IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

Prediction of PM_{10} using Random Forest regression and Landsat 8 data

P Durga Devi* and M Chandrakala

Department of Electronics and Communication Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

ABSTRACT: Air quality monitoring is difficult in remote areas. The ground stations can record the pollution parameters upto some limited area only. It may not be possible to establish pollution monitoring stations in all the areas. In this regard, satellite images can help us to monitor remote areas. In this paper, we have taken the reflectance values of Landsat 8 images for the years from 2016-2021 as input parameters to the regression model, random forest. The random forest algorithm is tuned with hyperparameters and k-fold cross validation is also used to get more accurate results. Finally, the predicted values of PM10are compared with the true values of PM10 that are collected by ground stations.A high correlation is established between satellite reflectance values and ground data. The coefficient of correlation that we obtained is, R2=90.08. The evaluation metrics, MAE, RMSE, and RRMSE are 6.24, 8.47, and 6.58 respectively.

KEYWORDS: Particulate matter, Remote sensing, Random Forest, Landsat 8.

I. INTRODUCTION

Air pollution is being increased drastically with the rapid growth of industries in the major cities of India. Hyderabad is identified as the fourth most polluted city in the country. The air pollutants cause skin diseases, and severe respiratory problems, sometimes leading to death also. $PM_{10}$is particulate matter with a particle of size 10 microns or less than that [1]. PM_{10} is more prone to accumulate on the surfaces of the bigger airways in the upper part of the lung. Particles that are deposited on the surface of the lung might cause lung inflammation and tissue damage. The incidence and mortality of COVID-19 are significantly correlated with cities with poor air quality. On the other hand, death rates were lower in cities with better air quality. The medical exaggeration of the condition may become more severe due to air pollution [2].

Air pollution can be measured by establishing ground stations. But it is a time-consuming process and not possible in remote areas. This paper is focused on predicting PM_{10} using satellite images. In the present work, the PM_{10} is predicted by training the machine learning model called random forest. A Random Forest is an ensemble of unpruned classification or regression trees produced by employing random feature selection in tree induction and bootstrap samples of the training data [3].

The PM_{10} of Principado de Asturias (Spain) was predicted by applying the random forest model to Landsat 5, and Landsat 8 data and obtained a relative value of root mean square error of 25.29% [4]. The reflectance values are obtained by passing the window of different sizes over Landsat TM images and a correlation is established using the regression model with an RMSE of 16 [5]. Wang, Zhaoxi et al. correlated the aerosol optical depth of MODIS data with acute health effects in public. They considered the data sets of outpatient visits, patient admissions, and mortality records in Beijing, China, and compared them with PM_{10} that is collected from ground stations [6]. Heidar Maleki et al. applied the ANN model to predict AQI and air quality health index and obtained RMSE is 59.9 [7]. The PM_{10} of the area Ankara, Turkey is predicted by training the machine learning models viz., LASSO, SVR, RF, kNN, xGBoost, and ANN. Among all of them, ANN has given the best result with RMSE=20.8, R^2=0.58, and MAE=14.4 [8].

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

II. DATA AND METHODS

Study area

Hyderabad is the capital city of Telangana state, India. The location map is shown in Fig. 1. The air pollution getting increased rapidly as it is an industrially fast-growing city. This paper is focused on predicting particulate matter (PM10) using remote sensing.

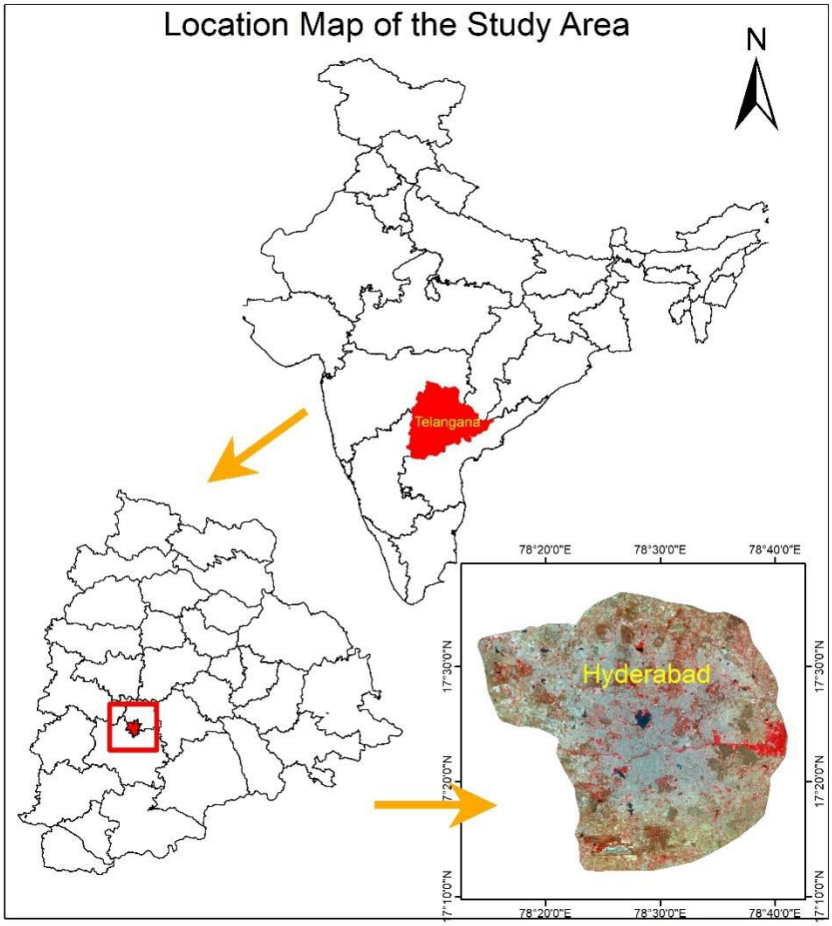

Fig. 1 Location map of the study area

Data acquisition

In-situ data of PM10 for the five years i.e., from 2017 to 2021 are taken from the Telangana pollution control board. The LANDSAT 8, level 2 satellite images for the respective months are taken from USGS explorer. For the present work, 40 images were taken. There is a total of eleven bands in LANSAT 8. But we have selected visible and infrared bands listed in Table1.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

Table1: Landsat 8 bands that are used in the present work

Bands	Wavelength (micrometers)	Resolution (meters)
Band 2 - Blue	0.45-0.51	30
Band 3 - Green	0.53-0.59	30
Band 4 - Red	0.64-0.67	30
Band 5 - Near Infrared (NIR)	0.85-0.88	30
Band 6 - SWIR 1	1.57-1.65	30
Band 7 - SWIR 2	2.11-2.29	30

Pre-processing of satellite data

The study area is extracted from the satellite image using the QGIS digitization tool. To extract the area of Hyderabad, we have considered the outer ring road as the boundary to the shapefile. The LANDSAT 8 image consists of DN values. These DN values are converted to surface reflectance using the scaling method. The reflectance values of the bands are used as independent values to predict the dependent parameter, PM10.

Random forest regression model

The regression methods are used for predicting the dependent parameter by taking the independent inputs. The commonly used linear regression model gives better results when the data is linear. But it is not suitable for non-linear data. In such cases, non-linear regression methods can be implemented. In this work, the machine learning algorithm; the random forest (RF) is used. The flow chart of the work is shown in Fig.2.

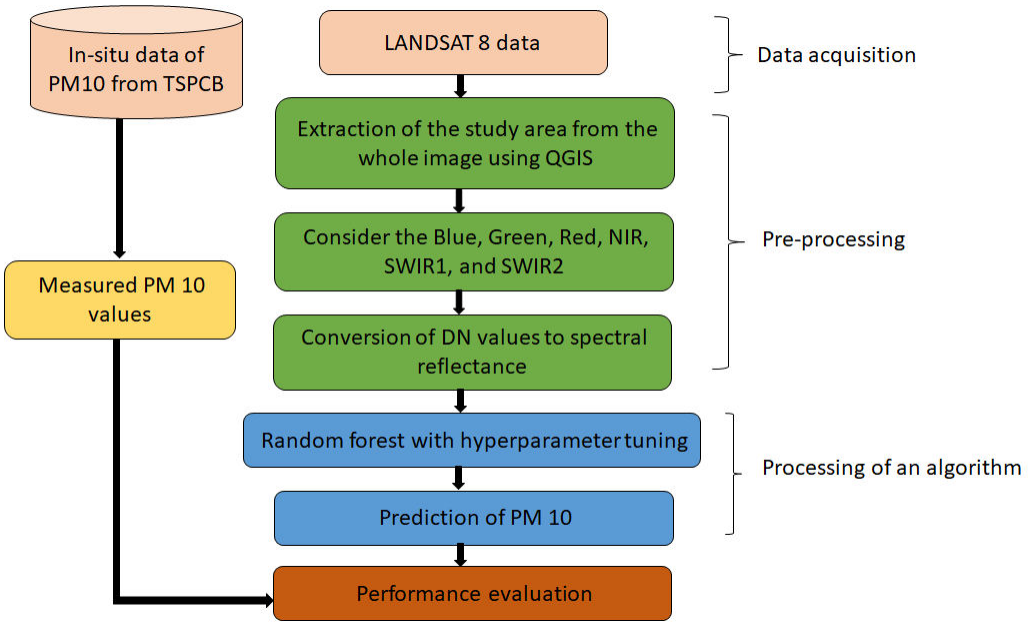

Fig. 2 The workflow of the RF algorithm

RF is an effective tool for regression and classification. To get more accurate results, the algorithm is tuned with hyperparameters, and 3-fold cross-validation is also implemented. Finally, the predicted values are compared with ground data. To evaluate the model, mean absolute, and root mean square errors are taken as performance metrics.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

III. RESULTS

The reflectance values of images are significantly correlated with the true values of PM10 that are collected from ground stations. The high reflectance indicates more PM10 i.e., more pollution. Here we compared the three images that are taken before lockdown, during the lockdown, and after the lockdown period i.e., April month of 2019, 2020, and 2021 respectively are shown in Fig. 3. The more the green shade, the less PM10 and vice versa. Among the three images, April 2020 image has more green shade compared to the other.

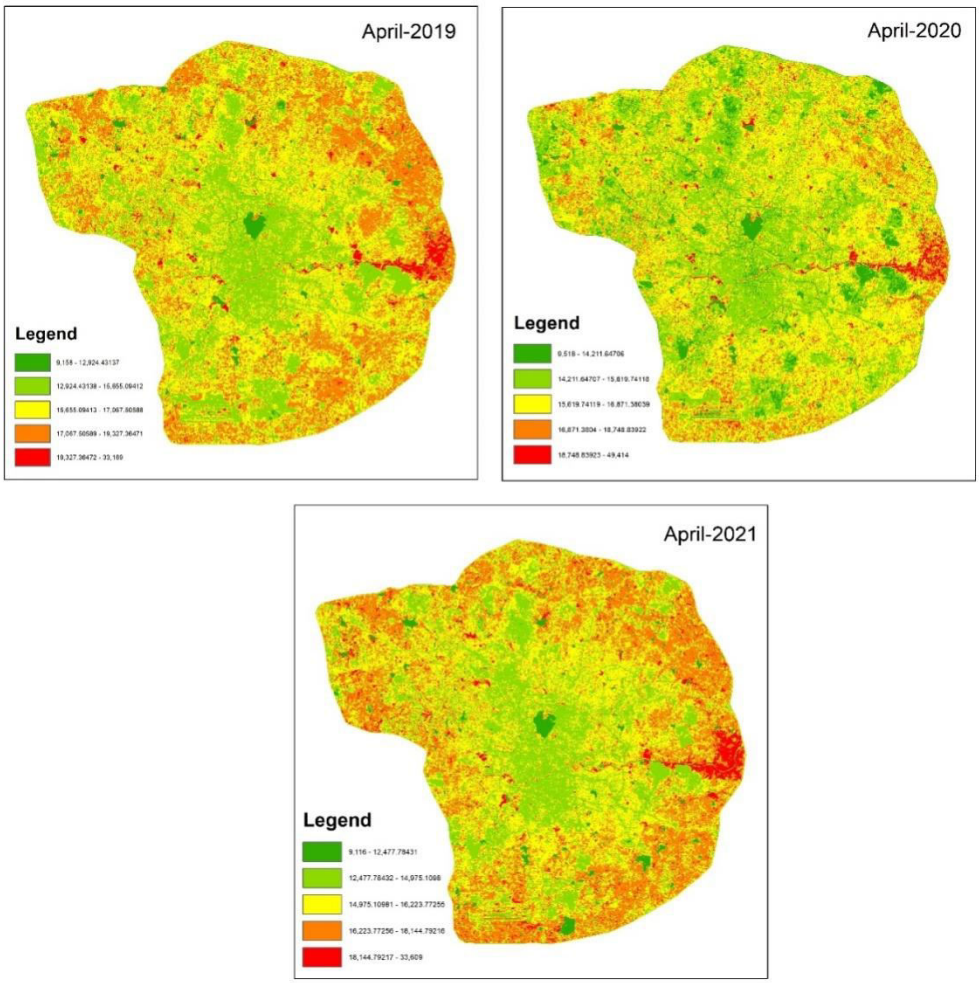

Fig. 3 Spectral reflectance maps of April 2019,2020,2021. (indicates lowest reflectance, indicates highest reflectance)

The random forest algorithm has given better results compared to the linear regression model. A high correlation is established between the spectral reflectance of satellite images and physically measured values by training the RF model. The regression plots of the RF model and the linear regression model are shown in Figs.4 (a) and (b)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

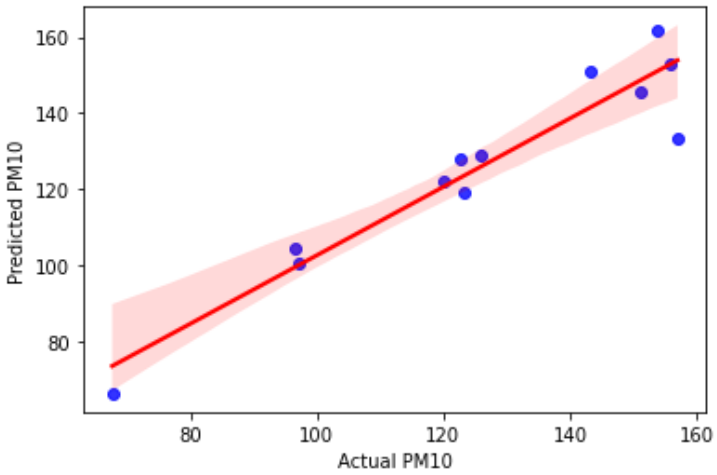

Fig.4(a)Regression plot for RF model

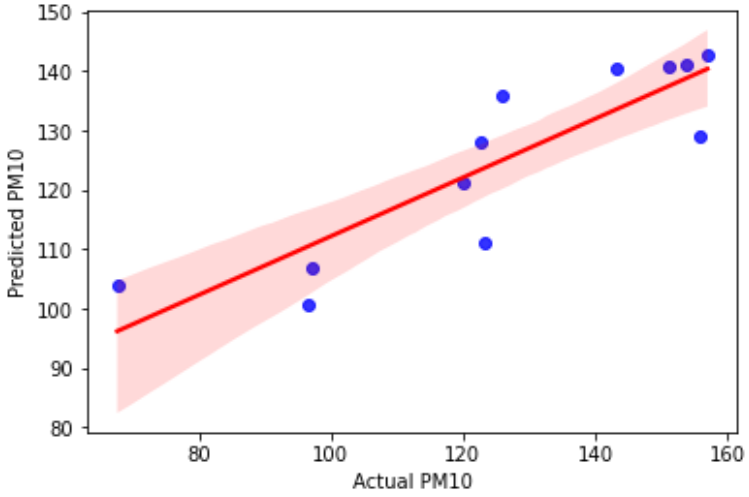

Fig.4(b) Regression plot for the linear regression model

It is clear from the two regression plots that, the points are closer to the line of fit in the RF model compared to the other. The coefficient of regression (R2) is 90.08 whereas, for the linear model, it is 66.75 only.

The comparison plots are shown in figures 5(a) and (b). From figure a, it is observed that the gap between measured and predicted values are less in the RF model.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

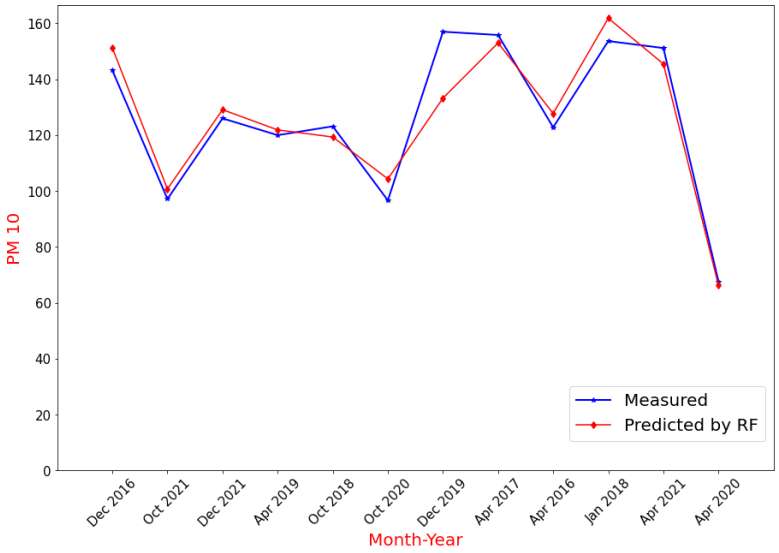

Fig.5(a)The comparison graph of in-situ data and RF predicted values of PM_{10}

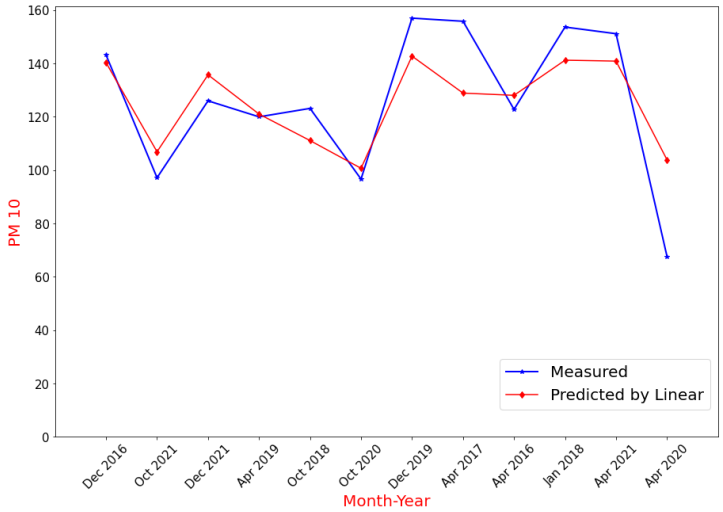

Fig.5(b) The comparison graph of in-situ data and linear regression predicted values of PM_{10}

Any model can be evaluated based on its error metrics. The evaluation metrics are given in table 2. Here the RF model is compared with the conventional linear regression model in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Relative value of RMSE (RRMSE).

Table2: Performance evaluation metrics

Regression model	MAE	RMSE	RRMSE(%)
RF	6.24	8.47	6.58
Linear	12.07	15.51	12.31

IV. CONCLUSION

In the present work, the RF model with hyperparameter tuning and 3-fold cross-validation has given a better result compared to the linear model. Almost 48 % reduction in mean absolute error, 45% reduction in RMSE, and 46.5%

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106329|

reduction in RRMSE. So, it is concluded that the remote sensing data has a significant correlation with in situ data of air pollutants.

ACKNOWLEDGMENT

The in-situ data of the air pollutant, PM10 is taken from the TSPCB official website. So, the authors would like to thank them.

REFERENCES

[1] World Health Organization. Ambient (Outdoor) Air Quality and Health: Fact Sheet. 2016. Available online: http://www.who.int/mediacentre/factsheets/fs313/en/ (accessed on 31 May 2018).

[2] H. R. Naqvi, M. Datta, G. Mutreja, M. A. Siddiqui, D. F. Naqvi, and A. R. Naqvi, "Improved air quality and associated mortalities in India under COVID-19 lockdown," *Environ. Pollut.*, vol. 268, no. 2, p. 115691, 2021, doi: 10.1016/j.envpol.2020.115691.

[3] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003, doi: 10.1021/ci034160g.

[4] V. M. Fernández-Pacheco *et al.*, "Estimation of PM10 Distribution using Landsat5 and Landsat8 Remote Sensing," p. 1430, 2018, doi: 10.3390/proceedings2231430.

[5]H. S. Lim, M. Z. MatJafri, K. Abdullah, and C. J. Wong, "Air Pollution Determination Using Remote Sensing Technique", in Advances in Geoscience and Remote Sensing. London, United Kingdom: IntechOpen, 2009 [Online]. Available: https://www.intechopen.com/chapters/9534 doi: 10.5772/8319

[6] Z. Wang *et al.*, "Acute health impacts of airborne particles estimated from satellite remote sensing," *Environ. Int.*, vol. 51, pp. 150–159, 2013, doi: 10.1016/j.envint.2012.10.011.

[7] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean Technol. Environ. Policy*, vol. 21, no. 6, pp. 1341–1352, 2019, doi: 10.1007/s10098-019-01709-w.

[8] A. Bozdağ, Y. Dokuz, and Ö. B. Gökçek, "Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey," *Environ. Pollut.*, vol. 263, 2020, doi: 10.1016/j.envpol.2020.114635.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details