 (%)
CNN	6.0549	24.19	11.1219	78.75
LSTM	6.5831	25.69	11.9246	77.63
MLR	7.2866	31.31	11.9516	75.46
SARIMAX	6.7979	33.74	9.7596	82.48
CNN-LSTM	0.8560	2.47	3.0672	98.38

As shown in the evaluation results in Table 3, the CNN-LSTM model outperforms other models in forecasting monthly PM2.5 concentrations in Northern Thailand, achieving the best results across all

evaluation metrics, with an MAE of 0.8560, MAPE of 2.47%, RMSE of 3.0672, and R2 of 98.38%. Additionally, when considering individual models, the CNN model performs better than LSTM, MLR, and SARIMAX, demonstrating that the CNN model is highly effective in forecasting image datasets.

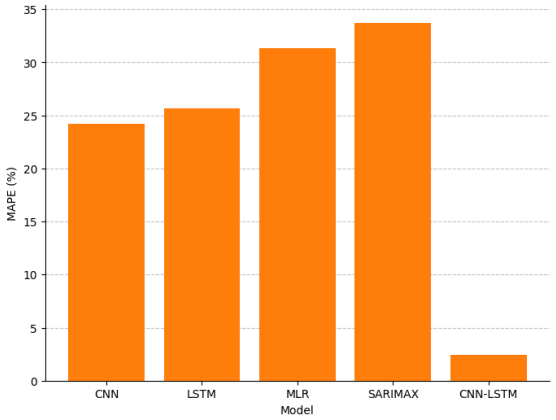

Figure 4: MAPE (%) comparison across different models.

Figure 5 presents the actual and forecasted PM2.5 concentrations for the test dataset using the CNN-LSTM model. The findings indicate that the forecast values are in close agreement with the actual values, showing the high accuracy and effectiveness of the model in forecasting PM2.5 concentrations for satellite imagery.

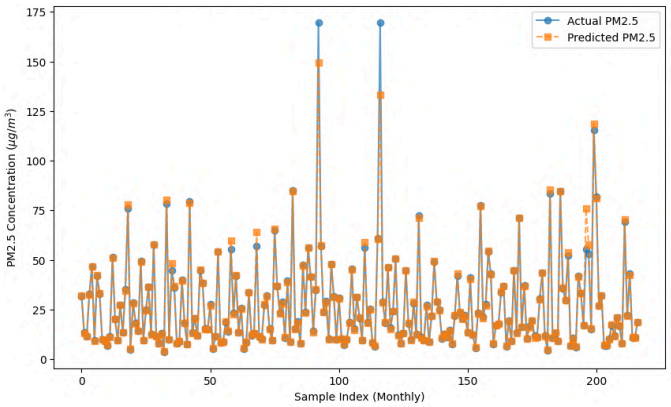

Figure 5: Comparison of actual values and forecasted values of PM2.5 for the CNN-LSTM model.

IV. Conclusion

The use of a hybrid deep learning model for estimate PM2.5 concentrations from satellite imagery has extended significant attention from researchers in recent years. This is due to the current limitations in air quality monitoring, such as high costs, inadequate coverage of monitoring devices, or even geographical limitations that may result in inaccurate readings from existing air quality monitoring equipment. Therefore, this research focuses on utilizing publicly available satellite imagery, which can be accessed at any time and covers the required study area comprehensively, to forecast PM2.5 concentrations. A hybrid deep learning model combining CNN and LSTM is applied to real satellite image data of environmental variables influencing PM2.5 in Northern Thailand. The model offered is in comparison with CNN, LSTM, MLR, and SARIMAX models. The evaluation results demonstrate that the CNN-LSTM hybrid model achieves superior performance compared to the other models. Future research may further enhance the model and evaluate the impact of additional environmental variables to improve forecasting accuracy.

Acknowledgment
This research was supported by the Government of Canada, Canada-ASEAN Scholarships and Educational Exchanges for Development (SEED 2023-2024).

References

1. Ahmed S, Khan MA, Rehman S. Estimation of ground PM2.5 concentrations in Pakistan using convolutional neural network and multi-pollutant satellite images. Remote Sens 2022;14(7):1735.
2. Kristiani E, Lin H, Lin J-R, Chuang Y-H, Huang C-Y, Yang C-T. Short-term prediction of PM2.5 using LSTM deep learning methods. Sustainability 2022;14(4):2068.
3. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. Advances in Neural Information Processing Systems, 2, 1-4.
4. Li, T., Hua, M., & Wu, X. (2020). A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). IEEE Access, 8, 26933-26940.
5. Liu, Y., Zhang, Z., Liu, X., & Liu, J. (2022). Hybrid deep learning model for PM2.5 prediction using satellite images and ground-level observations. Atmospheric Environment, 271, 118999.
6. NASA. (2023). MODIS 061 MCD19A2 Granules: Aerosol Optical Depth (AOD) data [Dataset]. NASA Goddard Space Flight Center. https://developers.google.com /earth-engine/datasets/catalog/MODIS_061_MCD19 A2_GRANULES
7. NASA. (2023). Total Ozone Mapping Spectrometer (TOMS) merged ozone data [Dataset]. NASA Goddard Space Flight Center. https://developers.google.com/earth-engine/datasets/catalog/TOMS_MERGED
8. Pollution Control Department. (n.d.). Air quality monitoring system. Air4Thai. Retrieved August 26, 2024, from http://air4thai.pcd.go.th/webV3/#/Home
9. The World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization. https://www.who.int/publications/i/item/9789240034228
10. University of Idaho. (2023). TerraClimate: Monthly climate and climatic water balance for global terrestrial surfaces [Dataset]. https://developers.google.com/ earth-engine/datasets/catalog/IDAHO_EPSCOR_TER RACLIMATE
11. UNEP. (2022). Applications of Remote Sensing for Air Pollution Monitoring in Thailand: An Early Warning for Public Health. Springer. https://link.springer.com /chapter/10.1007/978-981-19-8765-6_1
12. Zhang, X., et al. (2021). Spatio-temporal PM2.5 concentration prediction using a hybrid CNN-LSTM model with satellite data. Remote Sensing of Environment

Author's biography

C. Potavijit (Chutinun Potavijit) obtained her Bachelor's degree at Department of Mathematics, Faculty of Science, Chiang Mai University. Currently, she is a master's student in Applied Statistics and Data Analytics at Chiang Mai University. Her current research interests are hybrid deep learning models.

P. Pattarapanitchai (Parichart Pattarapanitchai) received her Bachelor's and Master's degrees in Statistics from Silpakorn University. She then obtained her Ph.D. in Statistics from Thammasat University. Currently, she is a lecturer in Statistics at Chiang Mai University. Her recent research focuses on statistical modeling in medical and forensic sciences.

C. Nontapa (Chalermrat Nontapa) is a Lecturer at the Department of Statistics, Faculty of Science, Chiang Mai University. His research focuses on Time Series, Machine Learning and Optimization. This forecast technique will be applied for hybrid model in time series data.