eISSN: 2976-1190

Journal of Engineering and Sciences

Vol. 3 Issue 1 May 30, 2024 / **DOI: https://doi.org/10.3126/jes2.v3i1.66239**

Journal homepage: Journal of Engineering and Sciences (nepjol.info)

Publisher: Research Management Cell, Pashchimanchal Campus, Lamachour, Pokhara

Applying machine learning algorithms to estimate PM 2.5 using satellite data and metrological data

Ishwor Thapa[1]*, Bidur Devkota[1]

Gandaki College of Engineering and Science, Pokhara University, Pokhara, Nepal

(Manuscript Received 09/03/2024; Revised 25/05/2024; Accepted 26/05/2024)

Abstract

Air pollution, particularly fine Particulate Matter (PM 2.5), poses significant health risks and environmental challenges worldwide. Therefore, it is essential to monitor air pollution to act on it. In this study, PM 2.5 was estimated using meteorological data and Sentinel-5P air pollution data using machine learning algorithms. The Sentinel-5P data are and the meteorological data utilized are air temperature, Relative Humidity (RH), and Wind Speed (WS). The three Air Quality Monitoring (AQM) stations in Kathmandu, Nepal, were chosen as a study area for this research. The effectiveness of several machine learning methods, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF), were evaluated. Both RF and XGBoost consistently performed better than SVM and KNN in terms of PM 2.5 estimation accuracy. RF got the highest R^2 value of 0.80 and SVM with the lowest R^2 value of 0.62 in the Sentinel-5P dataset only. The addition of meteorological data further improved the model's performance. After including metrological data in Sentinel-5P data the RF demonstrated the maximum R^2 score of 0.816 and XGBoost with R^2 score of 0.814. Hence, this study demonstrated machine learning algorithms can be used to estimate PM 2.5 by utilizing satellite and meteorological data, providing important information for air quality monitoring and management.

***Keywords*:** Google Earth Engine; Kathmandu-Nepal; PM 2.5; Sentinel-5P

1. Introduction

Air pollution is a serious environmental problem that affects human health and the environment worldwide. PM2.5 refers to fine particulate matter and has a diameter of 2.5 micrometers or smaller. PM 2.5 is a tiny particle suspended in the air that can come from a variety of sources like construction, industrial work, forest fires, dust, and so on. Exposure to PM2.5 has been linked to a range of health problems, particularly respiratory and cardiovascular issues. When people are exposed to high levels of PM2.5 for an extended period, it can cause or worsen respiratory diseases such as asthma, bronchitis, and other chronic obstructive pulmonary diseases (COPD). Additionally, it may increase the risk of heart attacks, strokes, birth defects and premature death [1].

One of the most common and accurate methods for monitoring air quality is through air quality monitoring stations. However, measurements are only available in the surrounding area of the stations [2]. Through air quality monitoring, air pollutant concentration data are obtained to determine whether the concentration levels are good, unhealthy for sensitive groups, or at emergency levels. Due to the high personnel, infrastructure, and financial demands associated with their establishment, operation, and maintenance, these stations are dispersed widely [3].

On the other hand, air quality monitoring is also conducted nowadays by utilizing geospatial and remote sensing techniques to gather air quality data over large areas. It is convenient to use the satellite remote sensing data information inference method because of the widespread remote sensing satellite coverage, large area synchronized observation that can be performed in a short amount of time, handy and quick access to the real-time global range of all kinds of natural phenomena, and efficiency of the PM2.5 measurement [4-5]. Air quality data is either directly obtained from sensors on various satellite platforms or derived from satellite images using regression analysis or dispersion models. As computer technology advances, machine learning techniques are being used extensively to predict PM2.5 concentrations. Classical machine learning methods using decision trees, random forests, support vector machines, and artificial neural networks have

*Corresponding author. Tel.: +977- 9826148325,
E-mail address: ishwor_msc03_2021@gces.edu.np

demonstrated good predictive ability for PM2.5 concentrations [6-7]. To estimate PM2.5 concentrations, machine learning methods often combine several data sources, such as meteorological, air pollution, spatio-temporal, land use, and satellite remote sensing information [8].

The level of particle air pollution in Nepal frequently ranks among the worst in the world. Nepal was ranked 177th out of 180 nations in 2016 and was ranked 162 in the 2022 Environmental Performance Index (EPI) for air pollution [9]. Air quality monitoring in Nepal is conducted through ground monitoring station data regulated by the Department of Environment under the Ministry of Forests and Environment. According to the Department of Environment of Nepal, 27 air quality monitoring stations measure the following significant parameters: PM1, PM2.5, PM10, and Total Suspended Particulates (TSP) [10]. Some of the limitations of these ground stations are limited coverage, spatial variability, and expensive and complex equipment. These few stations are not well sufficient for measuring air quality throughout the country. However, the technique of using satellite remote sensing to estimate air pollution concentrations has the benefit of being highly effective and inexpensive [11].

1.1 Related works

Based on the data source, the related works can be classified into three groups. The methods in the first group used the historical PM readings from the ground-based air quality monitoring stations for PM estimation. Several machine learning algorithms linear regression, K-neighbors, decision tree, RF, gradient boosting, CNN, and LSTM were used, to estimate the PM value in the current day or future days [12-13]. In addition to the PM measurements from the available stations, the second group uses satellite-derived data such as aerosol optical depth (AOD). Moreover, several studies have included meteorological data (temperature, humidity, wind speed, etc.) [14-15]. In the third group, instead of using the satellite-derived products, the satellite images are directly used [16]. However, only a few studies have used air pollutant concentrations (SO2, CO, NO2, and O3) for PM 2.5 computations [17].

The proposed study is one of the first in Nepal to directly estimate PM2.5 concentrations using Sentinel-5P pollution data. One of the worst affected areas in Nepal is the Kathmandu Valley, located in the mid-hills of Nepal at a latitude of 27.7◦N and a longitude of 85.3◦E. Most of the studies regarding air pollution in Kathmandu Valley are based on air pollution stations and satellite-derived products (AOD) [18]. The use of AOD and other atmospheric products has some difficulties such as computational burden and uncertainties due to aerosol model selection and cloud screening schemes. AOD products have coarse spatial resolutions which make them a good candidate for monitoring of global distribution of aerosols on large scales and wide coverage. However, the AOD products are not capable to estimate PM 2.5 concentration in smaller area. The purpose of this study is to develop low-computing and simple machine-learning models for the estimation of PM 2.5 concentration. It uses free Sentinel-5P pollution data to estimate PM in small areas, such as cities.

The proposed study offers several key contributions. Since there are fewer air quality monitoring stations in Nepal and most of the stations are not operating properly, this study will help enhance the analysis of the near-ground PM2.5 pollution situation. It will also be helpful to estimate the ground-level PM 2.5 using satellite images where there is no availability of ground-level AQM stations. Rather than using AOD products, Sentinel-5P air pollution data is directly used. We compared the effectiveness of various machine learning models on Sentinel-5P datasets, both with and without metrological data.

2. Materials and Method

2.1 Study Area

For this study, we have selected Kathmandu, Nepal, as our study area. There are seven air quality monitoring stations located around the Kathmandu Valley: in Dhulikhel, Ratnapark, Sankhapark, Bhaisepati, Pulchowk, Bhaktapur, and Kirtipur. The research locations and PM 2.5 monitoring sites in Kathmandu are shown in Figure 1.

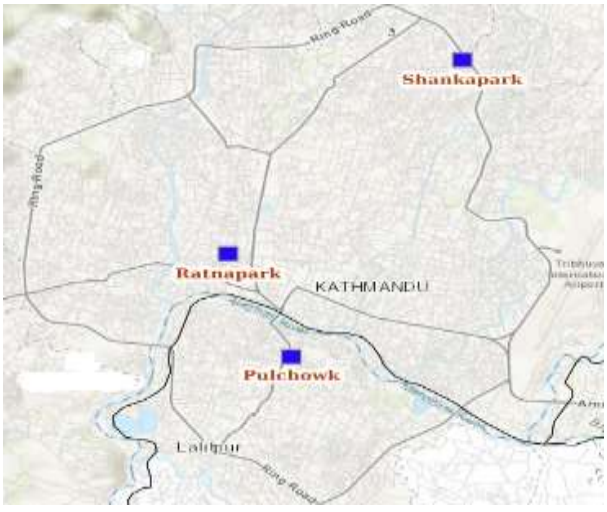

Figure 1: AQM stations used for this research

2.2 Datasets

The dataset used in this study includes air pollution data from the Sentinel-5P satellite, PM2.5 along with meteorological factors and data from ground stations. Two datasets were created based on ground stations, as shown in Table 1. Dataset 1 contains Ratnapark station data, while Dataset 2 consists of three stations (Ratnapark, Shankapark, and Pulchowk) data.

Table 1. Description of dataset 1 and dataset 2

Dataset 1	Ratnapark station	Sentinel-5P only
		Sentinel-5P and metrological data
Dataset 2	Ratnapark, Shankapark, Pulchowk station	Sentinel-5P only
		Sentinel-5P and metrological data

.2.2.1 Ground station AQM data

Daily average PM 2.5 data were collected from the Department of Environment office located at Forest Complex, Babarmahal, Kathmandu [19]. AQM Data from Ratnapark, Shankhapark, and Pulchowk stations were used in this study. Table 2 shows the summary of AQM station data and the overall total data that was used, excluding the missing value.

Table 2. Summary of AQM station data

Station	Latitude, Longitude	Time Period	Overall data accessibility
Ratnapark	27.7, 85.31	2019-1-1 to 2021-12-31	886
Shankapark	27.7328252, 85.342826	2019-2-12 to 2020-5-10	601
Pulchowk	27.682581, 85.318841	2019-1-1 to 2020-2-20	415

2.2.2 Metrological data

To check the performance and effectiveness of PM 2.5 estimation with and without metrological datasets, the minimum and maximum meteorological data, such as air temperature, RH and WS, were used. These data sets were obtained from the Department of Hydrology and Meteorology, Babarmahal, Kathmandu, Nepal, following the completion of the metrological data request payment procedure [20].

2.2.3 Satellite data

Sentinel-5P is the first mission of the Copernicus air pollution control program. Sentinel-5P can be used to detect gases such as NO2, SO2, CH4, HCHO, AI, CO and O3.

Google Earth Engine (GEE) has been used by several researchers for Air Pollutants (AP) retrieval [21-22]. GEE is a cloud-based platform developed by Google for planetary-scale for analyzing environmental data. It provides a huge collection of satellite imagery, geospatial datasets, and computational capabilities that helps to visualize, analyze, and process remote sensing data for a wide range of applications. Therefore, AP retrieval from the Sentinel-5P satellite has been done in this study using GEE.

2.3 Methodology

Data on NO_2, SO_2, CH_4, HCHO, AI, CO, and O_3 pollutant levels from Sentinel-5P air pollution were retrieved using GEE. The GEE platform was used to extract Sentinel-5P air pollution data, providing a strong toolkit for accessing and processing satellite imagery. To begin the extraction process, a region of interest (ROI) was defined. This defines the geographic area from which the data were gathered. The Sentinel-5P dataset was filtered based on the specified ROI, date range, and desired pollutants. The selected bands were then aggregated by calculating the mean value for each band across all available images within the dataset over the specific time period. Additionally, the dataset was clipped to the ROI to retain only the data relevant to the study area. Finally, the processed data was exported in CSV format. Data preprocessing was done on the gathered data from Sentinel-5P, AQM stations, and meteorological sources, as shown in Figure 2.

2.3.1 Data preprocessing

To enable temporal analysis and modeling, the date values in the date column of the gathered datasets were transformed into a standard date format that included the day, month, and year. Missing values in the datasets were handled using linear interpolation techniques to fill in the gaps and ensure continuity in the data [23]. This approach effectively addresses missing data issues while preserving the temporal and spatial integrity of the datasets.

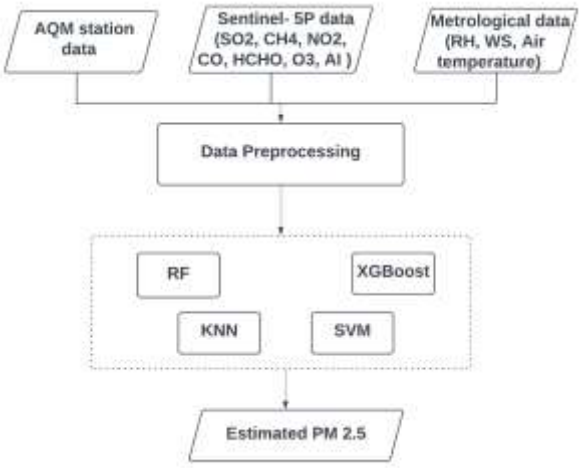

Figure 2: Methodology used for the estimations of PM 2.5

2.3.2 Algorithm selection

In this study, we have used four different algorithms, RF, SVM, XGBOOST and KNN. RF is a part of the ensemble learning techniques, which improve prediction performance by combining several models. To do regression tasks, RF builds a large number of decision trees during training and outputs the average prediction of each tree [24]. SVM works by finding the optimal hyperplane that best separates the data points into different classes or predicts continuous values. It accomplishes this by maximizing the margin between the hyperplane and the nearest data points (support vectors) [25]. XGBoost belongs to the ensemble learning category and is based on the gradient boosting framework. It combines the predictions of multiple weak models to create a stronger, more accurate model. XGBoost employs a technique called tree boosting, where decision trees are added one at a time and their predictions are combined with the predictions of previously added trees [26].

KNN is based on the principle that similar data points tend to have similar target values. In KNN, the prediction for a new data point is made based on the majority class (for classification) or the average of the values of its k nearest neighbors (for regression) in the feature space [27].

2.3.3 Hyperparameter Tuning

Hyperparameters play an important role in determining the complexity, flexibility, and generalization ability of the model. In this study, we have used grid search cross-validation for hyperparameter tuning [28]. It involves systematically searching through a specified parameter grid and evaluating each combination of hyperparameters to identify the optimal configuration. One of the main features of grid search CV is its extensive search capability and simplicity.

Table 3: Best hyperparameters for Dataset 1 and Dataset2

Algorithms	Best Hyperparameters (Dataset 1)	Best Hyperparameters (Dataset 2)
RF	'max_depth': None 'min_samples_split': 2 'n_estimators': 200	'max_depth': 10 'min_samples_split': 2 'n_estimators': 100
XGBOOST	learning_rate': 0.3 max_depth': 7 n_estimators': 500	learning_rate': 0.01 max_depth': 5 n_estimators': 1000
SVM	C=10 epsilon=0.5 gamma='auto'	C=10 epsilon=0.2 gamma='auto'
KNN	metric='manhattan' n_neighbors=3 weights='distance'	metric='manhattan' n_neighbors=3 weights='distance'

3. Results and Discussion

This study used evaluation metrics like the coefficient of determination (R^2) and Root Mean Square Error (RMSE). The dataset was divided into two parts, with 80% allocated for model training and 20% reserved for testing.

In this section, the PM 2.5 modeling performances of RF, XGBoost, KNN, and SVM are compared. The performance of these models on the Ratnapark station dataset 1 is presented in Table 4. The first model performance was evaluated on Sentinel-5P data only, and after that, it was again evaluated on Sentinel-5P data, including metrological data.

The results reveal that RF and XGBoost demonstrated superior performance compared to SVM and KNN. Across both, RF performed slightly better than the others. In Sentinel-5P data only, the RF obtained R^2 and RMSE of 0.82 and 13.17, respectively, whereas including metrological data in Sentinel AP data, R^2 and RMSE were 0.75 and 15.60, respectively. The results of dataset 1 indicated that SVM performed poorly, with the lowest R^2 values of 0.62 and 0.67 and the highest RMSE values of 19.38 and 18.07. Additionally, the addition of meteorological data in dataset 1 appears to have a significant effect on model performance.

Table 4. Model performance on Ratnapark dataset 1 with and without metrological data

Algorithm	Sentinel-5P		Sentinel-5P and metrological data	
	R^2	RMSE	R^2	RMSE
RF	0.82	13.17	0.75	15.60
XGBOOST	0.80	13.83	0.74	15.98
SVM	0.62	19.38	0.67	18.07
KNN	0.74	16.10	0.80	14.05

To see the model performance on datasets of different locations, Dataset 2 was prepared by combining three station data sets (Ratnapark, Shankapark, and Pulchowk). The model's performance was evaluated on Sentinel-5P air pollution data with and without metrological data. The performance of the model on dataset 2 is shown in Table 5. In Sentinel-5P data only, RF showed the lowest RMSE of 11.68, whereas after including the metrological data, RF showed the lowest RMSE of 11.36, indicating better accuracy in predicting PM2.5 concentrations.

Table 5. Model performance on dataset 2 with and without metrological data

Algorithm	Sentinel-5P		Sentinel-5P and metrological data	
	R^2	RMSE	R^2	RMSE
RF	0.80	11.68	0.816	11.36
XGBOOST	0.79	11.93	0.814	11.43
SVM	0.62	16.17	0.67	15.12
KNN	0.67	15.08	0.72	13.87

In dataset 2, for both cases of including and excluding metrological data with Sentinel-5P data, RF achieved the highest R^2 score of 0.816, indicating a good fit of the model to the data. XGBoost showed competitive performance with the RF in both cases. SVM had the highest RMSE values among all algorithms in both cases, indicating lower accuracy and larger errors in predictions. Also, SVM achieves relatively lower R^2 and RMSE scores compared to RF, XGBoost and KNN, suggesting a poorer fit of the model to the data. Overall, RF had the best performance across most metrics for both cases, followed by XGBoost and KNN, while SVM performed relatively poorly.

From the Table 4 result, it is seen that, when metrological data were added to dataset 1, the model performance seemed to have slightly declined. We have only used RH, wind speed, and temperature because other meteorological data such as wind direction, cloud cover, precipitation, and air pressure are not available from DOHM, Nepal. The model's performance might have been impacted in some way by the exclusion of these data. The linear interpolation method was used to handle the missing values; it may introduce some level of uncertainty or error in the data, especially if there are large gaps between observations. Also, in the combined dataset, the larger number of actual values available may help mitigate the impact of missing data. Additionally, having data from multiple stations may provide more robust and representative information about the underlying patterns and relationships in the data [29]. As we increased the dataset by combining three AQM stations (Ratnapark, Shankapark, and Pulchowk) data in dataset 2, the result of Table 5 shows that the model performance improved after including the metrological data. The average change in R^2 and RMSE values across all algorithms is +0.032 and -0.77, respectively, indicating an overall improvement in model performance on dataset 2 compared to dataset 1. It was seen that the size of the dataset had a great impact on the model performance.

With the help of the AQI breakpoint table presented in Table 6, a further analysis was carried out to figure out if the actual and expected levels of the Air Quality Index (AQI) would be identical based on the estimated PM2.5 values obtained. By looking at the results shown in Table 7, it is seen that our model can also be used to determine the AQI level. The predicted PM 2.5 value is in the same AQI category as the actual PM 2.5 category.

Table 6. AQI breakpoint table by US Environmental Protection Agency (EPA) [30]

PM 2.5 (µg/m³)	AQI	AQI category
0.0 - 12.0	0 - 50	Good
12.1 - 35.4	51 - 100	Moderate
35.5 - 55.4	101 - 150	Unhealthy for Sensitive Groups
55.5 - 150.4	151 - 200	Unhealthy
150.5 - 250.4	201 - 300	Very Unhealthy
250.5 - 500.4	301 - 400	Hazardous

Table 7. AQI level analysis based on predicted PM 2.5

PM 2.5 (Actual)	AQI Level (Actual)	PM 2.5 (Predicted)	AQI Level (Predicted)
12.7	Moderate	15.63	Moderate
64.51958	Unhealthy	64.33551	Unhealthy
83.64704	Unhealthy	81.72928	Unhealthy
45.77664	Unhealthy for sensitivity groups	39.62779	Unhealthy for sensitivity groups

3.2 Limitations of the Study

One limitation of the study is the availability of Sentinel-5P data from Google Earth Engine (GEE), which started in 2018. Additionally, data from air quality monitoring (AQM) stations showed inconsistencies due to device malfunctions. Consequently, the models were trained using limited datasets, which may not provide sufficient information for robust machine learning (ML) model training. Future work should focus on increasing the datasets and addition of other metrological data and relevant features. Additionally, checking the performance of deep learning algorithms could be a promising direction for further work.

4. Conclusion

This study illustrated the use of Sentinel-5P air pollution data and metrological data for the estimation of PM 2.5 using machine learning techniques. Our findings confirm that air pollution data obtained from Sentinel-5P can be used for the estimation of PM 2.5. Taking advantage of satellite data and a cloud platform like GEE is the most cost-effective and efficient method for AP retrieval. Over the four machine learning algorithms used, the performance of RF was found to be superior with the R^2 of 0.81 and RMSE of 11.36, while the performance of SVM was found to be the worst in all the scenarios. This study also confirmed that the addition of metrological data had a significant impact on model performances, and it was observed that there was an improvement in model performance after adding the metrological data. This study concludes that adequately trained machine learning models, utilizing sufficient data, hold promise for accurately estimating PM 2.5 levels at local scales. The integration of Sentinel-5P air pollution data and meteorological data presents an economically feasible solution, particularly in regions such as Nepal, where the establishment and maintenance costs of traditional air quality monitoring stations are costly.

References

[1] Xing, Y.F., Xu, Y.H., Shi, M.H., and Lian, Y.X. The impact of PM2.5 on the human respiratory system. *J Thorac Dis*, 8 (1) (2016) E69-E74.

[2] Batur, I., Markolf, S.A., Chester, M.V., Middel, A., Hondula, D., and Vanos, J. Street-level heat and air pollution exposure informed by mobile sensing. *Transp. Res. Part D Transp. Environ*, 113 (2022) 103535.

[3] Khedo, K., Perseedoss, R., and Mungur, A. A Wireless Sensor Network Air Pollution Monitoring System. *International Journal of Wireless & Mobile Networks*, 2 (2010) 10.5121/ijwmn.2010.2203.

[4] Engel-Cox, J., Holloman, C.H., Coutant, B.W., and Hoff, R.M. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, 38 (16) (2004) 2495-2509.

[5] Di, Q., Amini, H., Shi, L., et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130 (2019) 104909.

[6] Berrocal, V., Guan, Y., Muyskens, A., Wang, H., Reich, B., Mulholland, J., and Chang, H. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration. *Atmospheric Environment*, 222 (2019) 117130.

[7] Moursi, A., Shouman, M., Hemdan, E.E., and El-Fishawy, N. PM2.5 Concentration Prediction for Air Pollution using Machine Learning Algorithms. *Menoufia Journal of Electronic Engineering Research*, 28 (2019) 10.21608/MJEER.2019.67375.

[8] Wong, P.-Y., Lee, H.-Y., Chen, Y.-C., Zeng, Y.-T., Chern, Y.-R., Chen, N.-T., Lung, S.-C.C., Su, H.-J., and Wu, C.-D. Using a land use regression model with machine learning to estimate ground-level PM2.5. *Environ. Pollut*, 277 (2021) 116846.

[9] Air pollution in Kathmandu. (2020, March 23). *The Ecologist.* Retrieved from https://theecologist.org/2020/mar/23/air-pollution-kathmandu

[10] Department of Environment, Nepal. Main Report. (2021). Retrieved from https://doenv.gov.np/progressfiles/Main-report-60120-1660561765.pdf

[11] Devkota, B. and Neupane, P. Status of Air Pollution and its impacts in Nepal: A Review Study. (2020).

[12] Doreswamy, H., Harishkumar, K.S., Km, Y., and Gad, I. Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. *Procedia Computer Science*, 171 (2020) 2057-2066.

[13] Li, T., Hua, M., and Wu, X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). IEEE Access, (2020) 10.1109/ACCESS.2020.2971348.

[14] Wang, Z., Chen, L., Tao, J., Zhang, Y., and Su, L. Satellite-based estimation of regional particulate matter (PM) in Beijing using vertical-and-RH correcting method. *Remote Sensing of Environment*, 114 (2010) 50-63.

[15]Li, T., Shen, H., Yuan, Q., and Zhang, L. Geographically and temporally weighted neural networks for satellite-based mapping of ground-level PM2.5. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167 (2020) 178-188.

[16]Zheng, T., Bergin, M.H., Hu, S., Miller, J.D., and Carlson, D.E. Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach. *Atmospheric Environment*, 230 (2020) 117451.

[17]Song, Y.-Z., Yang, H.-L., Peng, J.-H., Song, Y.-R., Sun, Q., and Li, Y. Estimating PM2.5 concentrations in Xi'an City using a generalized additive model with multi-source monitoring data. PLoS ONE, 10 (2015) e0142149.

[18]Becker, S., Sapkota, R.P., Pokharel, B., Adhikari, L., Pokhrel, R.P., Khanal, S., and Giri, B. Particulate matter variability in Kathmandu based on in-situ measurements, remote sensing, and reanalysis data. *Atmospheric Research*, 258 (2021).

[19]Department of Environment, Nepal. Daily average PM2.5 data collection from the Department of Environment office at Forest Complex, Babarmahal, Kathmandu. (2019-2021).

[20]Department of Hydrology and Meteorology, Nepal. Meteorological data request payment procedure. (2019-2021).

[21]Qu, L.A., Chen, Z., Li, M., Zhi, J., and Wang, H. Accuracy improvements to pixel-based and object-based lulc classification with auxiliary datasets from Google Earth engine. *Remote Sens*, 13 (2021) 453.

[22]Shami, S., Ranjgar, B., Bian, J., Azar, M.K., Moghimi, A., Amani, M., and Naboureh, A. Trends of CO and NO2 Pollutants in Iran during COVID-19 pandemic using Timeseries Sentinel-5 images in Google Earth Engine. *Pollutants*, 2 (2022) 156-171.

[23]Noor, N., Abdullah, M.M.A.B., Yahaya, A.S., Ramli, N., and Fitri, N.F.M.Y. Estimation of Missing Values in Environmental Data Set using Interpolation Technique: Fitting on Lognormal Distribution. *Australian Journal of Basic and Applied Sciences*, 7 (2013) 336-341.

[24]Breiman, L. Random Forests. *Machine Learning*, 45 (2001) 5-32.

[25]Cortes, C. and Vapnik, V. Support-vector networks. *Chem. Biol. Drug Des*, 297 (2009) 273-297.

[26]Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016) 785-794.

[27]Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 (1) (1967) 21-27.

[28]Belete, D.M. and Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 44 (9) (2021) 875-886.

[29]Joel, L. and Doorsamy, W. A Review of Missing Data Handling Techniques for Machine Learning. (2022) 10.15157/IJITIS.2022.5.3.971-1005.

[30]U.S. Environmental Protection Agency (EPA). Technical assistance document for the reporting of daily air quality – the Air Quality Index (AQI). (2018) Retrieved from https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf