Environmental Health Engineering and Management Journal 2023, 10(2), 157–163
http://ehemj.com
doi 10.34172/EHEM.2023.18
Environmental Health Engineering and Management Journal
Original Article
Open Access Publish Free

A survey on air pollutant PM2.5 prediction using random forest model

Sherin Babu[1,2*], **Binu Thomas**[2,3]

[1]Department of Computer Science, Assumption College Autonomous, Changanacherry, Kottayam, Kerala, India
[2]School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
[3]Department of Computer Applications, Marian College Autonomous, Kuttikanam, Idukki, Kerala, India

Abstract

Background: One of the most critical contributors to air pollution is particulate matter ($PM_{2.5}$) that its acute or chronic exposure causes serious health effects to human. Accurate forecasting of $PM_{2.5}$ concentration is essential for air pollution control and prevention of health complications. A survey of the available scientific literature on random forest model for $PM_{2.5}$ prediction is presented here.

Methods: The scientific literature is extracted from Science Direct database based on a set of specified search criteria. The input features, data length, and evaluation parameters used in $PM_{2.5}$ prediction were analyzed in this study.

Results: The study shows that majority of the publications are aimed at the daily prediction of outdoor $PM_{2.5}$. Most publications base their $PM_{2.5}$ prediction on features aerosol optical depth (AOD) and boundary layer height (BLH). PM_{10} and NO_2 are the main air pollutants employed in the $PM_{2.5}$ estimation. Majority studies utilized input data lengths covering more than one year, and the effectiveness of prediction models are unaffected by the length of investigation. The coefficient of determination, R^2, is the primary evaluation parameter used in all publications. The majority of research study indicated R^2 values greater than 0.85, demonstrating the reasonable dependability and efficiency of random forest regression-based $PM_{2.5}$ prediction models.

Conclusion: The study demonstrates that the publications use a variety of meteorological and geological features for $PM_{2.5}$ estimation, depending on the context of the research as well as data accessibility. The findings demonstrate that it is hard to pinpoint the optimal model in any particular way.

Keywords: Air pollution, Air pollutants, Aerosols, Particulate matter, Machine learning

Citation: Babu S, Thomas B. A survey on air pollutant PM2.5 prediction using random forest model. Environmental Health Engineering and Management Journal 2023; 10(2): 157–163. doi: 10.34172/EHEM.2023.18.

Article History:
Received: 24 May 2022
Accepted: 10 December 2022
ePublished: 25 May 2023

***Correspondence to:**
Sherin Babu,
Email: sherinbabu@assumptioncollege.edu.in

Introduction

Air pollution is considered as a serious threat to public health across the world. It can adversely affect the length and quality of human life. In 2018, the World Health Organization (WHO) reported that about 90 percent of people around the world breathe polluted air (1). Particulate matter ($PM_{2.5}$) is a term used to describe fine, inhalable mixtures of solid and liquid particles with diameters smaller than 2.5 µm that can linger in the atmosphere for an extended period of time and pose major health risks (2,3). Fuel combustion and atmospheric chemical processes result in the formation of these particles. Fine particulate matter air pollutant ($PM_{2.5}$) is a significant public health problem, especially for older people and young children (4,5). $PM_{2.5}$ can penetrate deeply into the lungs, and hence, the exposure to high concentrations of $PM_{2.5}$ will cause respiratory and cardiovascular diseases (6-9). High concentrations of $PM_{2.5}$ in the lower atmosphere can lead to the formation of haze, which causes slight obscuration in the visibility, and thus, leading to road accidents and transportation delays (10). In a large number of recent time-series studies, the atmospheric particulate matter has been reported as a causal factor for morbidity (11-14). Studies have revealed that the exposure to fine particulate matter over a long term cause increased mortality rate (15-18). In 2017, the Global Burden of Diseases report ranked particulate matter out of a list of 84 risk factors as the sixth leading cause of human death (19). In light of these findings, environmental scientists and public health workers all around the world are becoming more concerned about the rising trend of particulate matter in the metropolitan areas. Air pollutant concentration predicting is an effective way of protecting public health by providing an early warning, and also, for taking precautionary actions and in turn, ensuring clean and fresh air in the future.

Machine learning (ML) systems possess the ability to learn automatically without specific programming and develop from experience (20). Today, in almost all fields of research, the application of machine learning techniques can be found, from plant identification to drug discovery. Machine learning techniques can identify patterns and associations underlying the large and complex datasets, and thus, generate knowledge from them (21-23). Increased computing capacity allowed the development of advanced machine learning algorithms such as multiple linear regression, artificial neural networks, support vector machines (SVM) regression, random forest regression (RFR), and deep learning models for accurate and efficient prediction of various air pollutants (24,25). Studies have shown that the prediction of air pollutant concentrations by machine learning algorithms resulted in higher prediction accuracy.

Random forest is a machine-learning algorithm for classification and regression, which uses an ensemble of decision trees, with ample strength in handling complex nonlinear relationships within variable (26,27). Random forest is a method of ensemble learning that provides high precision and interpretability for predictions (28-30). Random forest allows nonlinearities and interactions to be learned from the data without any need to explicitly model them, thus, enabling them to exhibit superior performance to traditional statistical models (31-33). RFR is a supervised learning algorithm that uses an ensemble learning method for regression. In RFR, each node is divided into two or more child nodes, using the best subset of predictors randomly selected at that node. The data in each child node is used to predict dependent variable values within that node. The results are then combined from all child nodes to generate final predictions (26,34,35).

There are air pollution prediction surveys that have been released with different emphases in recent years. But no research surveys are being carried out on the estimation of $PM_{2.5}$ using machine learning methods. One of the finest regression algorithms for features with non-linear correlations is RFR, which offers improved accuracy, reduction of overfitting, and performs well. Therefore, this investigation was conducted to get an overview of what research work has been done regarding the application of the random forest algorithm in the $PM_{2.5}$ prediction. This will help recognize the potential gaps in this research area and lead the new researchers in the field to understand the state of the art.

Materials and Methods

This survey aimed to get insight into what studies have been published in the domain of air pollutant $PM_{2.5}$ prediction and random forest technique. Before conducting the survey, the research questions were defined. For this survey, the following three research questions (RQs) were defined:

- RQ1 - Which are the input features or variables used in the scientific literature for $PM2_{.5}$ prediction using random forest technique?
- RQ2 - What is the input data length used in the scientific literature for $PM_{2.5}$ prediction using random forest technique?
- RQ3 - Which are the evaluation parameters used in the scientific literature for $PM_{2.5}$ prediction using random forest technique?

When research questions were ready, the database for conducting the study was selected. The database used in this study is Science Direct. The data for this study were retrieved on December 18, 2020. The search string used for extracting the relevant literature is [“$PM_{2.5}$” AND “prediction” AND “random forest”]. This string is searched by the title, abstract, and keywords. After the search process, the obtained results were filtered and assessed using a set of exclusion criteria.

- Exclusion criteria 1 – The publication is not a research article.
- Exclusion criteria 2 – The publication year is not 2019.
- Exclusion criteria 3 – The language of publication is not English.
- Exclusion criteria 4 – Full text of the publication is not available.

Results

On the basis of the search string, 27 research publications were extracted during the search process. Then, the exclusion criteria were applied, and only eight full text publications remained for further analysis. During the data analysis, all the extracted data were investigated thoroughly, and the research questions were answered accordingly. The resultant publications of the query are shown in Table 1. In this table, the title of the research articles and journal of publication of these articles are presented.

Bai et al (36) proposed a random forest-based $PM_{2.5}$ data mining framework for the improvement of $PM_{2.5}$ prediction accuracy in eastern China. In this study, Gaussian-kernel-based interpolators were built to use $PM_{2.5}$ information from nearby sites and near-term historical observations to estimate spatially and temporally lagged $PM_{2.5}$ terms. For more precise $PM_{2.5}$ mapping, the predicted prior $PM_{2.5}$ details and variables such as aerosol optical depth (AOD) and meteorological conditions were then integrated into RFR models. The study claimed that the presence of ground-based $PM_{2.5}$ neighborhood information could greatly enhance $PM_{2.5}$ mapping precision. For regions with no prior $PM_{2.5}$ knowledge or for regions with few $PM_{2.5}$ monitoring sites, the prediction model did not work either.

Bi et al (37) developed a $PM_{2.5}$ prediction model based on the random forest algorithm to estimate fully covered and high-resolution ground $PM_{2.5}$ in New York State in

Table 1. Resultant publications

Authors	Title of the article	Journal
Bai et al (36)	Advancing the prediction accuracy of satellite-based $PM_{2.5}$ concentration mapping: A perspective of data mining through in situ $PM_{2.5}$ measurements	Environmental Pollution
Bi et al (37)	Impacts of snow and cloud covers on satellite-derived $PM_{2.5}$ levels	Remote Sensing of Environment
Di et al (38)	An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution	Environment International
Li and Zhang (39)	Predicting ground-level $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach	Environmental Pollution
Nabavi et al (40)	Assessing $PM_{2.5}$ concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms	Atmospheric Pollution Research
Stafoggia et al (41)	Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model	Environment International
Tang et al (42)	Comparison of GOCI and Himawari-8 aerosol optical depth for deriving full-coverage hourly $PM_{2.5}$ across the Yangtze River Delta	Atmospheric Environment
Wei et al (43)	Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach	Remote Sensing of Environment

2015. The model takes into account satellite AOD and the impacts of snow and cloud covers on AOD for $PM_{2.5}$ predictions. The author argued that this is the first research work that considered the snow-AOD relationship for $PM_{2.5}$ modelling. In order to estimate the missing AOD, a daily gap-filling model with snow and cloud fractions and meteorological explanatory variables were developed using the random forest algorithm. By using this gap-filled AOD model in New York State, a daily AOD data set with a 1-km resolution was generated for 2015. Then, a random forest model based on the gap-filled AOD and covariates was built to predict fully covered $PM_{2.5}$ estimates. The study was able to ascertain the importance of cloud and snow parameters in estimating the air pollutant $PM_{2.5}$ and the discernible interactions between snow/cloud and AOD/$PM_{2.5}$. The drawback of the research is that it took into account only the coverage of snow and cloud, not the physical features of snow and cloud.

Di et al (38) developed an ensemble model that combined three machine learning algorithms called the neural network, random forest, and gradient boosting and predictor variables to estimate daily $PM_{2.5}$ concentrations at a resolution of 1 km × 1 km across the contiguous United States from 2000 to 2015. The three-machine learning algorithms were fed with satellite data, meteorological variables, land-use variables, elevation, chemical transport model predictions, land use data, and few other reanalysis data. Thus, the predicted values of $PM_{2.5}$ were obtained from each learner. The study also calculated spatially and temporally lagged $PM_{2.5}$ predictions from nearby monitoring sites and neighboring days and treated them as additional input variables along with the above-mentioned $PM_{2.5}$ predictions. Then, a generalized additive model that accounted for the geographic difference to combine $PM_{2.5}$ estimates from the neural network, random forest, and gradient boosting was used as an ensemble model to combine $PM_{2.5}$ estimation. The benefit of research work is that the combined $PM_{2.5}$ estimates of the generalized additive model from three-machine learning algorithms allowed each algorithm's contribution to differ by location.

Li and Zhang (39) proposed a hybrid remote sensing and machine learning model, termed as remote sensing-random forest, which incorporated AOD, weather variables and air pollution variables into a modelling framework to predict daily $PM_{2.5}$ values in Beijing-Tianjin-Hebei (BTH) region of China. The study aimed to predict the spatiotemporal distributions of daily $PM_{2.5}$ concentrations across the BTH region during 2015-2017. The study took into account the dynamic and large monitoring capacity of AOD and the benefits of the RF technique in the management of complex nonlinear relationships. In addition, meteorological and air contaminant variables to predict daily $PM_{2.5}$ concentrations were incorporated into a general structure. The authors claimed that the model provides decision support for air pollution control at a regional environment during haze periods.

Nabavi et al (40) suggested a model for the spatial estimation of $PM_{2.5}$ using 10-km merged dark target and deep blue (DB_DT)-dependent AOD and 1-km Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD over Tehran, Iran. The authors argued that the limitations of the ground-based $PM_{2.5}$ measurements constrained them to estimate $PM_{2.5}$ using satellite AOD-fed statistical models. The researchers used both the MAIAC AOD algorithm, which provided a good estimate of the recovery of aerosols over both dark and light surfaces, and the DB-AOD algorithm, which provided efficient aerosol recovery over bright surfaces. Afterwards, planetary boundary layer height (PBLH) and relative humidity were used for the normalization of AOD and correction of $PM_{2.5}$, respectively. Then, the performance of four machine-learning algorithms namely RF, gradient boosting, multivariate adaptive regression splines, and SVM were investigated in the spatial estimation of $PM_{2.5}$. The study established that RF model fed by normalized 10-km DB_DT AOD yielded the most accurate estimate of $PM_{2.5}$ over Tehran region. The authors concluded that the use of high-resolution MAIAC AOD could not improve the prediction ability of any of the machine-learning algorithms employed here compared to 10-km DB_DT AOD.

Stafoggia et al (41) developed a five-stage random forest model to predict daily PM_{10}, $PM_{2.5}$, and $PM_{2.5-10}$ concentrations at fine spatial resolution in Italy during 2013 to 2015. In stage 1, $PM_{2.5}$ and $PM_{2.5-10}$ concentrations were predicted where only PM_{10} data were available. Stage 2 dealt with the assignment of missing satellite AOD data using atmospheric ensemble model. In stage 3, a relationship between measured PM concentrations and satellite, land use and meteorological parameters was established. Stage 4 involved applying the stage 3 model to each 1-km^2 grid cell in Italy. Stage 5 aimed to improve predictions done at stage 3, by using additional information at a finer spatial resolution. The authors argued that they were successful in predicting the daily PM_{10}, $PM_{2.5}$, and $PM_{2.5-10}$ concentrations in Italy using this five-stage random forest model with high accuracy rate. The model's downside was the low performance for $PM_{2.5-10}$ estimation, in Southern Italy and during the summer months.

Tang et al (42) performed a comparative evaluation of the performance of the Geostationary Ocean Color Imager (GOCI) AOD and Himawari-8 AOD datasets in predicting the hourly $PM_{2.5}$ in Yangtze River Delta (YRD) region of China at a spatial resolution of 1 km for 2017. The comparative evaluation was done using the nonparametric approach with two random-forest sub-models. The full-coverage AOD dataset was generated with the first RF sub-model, followed by the second RF sub-model for the $PM_{2.5}$ estimation. The first RF-sub-model analysis showed that in 2017, AOD obtained from the GOCI and the Hiamwari-8 showed moderately similar trends across YRD. Similar performance was also shown by the second RF sub-model estimate of hourly $PM_{2.5}$ concentrations using the GOCI and Himawari-8.

Wei et al (43) estimated $PM_{2.5}$ concentrations based on the MAIAC-AOD product using a space-time random forest (STRF) model, across China for 2016. In order to produce 1-km daily $PM_{2.5}$ concentrations, the STRF model considered MAIAC-AOD data, along with meteorological conditions, land use and human activities. It was revealed that the STRF model was superior to those of commonly used regression models, in both model efficiency and predictive capacity.

Discussion

Based on the three research questions, a review of the current collection of scientific literature on the random forest model for $PM_{2.5}$ prediction was conducted. To address the research questions RQ1, RQ2, and RQ3, the input features used, the year of study, and the evaluation parameters employed in the publications were investigated and summarized. The RQ parameters and data extracted along with the article title are shown in Table 2.

All the 8 selected publications, except Stafoggia et al (41) considered $PM_{2.5}$ as a single dependent variable. Stafoggia et al (41) model estimated the particulate matter PM_{10} and $PM_{2.5}$. The features extracted are grouped to provide a clear description of the independent variables (features). The independent variables are grouped into meteorological, air pollutants, satellite-derived AOD, transportation and traffic, population, land use, normalized difference vegetation index (NDVI), and road data. All the publications employed more than one independent feature group for the estimation of $PM_{2.5}$. The survey found that the key feature used by all the selected publications was AOD. AOD is a measurement that tells us how much direct sunlight is prevented by particles such as smoke, dust and haze from reaching the ground. The next widely used feature group is meteorological features. The variables that define atmospheric chemistry are known as meteorological parameters. The most common meteorological variables utilized are temperature, wind speed, surface pressure, and relative humidity. PM_{10}, BLH, and NO_2 are the next prominent input features used by the majority of the publications. Only a few research works used air pollutants such as ozone, SO_2, CO, etc. for the estimation of $PM_{2.5.}$ The only research work that employed the input features of snow cover is the research by Bi et al (37).

The research input data length used in the selected papers was divided into three groups: the study period less than or equivalent to 1 year, the study period longer than one year and less than five years, and the study period longer than five years. Out of the eight selected publications, four papers utilized input data with a length spanning more than one year, but less than five years. Furthermore, the use of data covering a period of less than one year occurred in 2 papers, while only two studies utilized data with lengths of more than five years.

All the resultant publications were aimed to predict outdoor $PM_{2.5}$, and none of the papers concentrated on the estimation of indoor $PM_{2.5}$. Out of the 8 publications, 5 publications estimated daily $PM_{2.5}$ values and 3 publications estimated hourly $PM_{2.5}$ values. The studies by Li et al (39), Nabavi et al (40), and Tang et al (42) estimated hourly $PM_{2.5}$ values. All other five papers concentrated on the daily estimation of $PM_{2.5}$ values. Most of the selected publications measured the concentration of $PM_{2.5}$ in China, followed by the United States of America, Iran, and Italy. All the selected publications analyzed the model performance using the evaluation parameter R^2. R^2, known as goodness-of-fit, specifies the percentage of the variance in the dependent variable that is predictable from the independent variables. The study by Li and Zhang (39) reported the highest R^2 value ($R^2 = 0.93$), followed by the

Table 2. Research question parameters used in the publications

Reference	Features used	Year of study	Evaluation parameter
Bai et al (36)	24-h averaged $PM_{2.5}$, AOD, temperature at 2 m (T), wind speed components at 10 m (U and V, m/s), PBLH (m), and RH (%), NDVI, and tropospheric NO_2 column density	2015-2016	Coefficient of determination (R^2), MPE, RPE
Bi et al (37)	Daily $PM_{2.5}$, AOD, coverage of snow and cloud, air temperature, dew-point temperature, surface pressure, specific humidity, wind speed, visibility, PBLH, potential evaporation, downward shortwave radiation, and convective available potential energy known as CAPE, land-use parameters, population, distances to highways and major roads, elevation, NDVI and dummy variables for months and Julian days	2015	R^2, RMSE
Di et al (38)	24-h averaged $PM_{2.5}$, AOD, accumulated total precipitation, air temperature, downward shortwave radiation flux, accumulated total evaporation, PBLH, low cloud area fraction, perceptible water for the entire atmosphere, pressure, specific humidity at 2 m, visibility, wind speed, medium cloud area fraction, high cloud area fraction, and albedo, chemical transport model-based data, Land-use coverage types from the National Land Cover Database including barren land, forest, shrubland, herbaceous land, cultivated land, developed areas, and wetlands. Other factors – elevation, road density, restaurant density, elevation, and NDVI	2000-2015	R^2, RMSE
Li and Zhang (39)	Hourly ground-level $PM_{2.5}$, AOD, meteorological variables (air temperature, relative humidity, wind speed, wind direction and pressure), and air pollutant variables (SO_2, NO_2, CO, O_3)	2015-2017	R^2, RMSE
Nabavi et al (40)	Hourly $PM_{2.5}$, combined DB_DT AOD, combined MAIAC AOD, solar zenith angle, relative humidity, PBLH, road density, wind speed, visibility, minimum temperature, mean temperature, maximum temperature, elevation, and day of the year	2011-2016	R^2, RMSE, MRE
Stafoggia et al. (41)	Daily $PM_{2.5}$, air temperature, PBL (hh 00.00), Julian day, Barometric pressure, elevation, PBL (hh 12.00), wind (v component), AOD (470 nm), AOD (550 nm), month, latitude, administrative region, precipitations, longitude, wind (u component), distance from sea, resident population, distance from emission points, distance from highways, geoclimatic zone, density of local streets, PM_{10} emissions from point sources, % low development, NDVI, PM_{10} emissions from areal sources, day of week, distance from airport, % arable land, distance from major roads, light at night, % deciduous, % agricultural, density of major and minor roads, % shrub, % crops, desert dust advection, % high development, % evergreen, imperviousness surface areas,% pasture, and density of highways	2013-2015	R^2, RMSPE
Tang et al (42)	Hourly $PM_{2.5}$, hourly AOD data from the GOCI and Himawari-8 products, AERONET AOD data, elevation, NDVI, hourly PBLH, land use types, atmospheric pressure, relative humidity, precipitation, temperature, vapor pressure, and wind field, population density data, and road density	2017	R^2, RMSE, RPE
Wei et al (43)	Daily $PM_{2.5}$, daily MAIAC AOD, AERONET AOD, 2-m air temperature, total precipitation, evaporation, BLH, 10-m U/V wind components, relative humidity, surface pressure, wind speed and wind direction, land-related variables, and population-related variables	2015-2016	R^2, RMSE, MPE

Abbreviations: NDVI, normalized difference vegetation index; AOD, aerosol optical depth; PBLH, planetary boundary layer height; RH, relative humidity; MPE, mean prediction error; RPE, relative prediction error; MRE, mean relative error; MAIAC, Multi-Angle Implementation of Atmospheric Correction; RMSPE, root mean squared prediction error; RMSE, root mean square error.

study of Di et al (38) with $R^2 = 0.89$. The studies by Bai et al (36) and Tang et al (42) reported the prediction accuracy of $R^2 = 0.86$. The next major evaluation parameter used is RMSE. RMSE is a method by which the difference between a model's predicted values and their actual values can be measured. Of the 8 publications, 6 publications employed RMSE as an evaluation parameter and the lowest RMSE reported is 1.78 μg/m^3 for the studies by Bi et al (37).

Conclusion

$PM_{2.5}$ is an air pollutant that has a wide variety of adverse health effects on the general wellbeing. For air pollution control, mapping $PM_{2.5}$ concentration is thus of vital importance. A survey of random forest-based prediction models for $PM_{2.5}$ prediction was performed in this study. The study showed that depending on the scope and background of the research and the availability of data, the selected publications use a variety of input features, both meteorological and geological features for the estimation of $PM_{2.5}$. AOD is the most important input feature that is utilized in most research studies. The predominant air pollutant features used in the studies are PM_{10} and NO_2. It was discovered that BLH is a major meteorological input element in the RFR-based $PM_{2.5}$ prediction. In the majority of research studies, input data lengths that span more than a year were used. It is noteworthy that the length of the study has no effect on the accuracy and performance of the prediction models. The RFR-based $PM_{2.5}$ estimating models performed well, despite the investigation lasting between one and two years. The performance measuring metric that is most frequently employed across all publications is the coefficient of determination (R^2). The majority of studies reported R^2 values more than 0.85, showing that the RFR-based $PM_{2.5}$ prediction models are relatively reliable and effective. RMSE is another performance measuring metric that is frequently used in research publications. The results showed that no specific conclusion could be drawn as to what the best model is. This study provided a concise and comprehensive reference for researchers in the field of a random forest-based machine learning model for $PM_{2.5}$ prediction.

Acknowledgements

The authors gratefully acknowledge the service provided by the Mahatma Gandhi University Library, Kerala, as well as those who contributed to perform this study.

Authors' contribution

Conceptualization: Sherin Babu.
Data curation: Sherin Babu.
Formal analysis: Sherin Babu, Binu Thomas.
Funding acquisition: Sherin Babu.
Investigation: Sherin Babu, Binu Thomas.
Methodology: Sherin Babu.
Project administration: Binu Thomas.
Resources: Binu Thomas.
Software: Sherin Babu.
Supervision: Binu Thomas.
Validation: Binu Thomas.
Visualization: Sherin Babu.
Writing–original draft: Sherin Babu.
Writing–review & editing: Sherin Babu, Binu Thomas.

Competing interests

The authors declare that there is no conflict of interests.

Ethical issues

The authors confirm that all data acquired during the research are as stated in the paper, and no data from the study has been or will be published elsewhere.

References

1. World Health Organization (WHO). 9 Out of 10 People Worldwide Breathe Polluted Air, But More Countries Are Taking Action. Geneva: WHO; 2018. Available from: https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action. Accessed September 16, 2020.
2. Amarloei A, Fazlzadeh M, Jonidi Jafari A, Zarei A, Mazloomi S. Particulate matters and bioaerosols during Middle East dust storms events in Ilam, Iran. Microchem J. 2020;152:104280. doi: 10.1016/j.microc.2019.104280.
3. Yunesian M, Rostami R, Zarei A, Fazlzadeh M, Janjani H. Exposure to high levels of PM2.5 and PM10 in the metropolis of Tehran and the associated health risks during 2016-2017. Microchem J. 2019;150:104174. doi: 10.1016/j.microc.2019.104174.
4. Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, et al. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. J Expo Anal Environ Epidemiol. 2001;11(3):231-52. doi: 10.1038/sj.jea.7500165.
5. Elsunousi AAM, Sevik H, Cetin M, Ozel HB, Ozel HU. Periodical and regional change of particulate matter and CO2 concentration in Misurata. Environ Monit Assess. 2021;193(11):707. doi: 10.1007/s10661-021-09478-0.
6. Xing YF, Xu YH, Shi MH, Lian YX. The impact of PM2.5 on the human respiratory system. J Thorac Dis. 2016;8(1):E69-74. doi: 10.3978/j.issn.2072-1439.2016.01.19.
7. Yin P, Guo J, Wang L, Fan W, Lu F, Guo M, et al. Higher risk of cardiovascular disease associated with smaller size-fractioned particulate matter. Environ Sci Technol Lett. 2020;7(2):95-101. doi: 10.1021/acs.estlett.9b00735.
8. Bose S, Hansel NN, Tonorezos ES, Williams DL, Bilderback A, Breysse PN, et al. Indoor particulate matter associated with systemic inflammation in COPD. J Environ Prot (Irvine, Calif). 2015;6(5):566-72. doi: 10.4236/jep.2015.65051.
9. Aryal A, Harmon AC, Dugas TR. Particulate matter air pollutants and cardiovascular disease: strategies for intervention. Pharmacol Ther. 2021;223:107890. doi: 10.1016/j.pharmthera.2021.107890.
10. Sajjadi SA, Atarodi Z, Lotfi AH, Zarei A. Levels of particulate matters in air of the Gonabad city, Iran. MethodsX. 2018;5:1534-9. doi: 10.1016/j.mex.2018.11.001.
11. Lippmann M, Ito K, Nádas A, Burnett RT. Association of particulate matter components with daily mortality and morbidity in urban populations. Res Rep Health Eff Inst. 2000(95):5-72.
12. Yang Y, Guo Y, Qian ZM, Ruan Z, Zheng Y, Woodward A, et al. Ambient fine particulate pollution associated with diabetes mellitus among the elderly aged 50 years and older in China. Environ Pollut. 2018;243(Pt B):815-23. doi: 10.1016/j.envpol.2018.09.056.
13. Maher BA, Ahmed IA, Karloukovski V, MacLaren DA, Foulds PG, Allsop D, et al. Magnetite pollution nanoparticles in the human brain. Proc Natl Acad Sci U S A. 2016;113(39):10797-801. doi: 10.1073/pnas.1605941113.
14. Shi L, Zanobetti A, Kloog I, Coull BA, Koutrakis P, Melly SJ, et al. Low-concentration PM2.5 and mortality: estimating acute and chronic effects in a population-based study. Environ Health Perspect. 2016;124(1):46-52. doi: 10.1289/ehp.1409111.
15. Burnett RT, Pope CA 3rd, Ezzati M, Olives C, Lim SS, Mehta S, et al. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environ Health Perspect. 2014;122(4):397-403. doi: 10.1289/ehp.1307049.
16. Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, et al. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. Environ Health Perspect. 2009;117(6):957-63. doi: 10.1289/ehp.0800185.
17. Bartell SM, Longhurst J, Tjoa T, Sioutas C, Delfino RJ. Particulate air pollution, ambulatory heart rate variability, and cardiac arrhythmia in retirement community residents with coronary artery disease. Environ Health Perspect. 2013;121(10):1135-41. doi: 10.1289/ehp.1205914.
18. Madrigano J, Kloog I, Goldberg R, Coull BA, Mittleman MA, Schwartz J. Long-term exposure to PM2.5 and incidence of acute myocardial infarction. Environ Health Perspect. 2013;121(2):192-6. doi: 10.1289/ehp.1205284.
19. Gakidou E, Afshin A, Abajobir AA, Abate KH, Abbafati C, Abbas KM, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1345-422. doi: 10.1016/s0140-6736(17)32366-8.
20. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric. 2020;177:105709. doi: 10.1016/j.compag.2020.105709.
21. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x.

22. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Mark. 2021;31(3):685-95. doi: 10.1007/s12525-021-00475-2.
23. Zhong S, Zhang K, Bagheri M, Burken JG, Gu A, Li B, et al. Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol. 2021;55(19):12741-54. doi: 10.1021/acs.est.1c01339.
24. Shezi B, Jafta N, Sartorius B, Naidoo RN. Developing a predictive model for fine particulate matter concentrations in low socio-economic households in Durban, South Africa. Indoor Air. 2018;28(2):228-37. doi: 10.1111/ina.12432.
25. Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ Pollut. 2019;245:746-53. doi: 10.1016/j.envpol.2018.11.034.
26. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi: 10.1023/a:1010933404324.
27. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett. 2010;31(14):2225-36. doi: 10.1016/j.patrec.2010.03.014.
28. Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos Environ (1994). 2017;151:1-11. doi: 10.1016/j.atmosenv.2016.11.066.
29. James G, Witten D, Hastie T, Tibshirani R. Tree-based methods. In: James G, Witten D, Hastie T, Tibshirani R, eds. An Introduction to Statistical Learning: With Applications in R. New York, NY: Springer; 2013. p. 303-35. doi: 10.1007/978-1-4614-7138-7_8.
30. Rakhra M, Soniya P, Tanwar D, Singh P, Bordoloi D, Agarwal P, et al. Crop price prediction using random forest and decision tree regression:-a review. Mater Today Proc. 2021. doi: 10.1016/j.matpr.2021.03.261.
31. Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, et al. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ Sci Technol. 2017;51(12):6936-44. doi: 10.1021/acs.est.7b01210.
32. Brokamp C, Jandarov R, Hossain M, Ryan P. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ Sci Technol. 2018;52(7):4173-9. doi: 10.1021/acs.est.7b05381.
33. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat. 2009;63(4):308-19. doi: 10.1198/tast.2009.08199.
34. Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, et al. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environ Pollut. 2018;242(Pt A):675-83. doi: 10.1016/j.envpol.2018.07.016.
35. Dudek G. Short-term load forecasting using random forests. In: Filev D, Jabłkowski J, Kacprzyk J, Krawczak M, Popchev I, Rutkowski L, et al, eds. Intelligent Systems'2014. Cham: Springer; 2015. p. 821-8. doi: 10.1007/978-3-319-11310-4_71.
36. Bai K, Li K, Chang NB, Gao W. Advancing the prediction accuracy of satellite-based PM2.5 concentration mapping: a perspective of data mining through in situ PM2.5 measurements. Environ Pollut. 2019;254(Pt B):113047. doi: 10.1016/j.envpol.2019.113047.
37. Bi J, Belle JH, Wang Y, Lyapustin AI, Wildani A, Liu Y. Impacts of snow and cloud covers on satellite-derived PM2.5 levels. Remote Sens Environ. 2019;221:665-74. doi: 10.1016/j.rse.2018.12.002.
38. Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environ Int. 2019;130:104909. doi: 10.1016/j.envint.2019.104909.
39. Li X, Zhang X. Predicting ground-level PM2.5 concentrations in the Beijing-Tianjin-Hebei region: a hybrid remote sensing and machine learning approach. Environ Pollut. 2019;249:735-49. doi: 10.1016/j.envpol.2019.03.068.
40. Nabavi SO, Haimberger L, Abbasi E. Assessing PM2.5 concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms. Atmos Pollut Res. 2019;10(3):889-903. doi: 10.1016/j.apr.2018.12.017.
41. Stafoggia M, Bellander T, Bucci S, Davoli M, de Hoogh K, De' Donato F, et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013-2015, using a spatiotemporal land-use random-forest model. Environ Int. 2019;124:170-9. doi: 10.1016/j.envint.2019.01.016.
42. Tang D, Liu D, Tang Y, Seyler BC, Deng X, Zhan Y. Comparison of GOCI and Himawari-8 aerosol optical depth for deriving full-coverage hourly PM2.5 across the Yangtze River Delta. Atmos Environ. 2019;217:116973. doi: 10.1016/j.atmosenv.2019.116973.
43. Wei J, Huang W, Li Z, Xue W, Peng Y, Sun L, et al. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens Environ. 2019;231:111221. doi: 10.1016/j.rse.2019.111221.