Random Forest Feature Selection for PM10 Pollution Concentration

Habeeb Balogun[1], Hafiz Alaka[1], Christian Egwim[1] and Ajayi Saheed[2]

[1] *Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom.*

[2] *School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, LS2 8AG, United Kingdom.*

Abstract

There are already countless articles on strategies to limit human exposure to PM10 pollutants because of their disastrous impact on the environment and people's well-being in the United Kingdom (UK) and around the globe. Strategies such as imposing sanctions on places with higher levels of exposure, dissuading non-environmentally friendly vehicles, motivating the use of bicycles for transportation, and encouraging the use of eco-friendly fuels in industries. All these methods are viable options but will take longer to implement. For this, efficient PM10 predictive machine learning is needed with the most impactful features/data. The predictive model will offer more strategic avoidance techniques to this lethal air pollutant, in addition to all other current efforts. However, the diversity of the existing data is a challenge. This paper tries to solve this by bringing together traffic information, pollution concentration information, geographical/built environment information, and meteorological information. Furthermore, this paper applied random forest, which outperformed the decision tree and XGBoost in selecting the most impactful features. As part of the discovery from this research work, it is now clearly discovered that the height of buildings in a geographical area has a role to play in the dispersion of PM10.

Keywords: Air pollution, Feature selection, Machine Learning, PM10, Random Forest

1 INTRODUCTION

The United Kingdom (UK) and other parts of the world have experienced fast growth, development, and industrialization in recent years, causing severe air pollution in most major cities. (DEFRA, 2018; DfT & DEFRA, 2017). Atmospheric pollution harms humans, the environment and ecological systems. (Fortelli et al., 2016). It causes eye pain, throat irritation, lung cancer, and damage to vital parts of the human body such as the brain, kidneys, and heart. (Abdul Halim et al., 2018; WHO Regional Office for Europe OECD, 2015). These illnesses and other consequences of human exposure to poor air quality are linked to the over 40,000 premature deaths in the UK each year. (Public Health England, 2019) and the over 8 million deaths worldwide (Nethery & Dominici, 2019). Aside from the negative health effects, air pollution costs the UK and the rest of the world a fortune to manage, negatively impacting the UK economy (Bai et al., 2018) and the world's economy (Myllyvirta, 2020).

Sulphur dioxide (SO2), nitrogen dioxide (NO2), particulate matter (PM2.5 and PM10), and carbon dioxide (CO2) are among the killer air pollutants (Bai et al., 2018). Because of its composition and size, PM10, the inhalable particulate matter composed of sulphates and other chemical elements/compounds with a diameter of less than 10 microns, remains one of the deadliest pollutants. PM10 exists in the atmosphere as a mixture of solid and liquid granular elements, and It absorbed by the human body (WHO Regional Office for Europe OECD, 2015). PM10 is from various places, including buildings, roads, agricultural operations, wind erosion, vehicular and industrial emissions. (Sánchez et al., 2020). Because of the adverse health effects of PM10, research on its health impact has increased significantly; for example, PM10 is carcinogenic (Dehghan et al., 2018). The harmful effects of PM10 on health and the environment necessitate the urgent need for researchers to predict/forecast the concentration of the deadly pollutant. These forecasts will enable people to make informed decisions about whether or not to visit areas with higher concentrations of this pollutant (Bai et al., 2018).

In recent years, there has been a great deal of research into PM10 pollutant prediction models. This research interest is due to the number of benefits associated with the machine learning (ML)-based predictive models. However, these ML models gradually improved through mathematical, statistical, and computational techniques, such as land-use regression (Shahraiyni & Sodoudi, 2016) has remain not practically efficient due to dependent on the features/data used in developing such model.

A variety of data sources influences PM10 air pollution: solid and gas particles suspended in the air, artificially-made (e.g., vehicle, industry, agricultural and other chemical involved activities), meteorological (e.g., Temperature, humidity, pressure, wind, rainfall, among others), geographical (e.g., happenings around certain regions, for instance, market or industrial or education among others), Built environment (e.g., building height, building size, and building type). However, because of the variety of PM10 sources and available data, obtaining the most efficient and effective ML is difficult. Therefore, in this research, we combine PM10 sources/features and look at specific features that can help develop effective predictive machine learning models. Overall, the objective for this paper is as follows:

i. To preprocess a reasonably large data of PM10 from IoT sensors with time-corresponding weather, built environment information, and traffic data

ii. To investigate features that impact the prediction of PM10 using gradient boosting feature selection method

The remaining of this paper is arranged as thus. Section two briefly discussed the feature engineering methods that exist, and this forms the literature review for the research. Section three describes the mathematical and technical knowledge behind the implementation of the random forest feature selection method. Section four discusses how the massive data set obtained from the IoT emission sensor installed across 14 different UK locations was merged with traffic information from Tom-tom, meteorological data from the Open weather, and building property data from the Digi Map. Also, the result of the RF feature selection and features selected are discussed in this section. Finally, Section five concludes the study by offering suggestions for future research.

2 LITERATURE REVIEW

The dimensionality of features influences many powerful machine learning algorithms (Hafiz et al., 2015). Reduced dimensionality, selecting the most impactful features from the original set of features as the new input features, has been shown to improve the performance of prediction models (Balogun, Alaka and Egwim 2021).

Feature selection (FS) is a crucial stage in implementing machine learning algorithms in several areas; it decreases the number of features to a bare minimum, allowing for a more efficient and less computational cost model (Guyon and Elisseef, 2003). FS is divided into three categories: filter, wrapper, and embedded method. See figure 1 for the classification

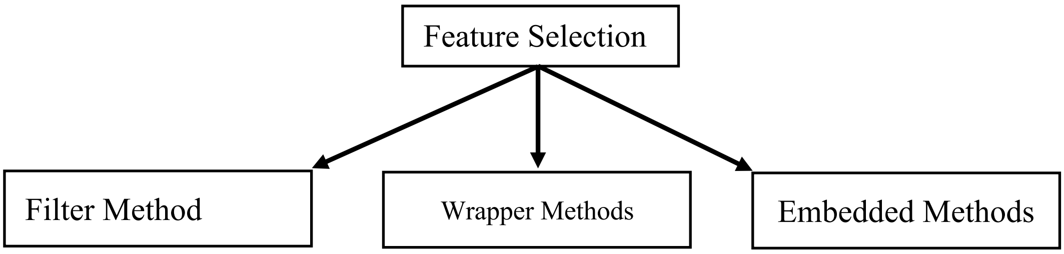

Figure 1 Feature Selection Classification

The filter technique selects features based on the dataset's statistical properties, delivering feature ranking as an output, regardless of the model. The research by (Jović et al., 2015) shows an example of various standard filter methods. Though filter methods are simple to use because of their low computational cost, the wrapper approach is superior due to its search strategy based on a modelling algorithm (Jović et al., 2015). In addition, wrapper approaches are devised to analyze specific feature subsets with learning algorithms.

Embedded FS methods, unlike filter and wrapper methods, choose subset features during the algorithm modelling implementation. This method combines the benefits of the filter and wrapper methods in terms of performance and computational cost. They carefully maintain each iteration of the model training process and extract features that contribute the most to training for specific iterations. Out of the many embedded FS that exist, random forest, a select from model method that uses estimators with important features function, is used in this research. Other most frequently used embedded FS is the Lasso and the Elastic Net.

For the benefits associated with the embedded Feature selection method, this paper investigates relevant features impactful in the development of the PM10 predictive model from the diverse set of data collected and preprocessed.

3 METHODS

Input Data and Preprocessing

For five months, the sensors recorded PM10 concentration data every ten seconds (i.e., December 2019 and April 2020). Over eight billion observations were collected for this period, which was tremendous, necessitating the use of a big data platform. This study uses the Amazon web service (AWS) cloud/Big data platform for implementation and data exchange. We developed a middleware framework used as a data transit platform and deployed it on an Elastic Computing Cloud two (EC2) instance. The middleware is then connected to an AWS relational database, where the massive data is stored. Because of the large amount of data in this study, we employed the AWS EC2 infrastructure to execute the big data analytics. In addition, the traffic, weather and building information/feature covering the same time and location as the data from the sensors were retrieved through application program interface (API) calls. The weather data from Open Weather, traffic data from Tom-tom, building features from Digi Map matching approximate locations where sensors are installed for a similar time was united, stored, and preprocessed. Weather and traffic information was provided hourly and then matched with PM10 concentration data from an IoT sensor, yielding (24hrs x 30days x 5months x 14 IoTs) data points. Table 1 shows all the features united, description and their sources.

Table 1 Features united from the different sources

S/N	Features	Source	Description
1	Ambient Humidity	open weather's API	Weather information measured and recorded at a particular time, usually hourly.
2	Ambient Pressure		
3	Ambient Temp		
4	Humidity		
5	Temp		
6	Road Type	Tom-tom's API	Type is either a Major or a Minor Road
7	Speed		The recorded speed for all the vehicles
8	Road area	DIGI geospatial Map	Total road/footpath area around the sensor
9	Low-rise building		The count of the different classes of the buildings in terms of floor and height (He et al., 2019)
10	Multi-story building		
11	Middle-rise building		
12	Small high-rise building		
13	High-rise building		
14	Ultra-high rise building		

15	Green area		The proximity of the green area to the sensor
16	hour	IoT Emission Sensor	The hour the sensor retrieves data
17	weekday		The day of the week, data is retrieved
18	holiday		Holiday or not on the day data was retrieved
19	Longitude		Longitude of the installed sensor
20	Latitude		Latitude of the installed sensor
21	Location name		The Name of a point at which the sensor was installed
22	Sensor		The ID assigned to the sensor
28	Pm10		The concentration of PM10 retrieved by the sensor

As part of the preprocessing, we dig into exploration of the united data. For instance, the exact location all the 14 installed sensor is presented in Figure 2.

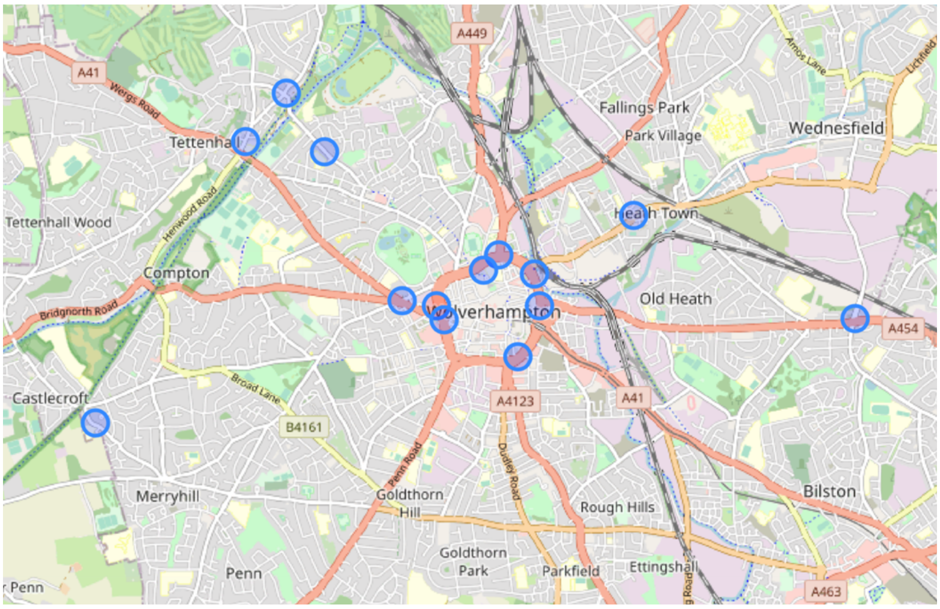

Figure 2 Map Showing the 14 PM10 IoT Monitoring sensors deployed at the UK

It was discovered, at around the 14 installed sensors, most of the buildings are not beyond six floors, leaving the data for mid-rise (MRB), Small high rise (SHRB), High rise (HRB), and ultra-high(UHRB) buildings empty. Thus, upon computing correlation analysis, the whole of the building features (MRB, SHRB, HRB, UHRB) did not correlate with the PM10 and was dropped. As for the once, with some degree of association with the PM10 concentration, Figure 3 shows what the correlation looks like.

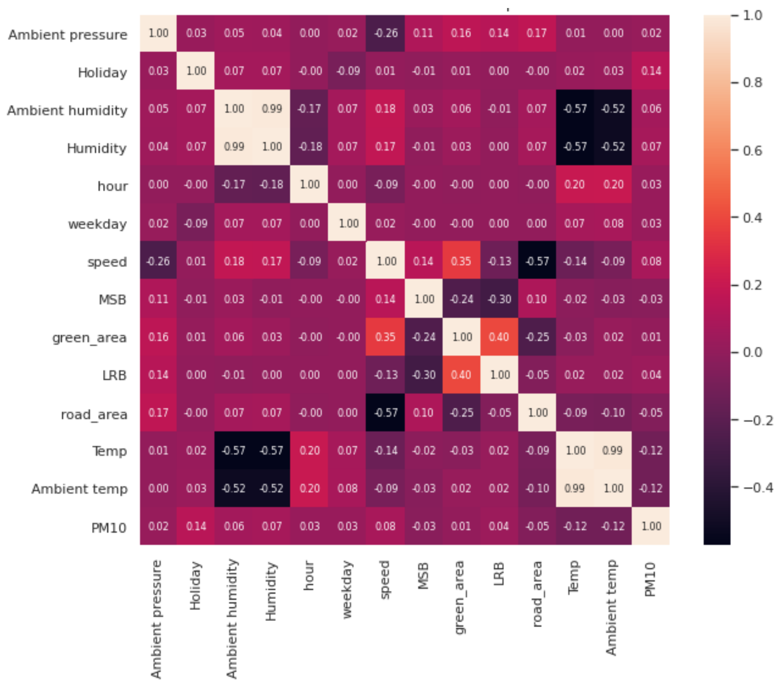

Figure 3 Feature correlation Heatmap

Figure 2, we can note there is a minimal relationship between virtually all the features united in this study and the PM10 concentration. This correlation, a filter FS, may misguide our judgement in selecting subsets impactful in developing an efficient predictive Model for PM10 pollutants. This correlation is what necessitates the use of Random Forest embedded FS.

Random Forest FS

RF is one of the most effective supervised learning algorithms easy to implement. The efficiency lies around the number of trees built. However, it can be used in the prediction of PM10. RF can also be used in selecting relevant subset features using the select from models. This select from model determines the threshold, assign boundary between the features to be selected and the redundant features that are dropped in the long run, then sort all input features by Gini importance score in ascending order. Further, features with lower Gini importance scores below the threshold are eliminated. The selected features (i.e., features with importance score above threshold) will serve as the new subset features selected in the development of random forest algorithm; see the implementation of the RF-FS method.

```
Input A training set S:$(x_1, y_1)$…,$(x_n, y_n)$, $F$ features and number of trees in Forest $B$

    1. Select M trees from the dataset
    2. Construct a decision tree from the M tree
    3. Repeat step 1 and step 2, B times.
    4. At each node:
    5.            Construct f as a tiny subset of F
    6.            Split on best feature in f
    7. New records are given to the category that wins the most votes

Results: D selected features that have the highest accuracy
```

4 DISCUSSIONS OF RESULT

Performance Metrics

In regression, the method of performance can be measure using the mean absolute error (MAE), mean square error (MSE), and R Squared(R^2). These metrics briefly explained were used in this paper. MAE is the average absolute error between each actual dependent feature value and predicted dependent feature value. The MAE is.

$$\text{MAE} = {}^{1}/_{n} \sum_{i}^{n} |y_i - y_i^*| \tag{1}$$

R-squared is the coefficient of determination indicating the goodness of fit. R-square is.

$$R^2 = 1 - \frac{\sum_{i=0}^{n} \left(y_i - y_i^*\right)^2}{\sum_{i=0}^{n} (y_i - \bar{y}_i)^2} \tag{2}$$

The EVS score measures the variation (a measure of dispersion) of the test data set. The best possible score of EVS is 1.0, and is.

$$\text{EVS} = 1 - \frac{y_i - y_i^*}{y_i} \tag{3}$$

Where $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$

Experiment

This study uses others to select from the Model Feature selection method, Decision tree and gradient boosting. Each FS was evaluated, presented in table 2.

Table 2 Feature selection methods implemented and their performance metrics

FS/Metrics	*MAE*	*R-square*	*EVS*
Random Forest	*2.8406*	*0.5794*	*0.5964*
Decision Tree	*3.6685*	*0.4476*	*0.4476*
XGBoost	*4.5896.*	*0.2125*	*0.3374*

As discovered from the analysis in table 2, RF performed better than Decision and the XGBoost feature selection method. Scoring MAE of 2.8406 cycles, i.e., the model selected impactful features for prediction of PM10 within the average mean absolute error range of ±2.8406 cycles.

Since RF selects the best set of subset features in the prediction of PM10. As part of the advantages associated with the select from the model, we present the rank of the features by importance (see figure 4)

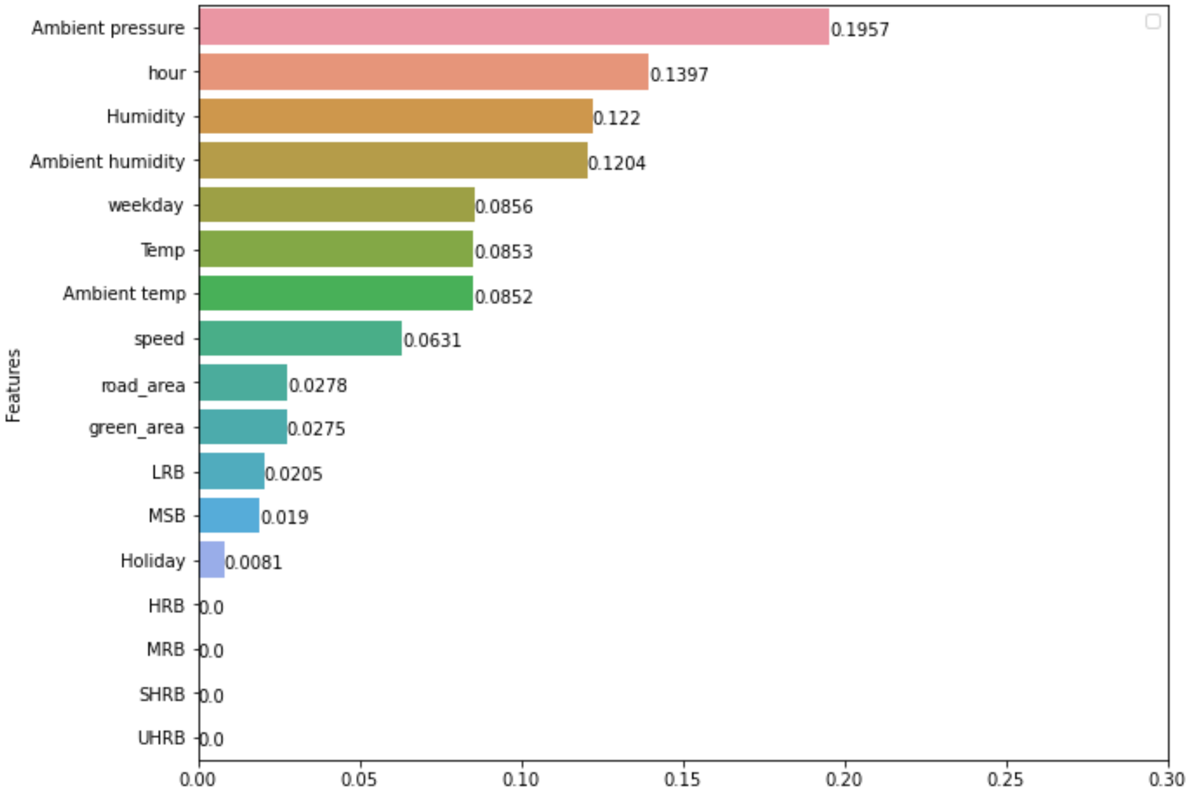

Figure 4 Visualizing Feature Importance by ranks

From the analysis, we can see that Ambient pressure is the most impactful feature in developing the PM10 predictive model. This is followed by the hour at which the PM10 concentration data is retrieved. This somehow makes sense, as the movement and dispersion of the pollutant may differ looking at the different hours of the day, say, for instance, the pollutant at rush hours when many are heading out for work, school and for other reasons will not be same for when people are not out for these reasons.

Other essential features ranked after an hour and relative pressure include the Humidity and relative humidity, weekdays, Temperature, Ambient Temperature, speed recorded, road area, green area, low rise and middle rise building around the sensor, and lastly, the sensor the holiday.

Inclusive of the dependent feature (i.e., PM10 concentration data), a total of 14 out of the 28 features united for this research were discovered impactful in the development of the PM10 predictive model.

5 CONCLUSIONS

Machine learning (ML)-based PM10 pollution prediction has several benefits, including counselling persons sensitive to air pollution. However, the numerous sources of data that influence this deadly air pollutant to reduce the efficiency of the many developed models. Most models predict using just the pollutant data, and some use weather data as independent, some use traffic data as independent considering vehicle as a significant source. The diversity of the existing data, which is a challenge, is what this paper tries to solve, bringing together traffic information, pollution concentration information, geographical/built environment information, meteorological information. The union of the many features/data will enable a more practical and efficient PM10 machine learning model that can accurately guide user decisions to air pollution, including the recent covid19 patients. In addition to data unity, we investigated specific features that can be used to develop efficient PM10 pollution ML models. The following set of inferences can bes formed after the study:

i. Random forest, a feature selection technique, outperforms decision tree and XGBoost Feature selection method
ii. The height and size of building around a geographical location contributes to the dispersion of PM10 pollutant
iii. Features such as Ambient pressure, hour, Humidity, relative humidity, weekdays, Temperature, Ambient Temperature, speed recorded, road area, green area, count of low rise and middle rise building around the sensor and lastly, the holiday are the most impactful features used in the development of PM10 predictive model.

Future studies should explore other feature selection methods and possibly compare the performance.

Reference:

Abdul Halim, N. D., Latif, M. T., Ahamad, F., Dominick, D., Chung, J. X., Juneng, L., & Khan, M. F. (2018). The long-term assessment of air quality on an island in Malaysia. *Heliyon*, *4*(12). https://doi.org/10.1016/j.heliyon.2018.e01054

Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, *15*(4), 1–44. https://doi.org/10.3390/ijerph15040780

Balogun, H., Alaka, H., & Egwim, C. (2021). Boruta-Grid-Search Least Square Support Vector Machine for NO2 Pollution Prediction Using Big Data Analytics and IoT Emission Sensors. *Applied Computing and Informatics*.

Chae, S., Shin, J., Kwon, S., Lee, S., Kang, S., & Lee, D. (2021). PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. *Scientific Reports*, *11*(1), 11952. https://doi.org/10.1038/s41598-021-91253-9

DEFRA. (2018). *Supplement to the UK plan for tackling roadside nitrogen dioxide concentrations*. *October*, 1–54. https://doi.org/10.1016/S0031-9406(07)60001-7

Dehghan, A., Khanjani, N., Bahrampour, A., Goudarzi, G., & Yunesian, M. (2018). The relation between air pollution and respiratory deaths in Tehran, Iran- using generalized additive models. *BMC Pulmonary Medicine*, *18*(1), 1–9. https://doi.org/10.1186/s12890-018-0613-9

DfT, & DEFRA. (2017). UK plan for tackling roadside nitrogen dioxide concentrations: Detailed plan. *Department for Environment Food & Rural Affairs Together with Department for Transport.*, *July*, 1–11.

Fortelli, A., Scafetta, N., & Mazzarella, A. (2016). Influence of synoptic and local atmospheric patterns on PM10 air pollution levels: a model application to Naples (Italy). *Atmospheric Environment*, *143*, 218–228. https://doi.org/10.1016/j.atmosenv.2016.08.050

Guyon, i., Elisseef, A. (2003). An introduction to variable and feature selection. *Machine Learning Research*, *3*, 1157–1182.

Hafiz, A., Lukumon, O., Muhammad, B., Olugbenga, A., Hakeem, O., & Saheed, A. (2015). Bankruptcy prediction of construction businesses: Towards a big data analytics approach. *Proceedings - 2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015*, 347–352. https://doi.org/10.1109/BigDataService.2015.30

He, S., Wang, X., Dong, J., Wei, B., Duan, H., Jiao, J., & Xie, Y. (2019). Threedimensional urban expansion analysis of valley-type cities: A case study of Chengguan District, Lanzhou, China. *Sustainability (Switzerland)*, *11*(20). https://doi.org/10.3390/su11205663

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Myllyvirta, L. (2020). *Quantifying the Economic Costs of Air Pollution from Fossil Fuels Key messages*. 2–13.

Nethery, R. C., & Dominici, F. (2019). Estimating pollution-attributable mortality at the regional and global scales: Challenges in uncertainty estimation and causal inference. In *European Heart Journal* (Vol. 40, Issue 20, pp. 1597–1599). Oxford University Press. https://doi.org/10.1093/eurheartj/ehz200

Public Health England. (2019). *Review of interventions to improve outdoor air quality and public health*.

Sánchez Lasheras, F., García Nieto, P. J., García Gonzalo, E., Bonavera, L., & de Cos Juez, F. J. (2020). Evolution and forecasting of PM10 concentration at the Port of Gijon (Spain). *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-68636-5

Shahraiyni, H. T., & Sodoudi, S. (2016). Statistical modeling approaches for pm10 prediction in urban areas; A review of 21st-century studies. *Atmosphere*, *7*(2), 10–13. https://doi.org/10.3390/atmos7020015

WHO Regional Office for Europe OECD. (2015). Economic cost of the health impact of air pollution in Europe: Clean air, health and wealth. *European Environment and Health Processes*, 1–54.

Wu, Z., Wu, X., Wang, Y., & He, S. (2020). PM2.5ĝ•PM10 ratio prediction based on a long short-Term memory neural network in Wuhan, China. *Geoscientific Model Development*, *13*(3), 1499–1511. https://doi.org/10.5194/gmd-13-1499-2020