Check for updates

REVIEW

Prediction of PM_{2.5} concentrations in Malaysia using machine learning techniques: a review [version 1; peer review: 1 approved, 1 approved with reservations]

Naveen Palanichamy¹, Su-Cheng Haw¹, Subramanian S²,
Kuhaneswaran Govindasamy¹, Rishanti Murugan¹

¹Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, 63000, Malaysia

²Department of Electrical Engineering, Annamalai University, Chidambaram, Tamil Nadu, 608002, India

V1 First published: 14 Dec 2021, 10:1279
<https://doi.org/10.12688/f1000research.73163.1>

Latest published: 14 Dec 2021, 10:1279
<https://doi.org/10.12688/f1000research.73163.1>

Abstract

Particulate matter (PM), an air pollutant that is detrimental to breathing, is either emitted or formed ambiently. The exposure of respiratory system towards PM_{2.5}, the fine particles of 2.5 micrometres diameter, causes complication for health. Thus, developing pollution control strategies requires the prediction of PM_{2.5} concentrations. Advancement of technology and computer science knowledge, machine learning (ML) algorithms are used for highly accurate prediction of air pollutant concentrations. Recently, air quality in Smart Cities of Malaysia has been getting worse due to industrialization, emissions from private motor vehicles, and transboundary haze pollution. Therefore, the forecasting of PM_{2.5} emissions to ensure they are within the statutory limits becomes necessary. Several machine learning methods have been implemented in existing research to predict air pollution concentrations in comparison to PM_{2.5}. However, very few studies have used ML techniques to predict air quality in Malaysia when compared with global studies. Hence, to create awareness on the ML techniques and promote further research in this area, this study reviews and highlights most of the existing ML techniques for the prediction of PM_{2.5}.

Keywords

Air Pollution, Particulate Matter (PM_{2.5}), Neural Network, Deep Learning, Decision Tree

Open Peer Review

Approval Status

	1	2
version 1
14 Dec 2021	view	view

1. **Nagender Aneja**, Universiti Brunei

Darussalam, Bandar Seri Begawan, Brunei

2. **D. Devaraj**, Kalasalingam Academy of

Research and Education, Krishnankoil, India

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research Synergy**
Foundation gateway.

Corresponding author: Naveen Palanichamy (p.naveen@mmu.edu.my)

Author roles: **Palanichamy N:** Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Haw SC:** Supervision, Writing – Review & Editing; **S S:** Writing – Review & Editing; **Govindasamy K:** Investigation; **Murugan R:** Investigation

Competing interests: No competing interests were disclosed.

Grant information: This work is funded by Multimedia University (Ref: MMUI/210006).

Copyright: © 2021 Palanichamy N *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Palanichamy N, Haw SC, S S *et al.* **Prediction of PM_{2.5} concentrations in Malaysia using machine learning techniques: a review [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2021, 10:1279 <https://doi.org/10.12688/f1000research.73163.1>

First published: 14 Dec 2021, 10:1279 <https://doi.org/10.12688/f1000research.73163.1>

Introduction

Air quality is important for human health, crops, vegetation, and aesthetic considerations, for example, visibility. Air pollution, in which the air is contaminated with a variety of dirt and chemicals, is detrimental to breathing and can cause a wide variety of health defects and issues. Bad air is a combination of both natural and human-made sources of perilous substances. It was estimated in 2016 that outdoor air pollution in rural and urban areas caused 4.2 million premature fatalities worldwide annually; exposure to PM is the reason for the mortality, which causes cancers, and respiratory and cardiovascular diseases [[https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)].

A complicated mixture of ultrafine particles and vaporized tiny molecules or liquid droplets is known as particulate matter (PM) [<https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>]. PM are categorized according to their sizes. PM_{2.5} has a diameter of 2.5 microns or less. PM_{2.5} is inhalable and can travel farther into our bodies and deposit into alveoli, eventually passing into the bloodstream. They may cause cardiovascular diseases, since they start to mix into the blood stream [<https://blog.breezometer.com/what-is-particulate-matter>].

The health impact of exposure to PM_{2.5} is a nightmare which involves different age groups of people.¹ The number of coughs counted by urban workers who are exposed to PM_{2.5} determines how badly their respiratory system is in danger. Usmani² reviewed a paper and mentioned that environmental indicators and PM_{2.5} have a great impact on Malaysian health services *via* infant mortality rate, fertility rate, and life expectancy. He also concluded that outdoor and indoor air quality can affect the health of school children who go to school every day.

Traffic pollution and industrialization are the root sources of emission of PM_{2.5}. In Malaysia, the main sources of PM_{2.5} are emissions from industrial growth, motor vehicles, and recently, transboundary haze pollution. Traffic-related air pollution (TRAP), happens due to the emissions from motor vehicles.^{1–3} S N Brohi *et al.*⁴ concluded that industries were the main contributors to PM in Malaysia, accounting for 32%. Haze incidents have been a crucial problem in Malaysia for decades. Jaafar *et al.*⁵ concluded that PM_{2.5} is presumed to be one of the most critical health hazards and should be continuously monitored during haze episodes.

Several empirical studies have identified air quality as a major concern in smart cities. The non-linear behaviour of air pollutants, combined with other significant regional factors, results in a highly complex system of air pollutant generation. According to research, capturing nonlinearity between air contaminants and their emission and dispersal sources is difficult in traditional deterministic models. As a result, to address the issue of capturing non-linearity trends in air pollution models and mitigating the impact of PM_{2.5}, ML approaches based on statistical algorithms that are reliable and widely used should be considered.

This literature review is structured as follows. After the justification for the prediction of PM_{2.5}, the steps involved in each ML technique are discussed first. The results and discussion of each ML approach used to predict PM_{2.5} concentrations are presented next. Finally, the conclusions discuss the use of ML for PM_{2.5} prediction.

Methods

The following phases comprise the review of ML approaches. The first step, to find related SCOPUS indexed papers, was to use keyword combinations to find the document; which were: {‘Particulate Matter’} AND {‘PM2.5’} AND {‘Prediction’} AND {‘Machine Learning’}. The paper publishing period was limited to 2017–2021, and the study was limited to journal and conference proceedings that were published in English. We ended up with 284 documents as a result.

After that, the studies were screened by looking at the title and abstract. Biological studies, social studies, and investigations into the relationship between PM_{2.5} and other air contaminants, among other topics, were omitted. The number of documents was reduced to 36. Finally, the papers that were unanimously deemed out of scope were excluded after reviewing the whole document. As a consequence, 20 manuscripts were chosen for further examination.

The five aspects are used to review the articles. The initial analysis is based on the ML type that was utilized. The researchers' method is the second point to consider. Third, the study's location as well as the dataset's characteristics. The fourth and fifth components deal with the evaluation method, performance measures such as root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), coefficient of determination (R²), and the performance of the algorithms under consideration.

Results and discussion

The findings of the review are explained in this section. To begin, the review's findings are based on the number of studies that used machine learning approaches to forecast PM_{2.5} levels. Figure 3 shows the number of conference and journal

articles produced in the last five years to predict $PM_{2.5}$. It shows that the number of studies has been increasing in recent years, and if this trend continues, a potentially higher number of documents might be expected by the end of 2021.

Next, the number of machine learning algorithms has increased significantly; however, this increase has not been evenly distributed globally. Figure 4 shows that the number of published studies in Eurasia and North America is significantly higher. While China (146 studies) and the US (109 studies) have the most published research, Malaysia only has three studies.

According to the study, supervised machine learning techniques like neural networks (NN), deep learning (DL), decision tree (DT), and others were commonly employed. The largest percentages are for DL and NN, followed by DT, regression, and support vector machine (SVM). The following categories are used to categorise the full descriptions of the selected papers.

Category 1: deep learning

Artificial neural networks, which are algorithms inspired by the structure and function of the brain, are used in deep learning. The papers that fall within this category are listed below.

Shahriar *et al.*⁶ conducted a study in three Bangladeshi air pollution hotspots to evaluate the effectiveness of two hybrid models for predicting daily $PM_{2.5}$ concentrations in terms of computational efficiency and accuracy. The models presented and compared with DT and Catboost were Autoregressive Integrated Moving Average (ARIMA)-Artificial Neural Network (ANN) and ARIMA-SVM. With lower MAE and RMSE and a better R^2 value, the CatBoost and ARIMA-ANN models fared well.

By comparing four prediction models, Yang *et al.*⁷ sought to anticipate PM several days in advance in 39 Seoul stations. The gated recurrent Unit (GRU) of a convolutional neural network (CNN) and the long short-term memory (LSTM) of a CNN were the two models suggested and compared. The CNN-LSTM model provided reliable prediction by capturing hidden patterns with low RMSE and MAE values. In Ref. 8 deep neural network (DNN) based hybrid model, DNN-LSTM, was proposed. The DNN-LSTM model outperformed the multiple additive regression trees (MART) and deep feedforward neural network (DFNN) models with a highest R^2 and lowest values of MAE and RMSE for 48-h predictions.

For the prediction of $PM_{2.5}$, Zhang *et al.*⁹ suggested a DL model including an auto-encoder (AE) and a bidirectional LSTM (Bi-LSTM). The proposed method incorporates data preprocessing to improve prediction accuracy, an AE layer to extract implicit features and increase training efficiency, and a Bi-LSTM layer to predict. The results indicate that the proposed model's prediction was better with low RMSE and higher R^2 values and a positive correlation does exist.

Liu *et al.*¹⁰ proposed a hybrid ensemble model using DBN, LSTM, and multilayer perceptron (MLP). To reduce the model's complexity, it uses complementary ensemble empirical mode decomposition (CEEMD) to extract the features from the data series. To produce the best forecast outcomes, the imperial competition algorithm (ICA) is used to alter the weights of the predictors. Two sets of hourly $PM_{2.5}$ concentrations from Shanghai are used to validate the model. According to the findings of the experiments, the proposed model performed better in terms of accuracy and resilience.

Category 2: neural network

The neural network is based on the idea of classifying input observations by linearly combining datasets. For lowering difficulties and errors, Jiang *et al.*¹¹ presented the group teaching optimization algorithm (GTOA): the extreme learning machine (ELM) method, which was based on a data-preprocessing strategy. Two-step decomposition is used to break down PM concentration data into high-frequency IMFs. GTOA is then used to optimise ELM. Over a 16-month period, the data covers hourly $PM_{2.5}$ levels in Beijing. The findings showed that the proposed model improved prediction accuracy while maintaining a low mean absolute error and root mean square error.

With daily records from cities in Finland and Brazil, Neto *et al.*¹² evaluated single and neural-based ensemble techniques. The values of MSE, MAPE, MAE, and RMSE were compared. MLP, the ensemble method, has the best overall performance. Suleiman *et al.*¹³ examined the performance of boosted regression trees (BRT), artificial neural networks (ANN), and support vector machines (SVM) models in forecasting roadside $PM_{2.5}$ concentrations in London at nineteen monitoring sites. For performance evaluation, RMSE and other metrics were used. In general, ANN performed better.

Ali Shah *et al.*¹⁴ assessed the performance of the Phase Space Reconstruction (PSR) technique, which captures multi-time scale information, using radial and linear Support Vector Regression (SVR), Feedforward Neural Network (FFNN), and

Random Forest (RF) ML. RMSE and MAE were used to assess the performance of ML techniques on a dataset collected in Saudi Arabia (Masfalah) over a 21-month period. The results showed that FFNN produced a reliable prediction.

Hung *et al.*¹⁵ conducted research to determine how transported smoke aerosols affect air quality in New York State. ANN was used in the method, and five models with distinct sets of predictors were considered during the summer seasons of 2012–2019. When the models were evaluated using RMSE and R^2 , the results revealed that smoke cases had higher average $PM_{2.5}$ concentrations than non-smoke cases.

Category 3: decision tree

In the decision analysis, a decision tree can be used to visually and explicitly depict decisions and decision making. Angelin Jebamalar *et al.*¹⁶ created a method that combined light gradient boosting and decision tree techniques. The model breaks the tree leaf by leaf using the best fit, allowing it to handle massive amounts of data with little memory and great speed. From 2017 to 2019, the $PM_{2.5}$ information for Chennai, India, was collected using IoT devices and stored in the cloud. When compared to RF, DT, and regression approaches, the suggested model outperformed with the lowest mean absolute error and root mean square error values.

Another study¹⁷ focused on $PM_{2.5}$ predictions from environmental sensor data streams using a stacked boosting ensemble (STBoost) model with z-score optimization techniques. The STBoost includes Light GBM as the meta regressor and base regressors such as XGBoost and GBM regression and RF to improve the prediction accuracy. The results indicate that the STBoost model outperformed with an accuracy score of 99.52 and an RMSE value of 0.1048.

To improve $PM_{2.5}$ perception, Luo *et al.*¹⁸ used an image-based technique that included CNN and GBM. Daily weather conditions, 6976 pictures, and hourly data from Shanghai were used to create the model (2016). With the proposed technique, the MAE, RMSE, and R^2 estimations of $PM_{2.5}$ are 3.56, 10.02, and 0.85, respectively.

Using a 1.5 years' dataset of Malaysia,¹⁹ examined the performance of MLP and RF models in predicting $PM_{2.5}$. Confusion matrix was utilised as a performance metric. MLP was outperformed by RF overall.

Category 4: regression

The regression model was used in a number of articles to predict particulate matter. Kleine Deters *et al.*²⁰ present a machine learning technique for predicting $PM_{2.5}$ based on pollution and meteorological data from two Ecuadorian cities over a six-year period. Using CGM in regression analysis, it was found that $PM_{2.5}$ could be predicted more accurately during extreme weather.

Kowalski and Warchałowski²¹ provide a comparison of machine learning techniques for predicting dust-type air pollution levels. Real-time hourly data from Krakow was utilised to train and test the models using MSE and R^2 over the course of a year. The best prediction approach, according to the findings, was a regression model.

Gu *et al.*²² introduced a recurrent air quality prediction (RAQP) model, which combines a recurrent framework and SVR. The model was tested on 180 hourly records from a small Chinese town. The RAQP model was found to be more successful than state-of-the-art air quality predictors with a low RMSE value.

Aljuaid *et al.*²³ compared numerous forecasting approaches and strategies based on mathematics and machine learning. The one-hour and five-minute datasets were created by combining and manipulating numerous sources of Danish data. For comparison, the researchers utilised multivariate (SVR, DT, and K-nearest neighbour) and univariate (auto-regression) techniques. SVR had the lowest RMSE and MAE values for the one-hour data set, and auto-regression for the five-minutes dataset, according to the findings.

Category 5: support vector machine

For $PM_{2.5}$ ground-level forecasting in the city of Bogotá, Mogollón-Sotelo *et al.*²⁴ presented SVM. Statistical validation was done using RMSE, *etc.* The SVM model predicts with greater accuracy. An ensemble empirical mode decomposition (EEMD), least square SVM (LSSVM), and PSR were proposed in Ref. 25 as alternative method for forecasting the following day. The empirical results reveal that the EEMD-PSR-LSSVM outperformed other models in terms of MAPE and RMSE values.

Conclusions

This study looked over 20 scientific papers that focused on machine learning algorithms for $PM_{2.5}$ prediction. The use of machine learning to forecast $PM_{2.5}$ has grown significantly in the previous five years, although only three research articles

have been published in Malaysia. There are also several international research programmes that concentrate on more than one type of particulate matter. For predicting PM_{2.5}, the ML techniques DL, NN, and DT are often utilised. Overall, it appears that supervised machine learning algorithms are employed to predict air pollution, notably PM_{2.5}. The review concludes that while there have been few researches in Malaysia using machine learning to forecast PM_{2.5}, the field can be expanded and the accuracy improved, as has been done globally using supervised machine learning methodologies.

Data availability

No data are associated with this article.

Author contributions

Palanichamy Naveen did the conception of the work, drafting the article, and revision to the final version. Kuhaneswaran and Rishanti did data collection, data analysis and interpretation, under the guidance of S Subramanian and their supervisors Su-Cheng Haw and Palanichamy Naveen. Palanichamy Naveen is the corresponding author for this paper.

Acknowledgements

Not applicable.

References

- Awang MF, Jalaludin J, Latif MT, *et al.*: **Exposure to PM_{2.5} in urban area and respiratory health symptoms among urban workers in Klang Valley**. *IOP Conference Series: Earth and Environmental Science*. IOP Publishing; 2019; Vol. **228**(No. 1): p. 012015.
- Usmani RSA, Saeed A, Abdullahi AM, *et al.*: **Air pollution and its health impacts in Malaysia: a review**. *Air Qual. Atmos. Health*. 2020; **13**(9): 1093–1118.
[Publisher Full Text](#)
- Ameer S, Shah MA, Khan A, *et al.*: **Comparative analysis of machine learning techniques for predicting air quality in smart cities**. *IEEE Access*. 2019; **7**: 128325–128338.
[Publisher Full Text](#)
- Brohi SN, Pillai TR, Asirvatham D, *et al.*: **Towards smart cities development: a study of public transport system and traffic-related air pollutants in Malaysia**. *IOP Conference Series: Earth and Environmental Science*. IOP Publishing; 2018, June; Vol. **167**(No. 1): p. 012015.
- Jaafar SA, Latif MT, Razak IS, *et al.*: **Composition of carbohydrates, surfactants, major elements and anions in PM_{2.5} during the 2013 Southeast Asia high pollution episode in Malaysia**. *Particuology*. 2018; **37**: 119–126.
[Publisher Full Text](#)
- Shahriar SA, Kayes I, Hasan K, *et al.*: **Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for Atmospheric PM_{2.5} Forecasting in Bangladesh**. *Atmos*. 2021; **12**(1): 100.
[Publisher Full Text](#)
- Yang G, Lee H, Lee G: **A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea**. *Atmos*. 2020; **11**(4): 348.
[Publisher Full Text](#)
- Karimian H, Li Q, Wu C, *et al.*: **Evaluation of different machine learning approaches to forecasting PM_{2.5} mass concentrations**. *Aerosol Air Qual. Res*. 2019; **19**(6): 1400–1410.
[Publisher Full Text](#)
- Zhang B, Zhang H, Zhao G, *et al.*: **Constructing a PM_{2.5} concentration prediction model by combining auto-encoder with Bi-LSTM neural networks**. *Environ. Model Softw*. 2020; **124**: 104600.
[Publisher Full Text](#)
- Liu H, Dong S: **A novel hybrid ensemble model for hourly PM_{2.5} forecasting using multiple neural networks: a case study in China**. *Air Qual. Atmos. Health*. 2020; **13**(12): 1411–1420.
[Publisher Full Text](#)
- Jiang F, Qiao Y, Jiang X, *et al.*: **MultiStep Ahead Forecasting for Hourly PM₁₀ and PM_{2.5} Based on Two-Stage Decomposition Embedded Sample Entropy and Group Teacher Optimization Algorithm**. *Atmos*. 2021; **12**(1): 64.
[Publisher Full Text](#)
- Neto PSDM, Firmino PRA, Siqueira H, *et al.*: **Neural-Based Ensembles for Particulate Matter Forecasting**. *IEEE Access*. 2021; **9**: 14470–14490.
[Publisher Full Text](#)
- Suleiman A, Tight MR, Quinn AD: **Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM₁₀ and PM_{2.5})**. *Atmos. Pollut. Res*. 2019; **10**(1): 134–144.
[Publisher Full Text](#)
- Ali Shah SA, Aziz W, Ahmed Nadeem MS, *et al.*: **A novel phase space reconstruction (PSR-) based predictive algorithm to forecast atmospheric particulate matter concentration**. *Sci. Program*. 2019; **2019**.
[Publisher Full Text](#)
- Hung WT, Lu CHS, Alessandrini S, *et al.*: **The impacts of transported wildfire smoke aerosols on surface air quality in New York State: A multi-year study using machine learning**. *Atmos. Environ*. 2021; **259**: 118513.
[Publisher Full Text](#)
- Angelin Jebamalar J, Sasi Kumar A: **PM_{2.5} prediction using machine learning hybrid model for smart health**. *Int. J. Eng. Adv. Technol*. 2019; **9**(1): 6500–6504.
- Jebamalar JA, Kamalakannan T: **Enhanced Stacking Ensemble Model in Predictive Analytics of Environmental Sensor Data**. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE; 2021, March; (pp. 482–486).
- Luo Z, Huang F, Liu H: **PM_{2.5} concentration estimation using convolutional neuranetwork and gradient boosting machine**. *J. Environ. Sci*. 2020; **98**: 85–93.
[Publisher Full Text](#)
- Murugan R, Palanichamy N: **Smart City Air Quality Prediction using Machine Learning**. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE; 2021, May; (pp. 1048–1054).
- Kleine Deters J, Zalakeviciute R, Gonzalez M, *et al.*: **Modeling PM_{2.5} urban pollution using machine learning and selected meteorological parameters**. *Int. J. Electr. Comput. Eng*. 2017; **2017**.
- Kowalski PA, Warchałowski W: **The comparison of linear models for PM₁₀ and PM_{2.5} forecasting**. *WIT Trans. Ecol. Environ*. 2018; **230**: 177–188.
- Gu K, Qiao J, Lin W: **Recurrent air quality predictor based on meteorology and pollution-related factors**. *IEEE Transactions on Industrial Informatics*. 2018; **14**(9): 3946–3955.
[Publisher Full Text](#)
- Aljuaid H, Alwabel N: **Air pollution prediction using machine learning algorithms**. *Int. J. Eng. Adv. Technol*. 2019; **8**(6 Special Issue 3): 160–164.
- Mogollón-Sotelo C, Casallas A, Vidal S, *et al.*: **A support vector machine model to forecast ground-level PM 2.5 in a highly populated city with a complex terrain**. *Air Qual. Atmos. Health*.

- 2021; **14**(3): 399–409.
[Publisher Full Text](#)
25. Niu M, Gan K, Sun S, *et al.*: **Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead PM2.5 concentration forecasting.** *J. Environ. Manag.* 2017; **196**: 110–118.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 24 January 2022

<https://doi.org/10.5256/f1000research.76794.r115274>

© 2022 Devaraj D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

D. Devaraj

Department of Electrical and Electronics Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu, India

Globally, in the major cities, air pollution is becoming a major issue. To adopt appropriate pollution control strategies Prediction of PM_{2.5} concentration is necessary. Recently, in the literature, machine learning algorithms like Decision tree algorithms, Artificial neural networks, and support vector machines have been applied for accurate prediction of PM_{2.5} concentration. In this review paper, the authors have reviewed recently published articles related to the prediction of PM_{2.5} using Machine learning algorithms. In this review paper, the authors have highlighted the strength and weaknesses of various algorithms reported in the literature. The reviewer has the following suggestions to further improve the quality of this paper:

- Include the review of "Deep learning" at the end after reviewing the papers related to the conventional machine learning algorithms, instead of reviewing that at the beginning.
- Include the research gap identified in this area at the end of the review.
- Also, the title of the section "Results and Discussion" may be changed appropriately.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning, Smart grid

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 06 January 2022

<https://doi.org/10.5256/f1000research.76794.r115273>

© 2022 Aneja N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nagender Aneja

Digital Science, Faculty of Science, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei

This manuscript examines and highlights the most available ML strategies for PM2.5 prediction to raise knowledge of ML techniques and encourage additional research in this field. The articles published in Malaysia proposed supervised machine learning algorithms to predict air pollution, notably PM2.5.

The articles have been categorized in deep learning, neural network, decision tree, regression, and support vector machine. This is a good analysis from the perspective of Malaysia, however, the search approach seems limited and missing some references e.g. Masood and Ahmad (2020)¹, Danesh *et al.*, (2020)², Doreswamy *et al.*, (2020)³. Authors can include similar references and see if the techniques are applicable for Malaysia also.

References

1. Masood A, Ahmad K: A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. *Procedia Computer Science*. 2020; **167**: 2101-2110 [Publisher Full Text](#)
2. Danesh Yazdi M, Kuang Z, Dimakopoulou K, Barratt B, et al.: Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. *Remote Sensing*. 2020; **12** (6). [Publisher Full Text](#)
3. Doreswamy, K S H, KM Y, Gad I: Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. *Procedia Computer Science*. 2020; **171**: 2057-2066 [Publisher Full Text](#)

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep Learning, NLP

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research