Data Science and Engineering (DSE) Record, Volume 6, issue 1.

ESTIMATION OF PM2.5 CONCENTRATIONS USING SATELLITE IMAGERY WITH MACHINE LEARNING TECHNIQUES

Piriya Boonchot[1]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
Piriya_b@cmu.ac.th

Abstract. Air pollution caused by fine particulate matter ($PM_{2.5}$) in Northeastern Thailand is a significant environmental concern. This study aims to identify the relationship between satellite-derived variables and $PM_{2.5}$ concentrations and to establish an effective machine learning model for $PM_{2.5}$ estimation. Sentinel-5P satellite data, comprising atmospheric variables including Carbon Monoxide (CO), Formaldehyde (HCHO), Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), Methane (CH_4), and Aerosol Index (AI), were analyzed alongside ground-based $PM_{2.5}$ measurements from 2018 to 2023. Based on Pearson correlation analysis of the atmospheric variables, it was found that Carbon Monoxide ($r = 0.72$) and Nitrogen Dioxide ($r = 0.51$) exhibited the strongest linear relationships with $PM_{2.5}$ levels. Based on statistical significance and regional source characteristics, five key variables (CO, NO_2, HCHO, AI, and O_3) were selected as input features to establish an effective machine learning model for $PM_{2.5}$ estimation. Several predictive algorithms were developed and evaluated, including Decision Tree Regression (DTR), Support Vector Regression (SVR), Polynomial Regression (PR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN). The results demonstrated that the CNN model achieved the superior performance, with the lowest Mean Absolute Error (MAE) of 7.87 $\mu g/m^3$ and the highest Coefficient of Determination (R^2) of 0.63. Although the model exhibited limitations in estimating peak concentrations during extreme haze episodes due to signal saturation, it demonstrated capability in monitoring seasonal trends and regional distribution. These findings highlight the efficiency of Deep Learning models and remote sensing data as valuable supporting tools for air quality monitoring in regions with limited ground-based observations.

Keywords: $PM_{2.5}$, Remote Sensing, Machine Learning, Deep Learning

1 Introduction

Air pollution, particularly fine particulate matter ($PM_{2.5}$), is a critical environmental issue affecting human health globally, including Thailand, where it is linked to serious health problems such as lung cancer and respiratory diseases (Ngamkaiwan, 2023). Due to their microscopic size of 2.5 micrometers or smaller, these particles can penetrate deep into the respiratory system, reaching the lungs and bloodstream (Fongsodsri et al., 2021). Northeastern Thailand, comprising 20 provinces, has faced intensifying $PM_{2.5}$ pollution, with a continuous increase in affected areas and pollution concentration levels (Kumharn et al., 2024). However, the existing ground-based air quality monitoring stations are currently insufficient to monitor the situation comprehensively. Specifically, there are only 17 monitoring stations distributed across the 20 provinces of Northeastern Thailand, which is considered inadequate coverage for the entire area. This results in data gaps and an inability to report a complete overview of air quality for the whole region. To address these limitations, satellite remote sensing technology has been applied in conjunction with Machine Learning to estimate $PM_{2.5}$ concentrations. This approach serves as an alternative tool capable of filling monitoring gaps, enabling continuous air quality monitoring with broader geographical coverage compared to ground-based stations. Generally, satellite sensors measure Aerosol Optical Depth (AOD), which correlates with $PM_{2.5}$ concentrations, and utilize algorithms to convert these values into ground-level $PM_{2.5}$ estimates (Lin et al., 2020; Yin et al., 2022). However, in the context of Northeastern Thailand, the haze problem is closely associated with seasonal factors and biomass burning, particularly during the dry season. This activity is a significant source that emits not only particulate matter but also precursor gases such as Carbon Monoxide (CO) and Nitrogen Dioxide (NO_2) (Suriyawong et al., 2023). Therefore, monitoring these gases via the Sentinel-5P satellite offers an effective approach that can help identify local pollution sources better than relying solely on general AOD values (Ahmed et al., 2022). While satellite data offers immense benefits, the relationship between satellite-derived gas data and ground-level $PM_{2.5}$ has complex characteristics. Deep Learning techniques have been proven effective in estimating and managing such data (Chen et al., 2023). Consequently, this research aims to develop and compare the performance of Machine Learning and Deep Learning models to identify the most suitable method for estimating $PM_{2.5}$ concentrations in the region.

2 Data and Methods

2.1 Study Area

The study focuses on the Northeastern region of Thailand (approx. 155,400 km^2), a highland basin characterized by a tropical savanna climate. As illustrated in Figure 1, the topography ranges from 140 to 200 meters above sea level. The region experiences distinct seasonal variations: a hot dry season (December–April) and a rainy season (May–October). Land use is predominantly agricultural, including rice, sugarcane, and maize cultivation, which significantly contributes to biomass burning activities.

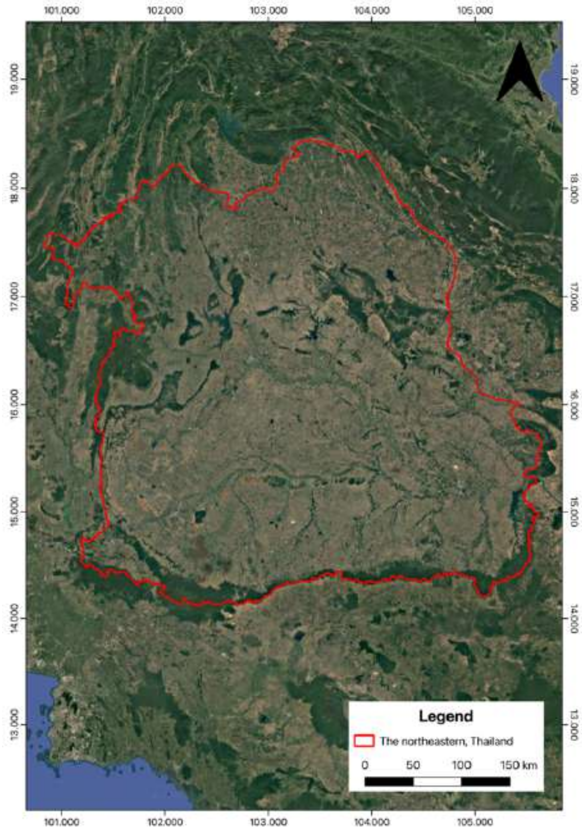

Figure 1: The study area of northeastern region, Thailand.

2.2 Data Collection

The dataset employed in this study spans the period from 2018 to 2023, integrating satellite-derived atmospheric parameters with ground-based air quality measurements to ensure temporal consistency. Satellite imagery was acquired via the Google Earth Engine (GEE) platform, specifically leveraging the Sentinel-5P TROPOMI instrument to extract key atmospheric variables, including UV Aerosol Index (AI), Carbon Monoxide (CO), Formaldehyde (HCHO), Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), and Methane (CH_4). Complementing this, ground-based air quality data comprising concentrations of $PM_{2.5}$, CO, O_3, NO_2, and SO_2 were obtained from the Pollution Control Department (PCD), collected from 17 monitoring stations distributed throughout the Northeastern region of Thailand.

2.3 Data Preprocessing

To prepare the dataset for model development, a Geographic Information System (GIS) approach was employed to extract pixel values from satellite imagery corresponding to the specific coordinates of ground monitoring stations. To ensure temporal consistency, a matching process was applied wherein ground-based $PM_{2.5}$ concentrations recorded between 13:00 and 14:00 were averaged to align with the Sentinel-5P overpass time of approximately 13:30. The preprocessing workflow proceeded with data cleaning, where invalid pixel values (≤ 0) indicative of cloud cover or water bodies were removed. Missing values were subsequently addressed using the KNN Imputer technique, selected for its ability to preserve local data structures and feature correlations. Following imputation, Min-Max Scaling was applied to normalize all features into a standardized range, ensuring stable gradient descent and faster convergence. Finally, the dataset was temporally partitioned, with data from 2018 to 2022 allocated for training to capture long-term patterns, while data from 2023 was reserved as an independent testing set for performance evaluation.

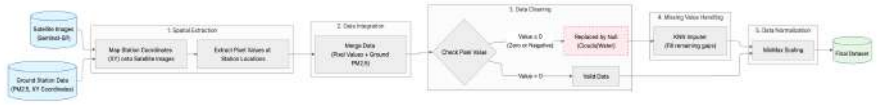
Satellite Images (Sentinel-5P)
Ground Station Data (PM2.5, XY Coordinates)
1. Spatial Extraction
Map Station Coordinates (XY) onto Satellite Images
Extract Pixel Values at Station Locations
2. Data Integration
Merge Data (Pixel Values + Ground PM2.5)
3. Data Cleaning
Check Pixel Value
Value ≤ 0 (Zero or Negative)
Replaced by Null (Clouds/Water)
Value > 0
Valid Data
4. Missing Value Handling
KNN Imputer (Fill remaining gaps)
5. Data Normalization
MinMax Scaling
Final Dataset

Figure 2: Data preprocessing workflow.

2.4 Statistical Analysis

To characterize the dependency structure between satellite-derived variables and ground-level $PM_{2.5}$, a two-stage statistical analysis was conducted prior to model development. First, correlation analysis was performed using Pearson, Spearman Rank, and Distance correlation metrics to comprehensively quantify linear, monotonic, and complex non-linear relationships, respectively. Subsequently, single-variable regression analysis was employed to evaluate the explanatory power of individual predictors. Various mathematical functions, including linear, polynomial, logarithmic, and exponential forms, were fitted to the data to determine the coefficient of determination (R^2). This preliminary assessment served to verify the complexity of the relationships and justify the necessity for advanced deep learning architectures over simple mathematical models.

Table 1: Types of statistical analyses.

Analysis Type	Method / Function
Correlation Analysis	Pearson Correlation
	Spearman Rank Correlation
	Distance Correlation
Regression Analysis	Linear Regression

Logarithmic Regression
Exponential Regression
Polynomial Regression (Degree 2)
Power Regression
Cubic Regression

2.5 Machine Learning Models

To establish a performance benchmark for $PM_{2.5}$ estimation, three traditional machine learning models Decision Tree Regression (DTR), Support Vector Regression (SVR), and Polynomial Regression (PR) were employed. Hyperparameter optimization was conducted using GridSearchCV in conjunction with 5-fold cross-validation to maximize predictive accuracy and mitigate overfitting. This process systematically evaluated the average error across various parameter configurations to identify the optimal settings for each algorithm, as detailed in Table 2.

Table 2: Hyperparameter search space for machine learning models.

Model	Hyperparameter	Description	Search Space / Values
Decision Tree	max_depth	Maximum depth of the tree	3, 5, 10, None
	min_samples_split	Minimum samples required to split a node	2, 5, 10
	min_samples_leaf	Minimum samples required at a leaf node	1, 2, 4
	max_features	Number of features to consider when looking for the best split	None, sqrt, log2
SVR	kernel	Specifies the kernel type to be used in the algorithm	rbf, poly, sigmoid
	C	Regularization parameter	0.1, 1, 10, 100
	gamma	Kernel coefficient	1, 0.1, 0.01, 0.001
Polynomial	fit_intercept	Whether to calculate the intercept for this model	True, False
	(Degree)	Polynomial degree (Fixed)	Fixed at Degree = 2

2.6 Deep Learning Models

Multilayer Perceptron (MLP): Implemented using Scikit-learn, the MLP architecture was optimized via exhaustive GridSearchCV with 5-fold cross-validation to minimize mean squared error. The search space covered hidden layer structures, activation functions, and solvers, as detailed in Table 3.

Table 3: Hyperparameter search space for MLP model optimization.

Hyperparameter	**Description**	**Search Space / Values Tested**
Hidden Layer Sizes	Architecture of hidden layers (nodes)	(6), (10), (50), (100), (50, 25), (100, 50)
Activation Function	Non-linear activation function	ReLU, Tanh, Logistic (Sigmoid)
Solver	Optimization algorithm	Adam, SGD
Alpha	L2 regularization penalty term	0.0001, 0.001, 0.01
Batch Size	Number of samples per gradient update	16, 32, 64, 128
Max Iterations	Maximum number of epochs	100, 500, 1000

One-Dimensional Convolutional Neural Network (1D-CNN): Developed using TensorFlow Keras, the model features a bottleneck architecture comprising three consecutive 1D-convolutional layers (filters: 128, 64, 32; kernel size: 2) to capture local temporal dependencies. To ensure stability and prevent overfitting, each block incorporates Batch Normalization, Dropout (0.2), and L2 Regularization ($\lambda = 0.01$). The extracted features are flattened and processed through two dense layers (64 and 32 units; Dropout 0.3) before a final linear output layer. Training utilized the Adam optimizer (lr = 0.001) and Huber Loss to mitigate the impact of outliers, employing EarlyStopping and ReduceLROnPlateau strategies. Unlike the MLP's exhaustive search, the CNN configuration was determined through empirical preliminary experiments due to computational constraints (Table 4).

Table 4: Architecture and hyperparameters of the proposed 1D-CNN model.

Layer / Stage	Configuration	Activation	Regularization / Technique
Input	Shape: (Time Steps, Features)	-	-
Conv1D (1)	Filters: 128, Kernel: 2	ReLU	BN, Dropout (0.2), L2 (0.01)
Conv1D (2)	Filters: 64, Kernel: 2	ReLU	BN, Dropout (0.2), L2 (0.01)
Conv1D (3)	Filters: 32, Kernel: 2	ReLU	BN, Dropout (0.2), L2 (0.01)
Flatten	-	-	-

Dense (1)	Units: 64	ReLU	Dropout (0.3), L2 (0.01)
Dense (2)	Units: 32	ReLU	Dropout (0.3), L2 (0.01)
Output	Units: 1	Linear	-
Training	Batch: 16, Epochs: 100	Optimizer: Adam (lr=0.001)	Loss: Huber (δ=1.0), Callbacks: EarlyStopping & ReduceLR

2.7 Model Evaluation

To comprehensively assess predictive performance and generalization capability, three standard statistical metrics were employed: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). To ensure the reliability of the results and mitigate bias associated with data partitioning, 5-fold cross-validation was applied to the machine learning and MLP models. The final performance assessment was based on the average metrics across all folds, with R^2 serving as the primary criterion for hyperparameter optimization and model selection.

3 Results

3.1 Correlation Analysis and Feature Selection Dataset

The dependency between satellite-derived variables and ground-level $PM_{2.5}$ was evaluated using Pearson, Spearman, and Distance correlations, with results summarized in Table 5. All relationships were statistically significant ($p < 0.05$). Carbon Monoxide (CO) exhibited the strongest positive correlation (Pearson $r = 0.72$), followed by Nitrogen Dioxide (NO_2) ($r = 0.51$). These strong associations align with the source characteristics of Northeastern Thailand, where biomass burning and open fires are predominant, co-emitting $PM_{2.5}$ along with CO and NO_2 products of incomplete combustion. Formaldehyde (HCHO) showed a moderate positive correlation ($r = 0.38$), reflecting its role in secondary aerosol formation, while Ozone (O_3) displayed a negative correlation ($r = -0.2$), likely attributable to atmospheric titration processes during high particulate events. Conversely, Methane (CH_4) and Sulfur Dioxide (SO_2) exhibited weak correlations ($r < 0.2$). This dissociation is attributed to mismatched emission sources, as CH_4 is primarily driven by agriculture and SO_2 by industrial activities, rather than the combustion sources driving $PM_{2.5}$ pollution in this region. Based on statistical significance and physical relevance, CO, NO_2, HCHO, and O_3 were selected as model inputs. Additionally, the UV Aerosol Index (AI) was retained despite a lower correlation ($r = 0.24$) due to its critical physical capability in detecting absorbing aerosols (smoke and dust) essential for identifying transboundary haze. Consequently, CH_4 and SO_2 were excluded to optimize model performance.

Table 5: Correlation coefficients between $PM_{2.5}$ concentration and satellite-derived variables.

Method	Correlation Coefficient (r)						
	AI	CH_4	CO	NO_2	O_3	HCHO	SO_2
Pearson	0.2353	-0.0705	0.7178	0.5119	-0.1978	0.3836	0.0585
Spearman	0.0653	-0.0825	0.6559	0.5161	-0.2486	0.3401	0.1051
Distance	0.1423	0.1004	0.6865	0.4969	0.2477	0.3544	0.0962

3.2 Performance of Single-Variable Regression

To evaluate the explanatory power of individual satellite predictors, single-variable regression analysis was conducted using Linear, Logarithmic, and Polynomial (degree 2) functions. As summarized in Table 6, traditional regression models yielded limited predictive accuracy. Even with non-linear polynomial fitting, Carbon Monoxide (CO), the most strongly correlated variable, achieved only a moderate R^2 of approximately 0.52, while other variables exhibited significantly lower values. These results indicate that simple mathematical functions are insufficient to capture the complex dynamics of $PM_{2.5}$ formation. The limited variance explained by these conventional statistical methods underscores the necessity for advanced Machine Learning and Deep Learning approaches capable of modeling high-dimensional and non-linear interactions effectively.

Table 6: Coefficient of determination between $PM_{2.5}$ concentration and satellite-derived variables.

Method	Coefficient of Determination (R^2)						
	AI	CH_4	CO	NO_2	O_3	HCHO	SO_2
Linear	0.0554	0.005	0.5152	0.262	0.0391	0.1472	0.0034
Exponential	0.0169	0.0049	0.464	0.198	0.0392	0.086	0.0042
Logarithmic	0.0154	0.003	0.4386	0.2593	0.0569	0.1232	0.0086
Polynomial	0.1013	0.0052	0.5247	0.279	0.055	0.1494	0.0069
Power	0.0053	0.0029	0.4357	0.227	0.0577	0.0799	0.0077
Cubic	0.0554	0.005	0.5152	0.262	0.0391	0.1472	0.0034

3.3 Comparative Performance of Deep Learning and Machine Learning Models

The performance evaluation of five distinct regression models Decision Tree Regression (DTR), Support Vector Regression (SVR), Polynomial Regression (PR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN) is summarized in Table 7. The experimental results indicate that the 1D-CNN architecture yielded the superior performance metrics, achieving the lowest Mean Absolute Error (MAE) of 7.87 and the highest Coefficient of Determination (R^2) of 0.63. This suggests that the hierarchical feature extraction capabilities of the CNN are most effective in capturing the complex, local patterns within the multivariate satellite data. Closely following the CNN, the MLP ranked second (R^2 = 0.62, MAE = 8.14), while SVR demonstrated competitive performance (R^2 = 0.61, MAE = 8.15). The proximity of SVR's results to the deep learning models implies that the relationship between satellite parameters and $PM_{2.5}$ is highly non-linear, which the SVR's kernel trick successfully modeled. However, the deep learning architectures retained a marginal advantage in handling high-dimensional complexity. In contrast, PR (R^2 = 0.59) and DTR (R^2 = 0.56) exhibited higher error rates, highlighting their limitations in generalizing continuous variable predictions for this specific domain.

Table 7: The Performance of Deep Learning and Machine Learning Models.

Model	MAE	MSE	R^2
DTR	8.74	181.26	0.56
SVR	8.15	160.34	0.61
PR	8.43	167.04	0.59
MLP	8.14	156.41	0.62
CNN	**7.87**	**152.63**	**0.63**

To further assess model robustness, a 5-fold cross-validation was conducted for the ML and MLP models (Table 8). The results reveal a significant distinction in stability. The MLP demonstrated the most robust framework, achieving the highest average R^2 of 0.55 with the lowest Standard Deviation (SD) of 0.12. Conversely, traditional models (SVR, PR, DTR) exhibited considerably higher volatility (SD ≈ 0.25–0.26), indicating sensitivity to data fluctuations. Note that while the CNN was optimized via extensive randomized experiments rather than k-fold cross-validation due to computational constraints, its superior accuracy on the independent test set, combined with the proven stability of the neural network family (MLP), confirms deep learning as the optimal approach for this study.

Table 8: 5-fold cross-validation.

Model	Average R^2	Standard Deviation R^2
DTR	0.46	0.26
SVR	0.53	0.26
PR	0.53	0.25
MLP	0.55	0.12

3.4 Application of the 1D-CNN Model for $PM_{2.5}$ Estimation

3.4.1 Overall Estimation Performance (2018–2023)

The scatter plot comparing actual versus estimated $PM_{2.5}$ concentrations across the entire study period reveals a positive correlation with an R^2 of 0.63 (Figure 3). The model demonstrates high reliability in estimating concentrations within the 'Very Good' to 'Unhealthy' range (0–75 µg/m³), where data density is highest. However, a systematic underestimation is observed during extreme pollution episodes (>75 µg/m³), a limitation likely attributed to satellite signal saturation. Despite this, the overall alignment confirms the model's robustness in monitoring long-term trends and seasonal variability.

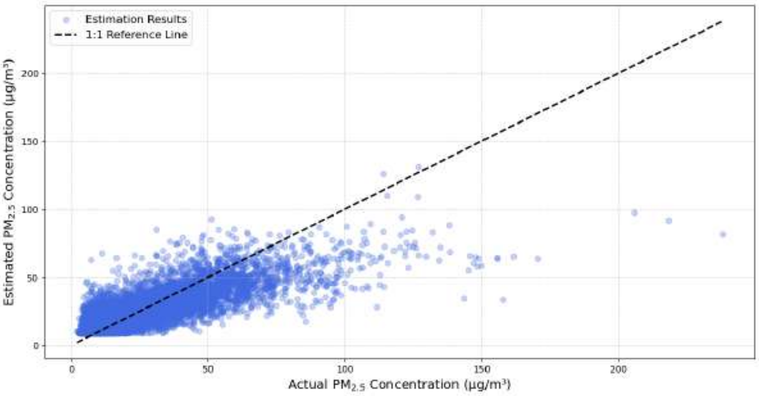

Figure 3: Scatter Plot Actual vs Estimation of $PM_{2.5}$ Concentrations (2018-2023).

3.4.2 Temporal and Seasonal Analysis (2023)

The time-series analysis for 2023 (Figure 4) confirms the model's ability to track seasonal dynamics, effectively capturing the haze period (December–April) and the washout season (June–September). Rainy Season (Jun–Oct): The model achieved peak performance with the lowest MAE of 4.22 µg/m³, accurately reflecting background pollution levels (Mean Est: 13.88 vs. Actual: 12.93 µg/m³). Winter Season (Nov–Feb):

Performance remained robust (Mean Est: 25.39 vs. Actual: 26.29 µg/m³; MAE = 9.19 µg/m³). Summer Season (Mar–May): This period presented the greatest challenge due to high volatility from biomass burning. The model tended to underestimate peaks, resulting in a higher MAE of 12.08 µg/m³. notably, the extreme peak on April 7 (238.04 µg/m³) was estimated at 81.75 µg/m³, highlighting a limitation in capturing magnitude during severe events while still successfully identifying the timing of pollution onset.

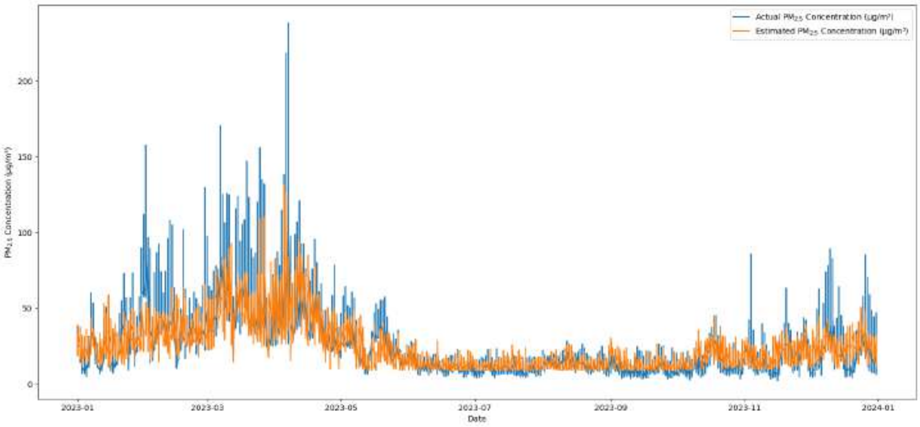

Figure 4: Estimated PM2.5 Concentrations of 2023.

3.4.3 Spatial Distribution Analysis

Spatial distribution maps generated for January–June 2023 (Figure 5) illustrate the regional progression of PM2.5. The model correctly identifies spatial patterns, showing low concentrations in January–February, followed by the emergence of hotspots along the Laos border in March due to transboundary haze. The pollution footprint peaks in April, covering the upper Northeastern region, before dissipating in May–June due to seasonal rains. Although the model captures the spatiotemporal evolution of the haze, it tends to conservative estimations, frequently categorizing concentrations within 'Very Good' to 'Moderate' levels (0–37.5 µg/m³), suggesting a need for further calibration to fully represent the magnitude of critical haze episodes.

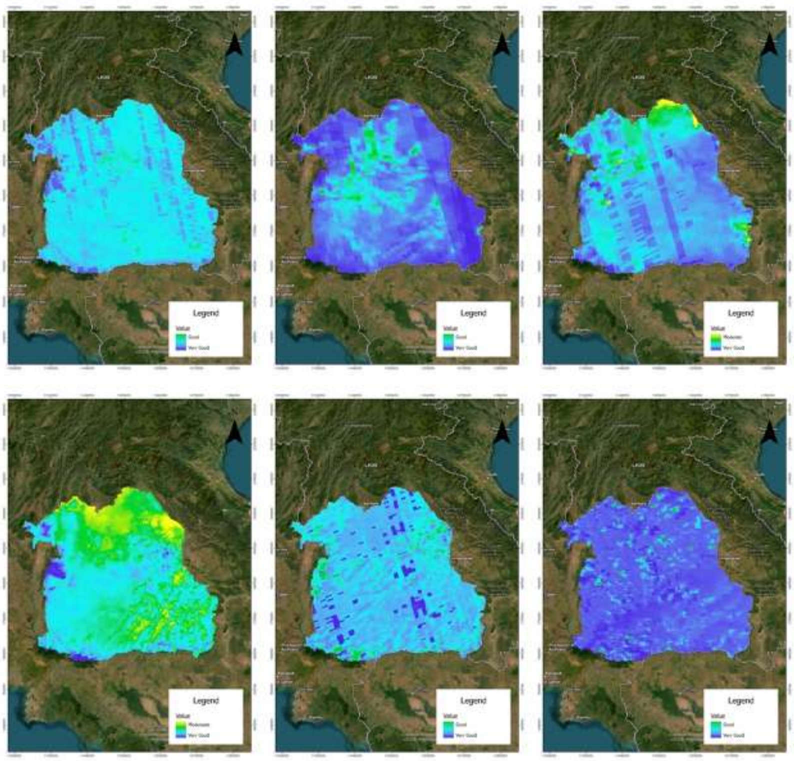

Figure 5: Spatial Distribution Maps of $PM_{2.5}$

4 Conculsion and Discussion

4.1 Conclusion

This study establishes a robust framework for estimating ground-level $PM_{2.5}$ in Northeastern Thailand using satellite-derived data. The investigation into variable relationships identifies Carbon Monoxide (CO) and Nitrogen Dioxide (NO_2) as the most critical indicators, reflecting the dominance of biomass burning emissions, whereas the inability of simple mathematical models to capture these dependencies justifies the necessity for advanced high-dimensional approaches. Consequently, the comparative analysis demonstrates that Deep Learning, specifically the 1D-Convolutional Neural Network (1D-CNN), is superior to traditional machine learning methods. With the highest accuracy ($R^2 = 0.63$) and lowest error rates, the 1D-CNN effectively overcomes feature engineering limitations by automatically extracting hierarchical representations from multivariate data. In practical application, the model proves highly effective for regional trend monitoring, accurately capturing seasonal transitions and background pollution levels, particularly during the rainy season where error rates are minimal.

However, a systematic underestimation was observed during extreme haze episodes, attributed to signal saturation, which limits the model's utility for precise peak detection. Despite this quantitative limitation regarding extreme magnitudes, the model successfully delineates spatial distribution patterns and transboundary pollution hotspots. Ultimately, this research confirms that while further calibration for extreme events is required, the proposed satellite-based deep learning approach serves as a valuable and cost-effective tool for air quality management in regions lacking extensive ground-based monitoring infrastructure.

4.2 Discussion

The Analysis of Relationships between Satellite Data and $PM_{2.5}$ The analysis of the relationship between $PM_{2.5}$ concentration and satellite-derived variables indicates that the correlation type provides critical insights into emission sources. Our study utilized Pearson correlation analysis, revealing that Carbon Monoxide (CO) and Nitrogen Dioxide (NO_2) exhibit the strongest positive linear relationships with $PM_{2.5}$ ($r = 0.7178$ and 0.5119, respectively). This methodological approach aligns with Li et al. (2017), who also employed Pearson correlation to analyze air pollution in China and found that CO and NO_2 exhibited significant correlations with particulate matter. This similarity underscores that CO and NO_2, as primary combustion byproducts, maintain a robust linear dependency with $PM_{2.5}$ across different geographical regions. However, a distinct divergence is observed regarding Sulfur Dioxide (SO_2). In our study, SO_2 showed a negligible linear correlation (Pearson $r = 0.0585$). This stands in sharp contrast to the findings of Li et al. (2017), who reported that all pollutants, including SO_2, exhibited significant correlations with one another in Chinese cities due to heavy industrialization. This disparity highlights the specific regional context of Northeastern Thailand. As described by Suriyawong et al. (2023), air pollution in this region is driven primarily by the open burning of agricultural residues (rice and sugarcane) rather than sulfur-rich industrial activities. Thus, the lack of SO_2 correlation in our study serves as a validation that the $PM_{2.5}$ sources in the study area are predominantly agricultural. Furthermore, Ozone (O_3) exhibits a statistically significant inverse correlation (Pearson $r = -0.1978$) in our study. This negative relationship is consistent with the general trends observed by Li et al. (2017), who noted that O_3 often displays a distinct temporal variation pattern compared to particulate matter and primary pollutants (CO, NO_2). This inverse association suggests the presence of atmospheric chemical interactions, such as the "titration effect," where ozone is consumed by NO in chemical reactions during periods of high pollution concentration, resulting in the negative correlation observed in our annual dataset. Regarding Methane (CH_4), our analysis using satellite-derived numerical values indicates a negligible correlation (Pearson $r = -0.0705$). This contrasts with Ahmed et al. (2022), who successfully employed satellite-derived CH_4 images in a Deep Learning model. The disparity suggests that while CNNs can extract complex features from CH_4 imagery, the direct numerical values used in this correlation analysis do not exhibit a strong linear relationship with ground-level $PM_{2.5}$.

Model Performance and Input Modality In terms of model development, the Convolutional Neural Network (CNN) emerged as the most suitable approach for estimating $PM_{2.5}$ concentrations, achieving the lowest MAE (7.87) and the highest Coefficient of Determination (R^2 of 0.63). This performance is attributed to the CNN's capacity to learn hierarchical features and capture complex dependencies within the data. Notably, the Support Vector Regression (SVR) demonstrated competitive performance with an R^2 of 0.61, closely trailing the 1D-CNN. This proximity suggests that the relationship between satellite-derived variables and ground-level $PM_{2.5}$ is highly non-linear but can be effectively approximated by kernel-based methods. The SVR, utilizing the Radial Basis Function (RBF) kernel, successfully mapped these non-linear interactions in a high-dimensional feature space without requiring deep architectural layers. However, the 1D-CNN achieved slightly superior results ($R^2 = 0.63$) because its convolutional layers are capable of extracting latent local patterns and hierarchical feature interactions within the input vector that a fixed kernel function might overlook. However, our model's performance remains lower than the P-CNN model reported by Ahmed et al. (2022). This discrepancy is primarily associated with the fundamental difference in input data. Ahmed et al. (2022) utilized multi-pollutant satellite images as inputs, allowing their model to leverage the capabilities of CNNs in extracting spatial textures and patterns. In contrast, our study employed numerical values extracted from satellite products. While efficient, this tabular format lacks the spatial context of raw imagery, thereby limiting the CNN’s ability to fully exploit its architecture for spatial feature extraction.

Impact of Meteorological Data and Data Precision Furthermore, the performance gap is explained by the complexity of input features when compared to Sun et al. (2019). Their PM25-DNN model achieved a higher R^2 of 0.84 by relying on a comprehensive dataset incorporating Aerosol Optical Depth (AOD) and seven meteorological variables. Conversely, our study operated under constrained data availability, utilizing only satellite-derived pollution parameters without meteorological variables. While Sun et al. (2019) emphasized that meteorological factors are significant drivers for $PM_{2.5}$ estimation, our results demonstrate that a CNN can still adequately capture air quality trends using solely satellite-derived pollutants ($R^2 = 0.63$), offering a viable solution for regions lacking weather stations. Finally, comparisons with Chen et al. (2023) and Ding et al. (2021) highlight the impact of data source precision. Both studies achieved higher predictive accuracies (R^2 of 0.88 and 0.96, respectively) by utilizing ground-based measurements for pollutants and meteorological factors. In contrast, our study relied on satellite-derived numerical values, which represent data from the entire atmospheric column rather than direct surface concentrations. This fundamental difference in data nature, combined with the absence of meteorological variables, explains the performance disparity. Nevertheless, the fact that our model achieved satisfactory performance using solely satellite-derived data suggests that this approach provides a feasible baseline for prediction, particularly valuable for resource-limited settings where ground-based measurements are sparse or unavailable.

5 Challenges and Limitations

1. Incomplete data, some air quality monitor stations and satellite images may have missing or erroneous data, along with discontinuous data collection, potentially affecting model performance.
2. Limitations of satellite imagery, satellite data has constraints in terms of spatial resolution, which may make it inappropriate for studies focusing on small area such as districts or villages.
3. Underestimation of Extreme Events, the model demonstrated a systematic underestimation of $PM_{2.5}$ concentrations during high-pollution episodes (e.g., biomass burning peaks). This suggests a limitation in the model's dynamic range or signal saturation issues when dealing with extreme values beyond the distribution of the training set.

6 Recommendations

1. Model improvement, further studies should explore advanced Deep learning model such as Hybrid Model or Transformer-based models to enhance prediction accuracy.
2. Incorporation of atmosphere variables, further research should consider integrating meteorological factors such as humidity, temperature, and wind speed alongside satellite data to improve model learning and performance.
3. Combined with additional satellites, further research should consider integrating images from other satellites to reduce missing values and increase the amount of data.

References

1. Ahmed, M., Xiao, Z., & Shen, Y. (2022). Estimation of Ground PM2.5 Concentrations in Pakistan Using Convolutional Neural Network and Multi-Pollutant Satellite Images. *Remote Sensing*, *14*(7), 1735. https://www.mdpi.com/2072-4292/14/7/1735
2. Chen, M.-H., Chen, Y.-C., Chou, T.-Y., & Ning, F.-S. (2023). PM2.5 Concentration Prediction Model: A CNN–RF Ensemble Framework. *International Journal of Environmental Research and Public Health*, *20*(5).
3. Ding, C., Wang, G., Zhang, X., Liu, Q., & Liu, X. (2021). A hybrid CNN-LSTM model for predicting PM2.5 in Beijing based on spatiotemporal correlation. *Environmental and Ecological Statistics*, *28*(3), 503-522. https://doi.org/10.1007/s10651-021-00501-8
4. Fongsodsri, K., Chamnanchanunt, S., Desakorn, V., Thanachartwet, V., Sahassananda, D., Rojnuckarin, P., & Umemura, T. (2021). Particulate Matter 2.5 and Hematological Disorders From Dust to Diseases: A Systematic Review of Available Evidence. *Frontiers in Medicine*, *8*.
5. Kumharn, W., Sudhibrabha, S., Hanprasert, K., Janjai, S., Masiri, I., Buntoung, S., Pattarapanitchai, S., Wattan, R., Homchampa, C., Srimaha, T., Pilahome, O., Nissawan,

W., & Jankondee, Y. (2024). Estimating hourly full-coverage PM2.5 concentrations model based on MODIS data over the northeast of Thailand. *Modeling Earth Systems and Environment*, *10*(1), 1273-1280. https://doi.org/10.1007/s40808-023-01839-7

6. Li, M., Tijian, W., Min, X., Bingliang, Z., Shu, L., Yong, H., & and Chen, P. (2017). Impacts of aerosol-radiation feedback on local air quality during a severe haze episode in Nanjing megacity, eastern China. *Tellus B: Chemical and Physical Meteorology*, *69*(1), 1339548. https://doi.org/10.1080/16000889.2017.1339548
7. Li, R., Cui, L., Li, J., Zhao, A., Fu, H., Wu, Y., Zhang, L., Kong, L., & Chen, J. (2017). Spatial and temporal variation of particulate matter and gaseous pollutants in China during 2014–2016. *Atmospheric Environment*, *161*, 235-246. https://doi.org/https://doi.org/10.1016/j.atmosenv.2017.05.008
8. Lin, C., Labzovskii, L. D., Leung Mak, H. W., Fung, J. C. H., Lau, A. K. H., Kenea, S. T., Bilal, M., Vande Hey, J. D., Lu, X., & Ma, J. (2020). Observation of PM2.5 using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring. *Atmospheric Environment*, *227*, 117410. https://doi.org/https://doi.org/10.1016/j.atmosenv.2020.117410
9. Ngamkaiwan, C. (2023). Secondary Green Crime: Bangkok’s PM2.5 Pollution and Policy Corruption. *International Journal for Crime, Justice and Social Democracy*.
10. Sun, Y., Zeng, Q., Geng, B., Lin, X., Sude, B., & Chen, L. (2019). Deep Learning Architecture for Estimating Hourly Ground-Level PM2.5 Using Satellite Remote Sensing. *IEEE Geoscience and Remote Sensing Letters*, *16*(9), 1343-1347. https://doi.org/10.1109/LGRS.2019.2900270
11. Suriyawong, P., Chuetor, S., Samae, H., Piriyakarnsakul, S., Amin, M., Furuuchi, M., Hata, M., Inerb, M., & Phairuang, W. (2023). Airborne particulate matter from biomass burning in Thailand: Recent issues, challenges, and options. *Heliyon*, *9*(3), e14261. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e14261
12. Yin, S., Li, T., Cheng, X., & Wu, J. (2022). Remote sensing estimation of surface PM2.5 concentrations using a deep learning model improved by data augmentation and a particle size constraint. *Atmospheric Environment*.