An Application Of Machine Learning With Boruta Feature Selection To Improve NO2 Pollution Prediction

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

Projecting and monitoring NO2 pollutants' concentration is perhaps an efficient and effective technique to lower people's exposure, reducing the negative impact caused by this harmful atmospheric substance. Many studies have been proposed to predict NO2 Machine learning (ML) algorithm using a diverse set of data, making the efficiency of such a model dependent on the data/feature used. This research installed and used data from 14 Internet of thing (IoT) emission sensors, combined with weather data from the UK meteorology department and traffic data from the department for transport for the corresponding time and location where the pollution sensors exist. This paper select relevant features from the united data/feature set using Boruta Algorithm. Six out of the many features were identified as valuable features in the NO2 ML model development. The identified features are Ambient humidity, Ambient pressure, Ambient temperature, Days of the week, two-wheeled vehicles(counts), cars/taxis(counts). These six features were used to develop different ML models compared with the same ML model developed using all united data/features. For most ML models implemented, there was a performance improvement when developed using the features selected with Boruta Algorithm.

Keywords: Boruta, Feature Selection, Machine Learning, NO2 Pollution, Prediction

1 INTRODUCTION

Air pollution is a principal causal agent for increased mortality, making air pollution a significant health concern in all developing/developed nations. For instance, air pollution is linked to over 25,000 deaths yearly in the UK (Public Health England, 2019) and over 8million deaths globally (Nethery & Dominici, 2019). In addition to death and other health issues caused by air pollution, the economic challenges caused by air pollution cannot be ignored. For example, the UK government spends approximately £40bn (WHO Regional Office for Europe OECD, 2015). Exciting research by the centre for research on energy and clean air (CREA) associates with over 1.5 billion absent from workdays, around 3.5million new asthma cases and around 1.5million preterm births to air pollution. These negative impacts are responsible for the increase in health care cost and decrease in economic productivity.

Air pollutants are toxic substance released into the air and are typical of two groupings: particulate matter and gases. Of the gases, Nitrogen dioxide (NO2) is debatably the most threatening to human existence as it has been recognized as a trigger for asthma (Kopparapu et al., 2021). Unfortunately, in the UK and most developing/developed nations, exposure to NO2 is inevitably impossible for people, and this is because they may find it difficult to change their habits. However, a strategic way in reducing exposure to this harmful substance will be to provide an explanatory system that informs people of the pollution level in a particular geographical location.

There is a usual understanding among researchers that NO2 originates from combustion processes such as fossil fuels from vehicles and industrial activities. However, the origin of NO2 can be distinctive from region to region. For example, the world noticed over 15% drop in global NO2 concentration during the Covid-19 pandemic.

Monitoring and predicting NO2 concentration have been studied for a long time, starting from 2001 (Kolehmainen et al., 2001). Several methodologies have been in existence and are generally simulation-based or data-driven machine learning-based. The simulation-based methods use professional knowledge and are primarily dependent on the Model's assumption. However, most of the simulated models' efficiency depends on human factors like the developer's experience, which may be imprecise and inaccurate.

A machine learning-built predictive model derived from statistics, applied mathematics widely used in computing and algorithms development. Machine learning models help solve regression or classification problems. Classification problems involve predicting from two or more classes, while regression intends to predict exact continuous value. In this research, predicting NO2 pollution is a regression problem due to the continuous nature of NO2 pollution data.

Studies on NO2 prediction models have thus justifiably increased since the turn of the millennium. However, if they are helpful to users vulnerable to pollution, e.g., coronavirus patients, the effectiveness of such models depends on the Model's performance. Lesser performance can be misguiding and could expose the user to a pollution hotspot, triggering life-threatening attacks.

Machine learning (ML) offers many advantages for the prediction of NO2 from data. Many such ML methods have been proposed for the prediction of NO2. However, the diversity of the existing data make it a challenge to get the most efficient and effective ML. There are multiple sources of data that influence NO2 air pollution: the pollution data, origin of NO2 pollution (e.g., vehicle, industry, other chemical involved activities), weather data (e.g., temperature, wind, rainfall, among others), geographical data (e.g., geolocation, activities around certain areas, for instance, trading

or manufacturing or schooling among others), Built environment (e.g., building height, building size, and building type). All these data have been proven to contribute to the dispersion of NO2 pollution. This paper establishes how a range of pollution-related data sets can be united and investigated to know exact features helpful in developing efficient NO2 pollution ML models.
In brief, this paper seeks to select relevant features helpful in developing an accurate ML model to predict NO2 pollution. The objectives are as follows:

(i) To unite enormous NO2 pollution data from IoT-emission-sensors with time-corresponding weather and traffic data
ii) To investigate the data, selecting valuable/relevant features using the Boruta Feature selection method
iii) To develop NO2 pollution ML models using features selected with and without Boruta

The rest of this paper is structured as follows. Section II provides a brief discussion on the related works. Section III focuses on the methods, including data sourcing, preparation, and pre-processing for analytics. Section IV presents a discussion. Finally, section V focuses on conclusions and further research opportunities.

2 LITERATURE REVIEW

The predictive capability of various machine learning depends on the features' dimensionality (Hafiz et al., 2015). Dimensionality reduction has been proven to help make predictive models perform better (Reddy et al., 2020). Of the reduction techniques, feature selection is selecting the most impactful features from the original set of features as the new input features. Thus, not all features impact the prediction, making feature/variable selection critical in developing/building machine learning predictive models.

Feature selection (FS) is a significant step in implementing machine learning algorithms across different domains; it aids in reducing the features to a minimum and thus provide efficient analysis. Aside from providing efficiency in the model implementation, FS can enrich simple and less complicating with a lesser computational cost model as a subset rather than the exclusive features used in implementation (Guyon and Elisseef, 2003). FS is vastly categorized as the filter, wrapper, and the embedded method. See figure 1 for the classification.

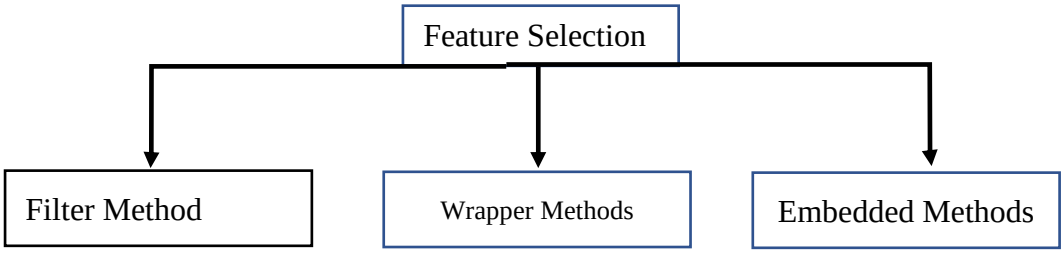

Figure 1 Feature Selection Classification

The filter method uses the statistical characteristics of the dataset, providing feature ranking as output, selecting features regardless of the Model. Example of the many standard filter methods can be seen in the research by (Jović et al., 2015). Though filter methods are easily employed due to their low cost for computation, the wrapper method is better because of their search strategy on a modelling algorithm (Jović et al., 2015). Furthermore, wrapper methods are designed to evaluate individual feature subset using learning algorithms (see Figure 2, showing the architectural design of wrapper FS method.

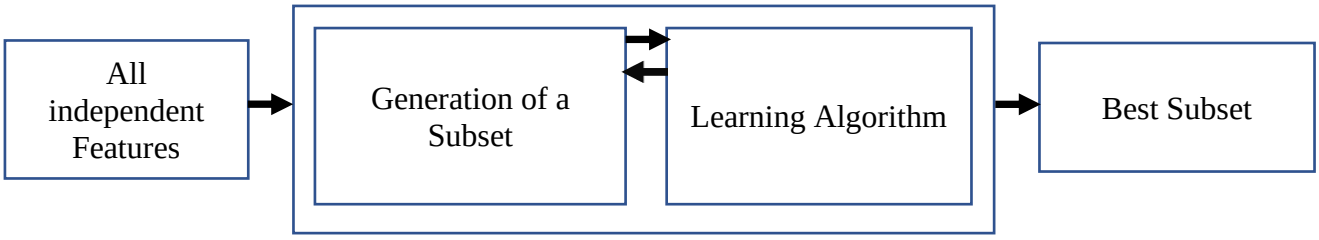

Figure 2 Wrapper Feature Selection

As a plus to the two methods, i.e., filter and wrapper FS methods, embedded FS methods select features during the algorithm modelling implementation. This method enjoins both performance and computational cost advantages from the filter and wrapper method. Figure 3 shows precisely how a subset is produced from learning algorithm performance. An interested reader on the embedded method is referred to (Huan and Lei, 2005).

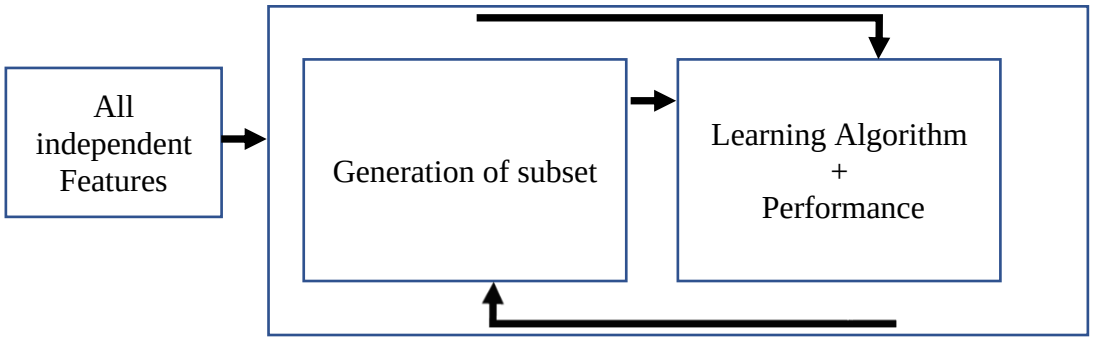

Figure 3 Embedded FS

This paper is looking into investigating relevant features helpful in developing the many NO2 ML models. Embedded FS is a good choice for FS. However, its efficiency is somewhat dependent on the learning algorithm in many cases. As such, this paper employs Boruta, a wrapper FS to select the best subset features. The Boruta FS was implemented using the sklearn package in Python 3.8.

3 METHOD

Data Description and Big Data Analytics

To monitor/reduce exposure to air pollution, most cities now deploy monitoring sensors for measuring traffic intensity, weather characteristics, and air quality of the environment. For this project, a total of 14 Internet of thing (IoT) monitoring sensors for NO2 and other pollutant concentration represented as blue circles were deployed across Wolverhampton City in the UK (see Figure 4)

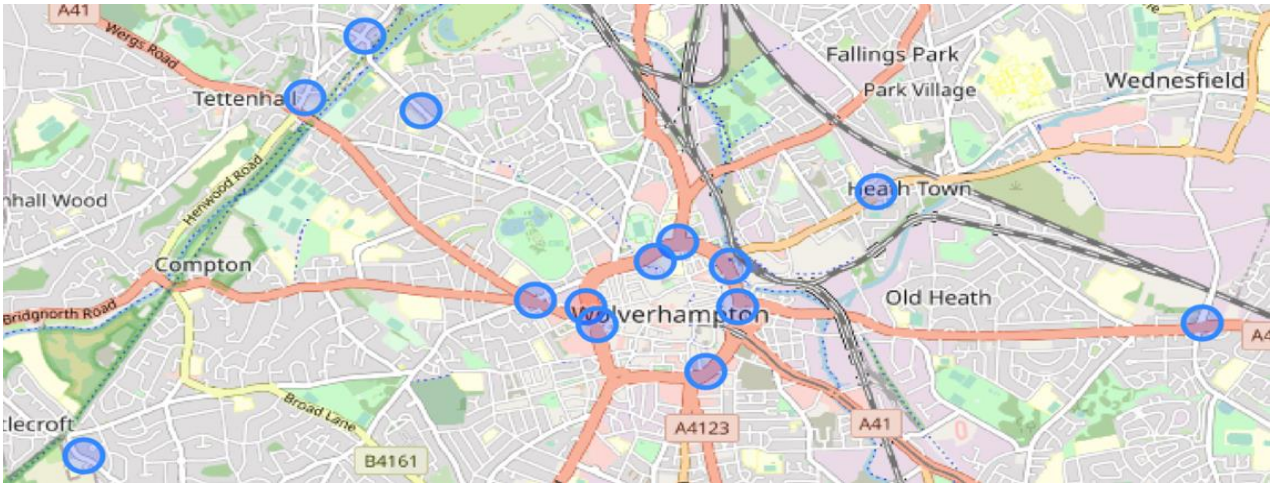

Figure 4 Map Showing the 14 NO2 IoT Monitoring sensors deployed

The data from the 14 IoTs monitoring sensors for NO2 pollutant concentration was the dependent variable. In contrast, weather, traffic data were the independent variables. The traffic data was sourced from the UK's Traffic data provider 'DfT' and included mainly vehicle types and road types (see Table 2). In addition, the weather data for a similar period was recovered from the UK Met Office. It included various weather variables like ambient pressure and humidity, among others (see table 2). Traffic and weather data were provided hourly, and each had over fifty thousand data points. The three data sources were matched, giving (24hrs x 30days x 5months x 14 IoTs) data points.

Table 2 Independent features after matching the three data sources

S/N	**Features**	**Unit**	**Data Source**
1	Ambient Humidity	RH	UKMETOFFICE
2	Ambient Pressure	Pa	
3	Ambient Temp	^{0}C	
4	Humidity	RH	
5	Temp	^{0}C	
6	Road Type	-	DFT
7	Link Length in Km	-	
8	Link Length in Miles	-	
9	Pedal Cycles	-	
10	Two Wheeled Motor	-	
11	Cars and Taxis	-	
12	Buses and Coaches	-	
13	Lgvs	-	
14	Hgvs 2 Rigid Axle	-	
15	Hgvs 3 Rigid Axle	-	
16	Hgvs 4 Or More Rig.	-	
17	Hgvs 3 Or 4 Articulate Axle	-	
18	Hgvs_5_Articulated_Axle	-	
19	Hgvs_6_Articulated_Axle	-	
20	All Hgvs	-	
21	All Motor Vehicles	-	
22	Zid	-	

23	Date	-	IoT
24	Holiday	-	
25	Day of the week	-	
26	X (3d Coordinates)	-	
27	Y (3d Coordinates)	-	
28	Z (3d Coordinates)	-	
29	NO2	$\mu g/m^3$	

Boruta Algorithm

The wrapper Boruta algorithm (WBA) FS method was implemented for feature selection in this study. WBA finds the importance of a feature by creating shadow features; see the implementation steps.

1. First, we introduced randomness to the union of all the input data set by creating rearranged replicas of all features (i.e., shadow features).
2. Then, we trained an RF classifier on the complete data set and relates a feature importance measure to evaluate the importance of individual features.
3. At every repetition, we checked if a feature has higher importance than the best of its shadow features (i.e., whether the feature has a higher Z-score than the maximum Z-score of its shadow features)
4. We constantly remove features that are deemed highly unimportant.
5. Finally, we stop the algorithm when all features reach a specified limit of random forest runs.

4 DISCUSSION OF RESULT

For this paper, Boruta FS was implemented to select impactful features useful in developing NO2 ML pollution models. This FS method was implemented with NO2 pollution as the dependent feature. Completing the implementation of Boruta FS, six out of the twenty-nine features were identified as most relevant. The non-redundant features identified include; ambient humidity, Ambient pressure, Ambient temperature, Days of the week, two-wheeled vehicles(counts), cars/taxis(counts).

In addition to the FS Methods implemented, the research took a step ahead to investigate the relevance of the features selection method by developing NO2 pollution ML models. As a result, ML models such as Random Forest (RF), Support vector machine (SVM), Decision tree (DT), XGBoost (XGB), Adaboost, Artificial neural network (ANN) and Linear Regression (LR) were identified as reliable and robust machine learning algorithms for prediction (Bilal & Oyedele, 2020; Choi et al., 2020; Mehtab & Sen, 2020; Purnus & Bodea, 2017) and developed using the features selected by the Boruta FS method.

Given that feature selection may not be entirely favourable to some algorithms (H. Alaka et al., 2018; Hafiz et al., 2015), we developed the predictive models for each algorithm in two ways to allow fairer comparison. The first was to develop the models using all the available variables before any feature selection processes. The results from this were recorded and compared (see Table 3). The second was to develop the models using the features selected with the Boruta algorithm. The results from this were also recorded and compared (see Table 3).

The results of the two-stream of comparison were investigated using the system of measurement: 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.

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

MSE is yet an alternative risk system of measurement that corresponds to the average of all the error squares between the predicted dependent feature value and the actual dependent feature value. MSE is.

$$\text{MSE} = {}^{1}/_{n} \sum_{i=0}^{n} (y_i - y_i^*)^2 \quad (2)$$

Unlike the risk system of measurement (i.e., MAE, MSE). R-squared is the coefficient of determination indicating the goodness of fit. R-square is.

$$R^2 = 1 - \frac{\sum_{i=0}^{n} (y_i - y_i^*)^2}{\sum_{i=0}^{n} (y_i - \bar{y}_i)^2} \quad (3)$$

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

As stated earlier, Table 3 shows the performance of the developed NO2 ML models with and without Boruta feature selection represented as WBA (With Boruta Algorithm) and WoBA (without Boruta Algorithm).

Table 3: NO2 Pollution ML models developed with and without feature selection

Model/Metrics	**MAE**		**MSE**		**R-square**	
	WoBA	**WBA**	**WoBA**	**WBA**	**WoBA**	**WBA**
KNN	7.6	7.59	152.9	131.93	0.73	0.77
RF	5.8	7.66	110.8	132.87	0.79	0.77
ANN	8.3	8.26	164	134.3	0.71	0.77
XGB	9.4	9.01	219	189.1	0.62	0.74
SVR	9.5	9.01	219.5	189.1	0.62	0.67
LR	9.5	10.3	219.5	199.98	0.62	0.62
DT	8.1	10.49	192	267.48	0.67	0.54
ADB	18.4	14.5	523	332.67	0.1	0.42

Figure 5 present the R-square score for all the NO2 pollution ML models implemented using features selected with the Boruta Algorithm and using all independent features united in the research.

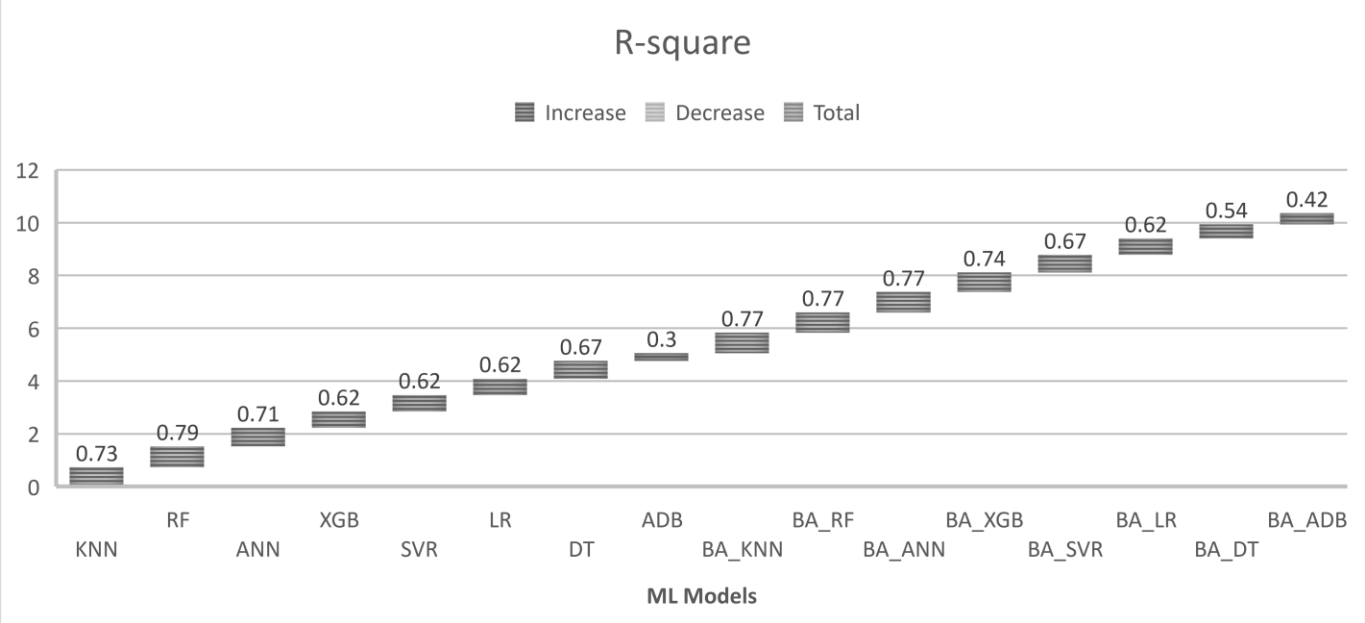

Figure 5 Rsquare score for all the two streams of ML developed

As discovered from the results of the NO2 Model developed, there was some slight improvement in the models developed without using the Boruta FS. For instance, Adaboost without FS method scored 30% for the R-square value and was improved by up to 12% when developed using features selected with Boruta FS. Another interesting example is that the KNN model improved by 2% simply by using features selected with Boruta. The same happens with XGBoost, Support Vector Machine.

However, the tree-based algorithms; Decision tree, and Random Forest depreciated in performance upon using features selected with Boruta FS. The depreciation is because tree-based ML models have ways of pruning features, picking, and splitting the trees in order of relevance, for instance, the Gini entropy and likes.

The assertion of the improvement caused by feature selection was proved in this paper; for instance, the first approach (i.e., Model's implementation without feature selection) got improved when the same models were developed using Boruta selected features.

5 Conclusions

NO2 pollution prediction using Machine learning (ML) offers many advantages, including guiding people vulnerable to air pollution. There are multiple sources of data that influence NO2 air pollution: the pollution data, origin of NO2 pollution (e.g., vehicle, industry, other chemical involved activities), weather data (e.g., Temperature, wind, rainfall, among others), geographical data (e.g., geolocation, activities around certain areas for instance, trading or manufacturing or schooling among others), Built environment (e.g., building height, building size, and building type). There are numerous such Model developed using diverse data. However, the diversity of the existing data makes it a challenge to get the most efficient and effective ML. In this paper, we unite pollution-related data sets and investigate exact features that are useful in developing efficient NO2 pollution ML models. At the end of the study, the following list of the inferences can be reached, including:

i. Boruta, a dimensionality selection technique, improves the performance of many ML model

ii. There is no need to subject independent features to Feature selection when developing Random forest and Decision tree
ii. The six most impactful features in NO2 ML model development are Ambient humidity, Ambient pressure, Ambient temperature, Days of the week, two-wheeled vehicles(counts), cars/taxis(counts).
iv. Surrounding weather features like ambient temperature, ambient humidity, and ambient pressure were identified as more relevant.

Future studies should explore the Embedded feature selection method and possibly compare the performance. Also, further research may look into all built environment properties and topography.

6 References

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., & Ajayi, S. (2018). Firms Failure Prediction Models. *IEEE Transactions on Engineering Management*, *PP*(4), 1–10. https://doi.org/10.1109/TEM.2018.2856376

Bilal, M., & Oyedele, L. O. (2020). Guidelines for applied machine learning in construction industry—A case of profit margins estimation. *Advanced Engineering Informatics*, *43*(March 2019), 101013. https://doi.org/10.1016/j.aei.2019.101013

Choi, J., Gu, B., Chin, S., & Lee, J. S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction*, *110*(May 2019), 102974. https://doi.org/10.1016/j.autcon.2019.102974

Dou, X., Liao, C., Wang, H., Huang, Y., Tu, Y., Huang, X., Peng, Y., Zhu, B., Tan, J., Deng, Z., Wu, N., Sun, T., Ke, P., & Liu, Z. (2020). Estimates of daily ground-level NO2 concentrations in China based on big data and machine learning approaches. *ArXiv*, *2*.

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

Huan, Liu., Lei, Y. (2005). Towards integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4).

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

Juhos, I., Makra, L., & Tóth, B. (2008). Forecasting of traffic origin NO and NO2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis.

Simulation Modelling Practice and Theory, *16*(9), 1488–1502. https://doi.org/10.1016/j.simpat.2008.08.006

Kamińska, J. A. (2019). A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions. *Science of the Total Environment*, *651*(2), 475–483. https://doi.org/10.1016/j.scitotenv.2018.09.196

Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment*, *35*(5), 815–825.

Kopparapu, R., Arney, G., Haqq-Misra, J., Lustig-Yaeger, J., & Villanueva, G. (2021). Nitrogen Dioxide Pollution as a Signature of Extraterrestrial Technology. *The Astrophysical Journal*, *908*(2), 164. https://doi.org/10.3847/1538-4357/abd7f7

Mehtab, S., & Sen, J. (2020). *Stock Price Prediction Using Convolutional Neural Networks on a Multivariate Timeseries*.

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*.

Purnus, A., & Bodea, C. N. (2017). A Predictive Model of Contractor Financial Effort in Transport Infrastructure Projects. *Procedia Engineering*, *196*(June), 746–753. https://doi.org/10.1016/j.proeng.2017.08.003

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, *8*, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

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.

Xu, Z., Huang, G., Weinberger, K. Q., & Zheng, A. X. (2014). Gradient boosted feature selection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 522–531. https://doi.org/10.1145/2623330.2623635