Boruta-grid-search least square support vector machine for NO_2 pollution prediction using big data analytics and IoT emission sensors

Habeeb Balogun, Hafiz Alaka and Christian Nnaemeka Egwim
Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield, UK

Received 21 April 2021
Revised 24 June 2021
5 July 2021
Accepted 7 July 2021

Abstract

Purpose – This paper seeks to assess the performance levels of BA-GS-LSSVM compared to popular standalone algorithms used to build NO_2 prediction models. The purpose of this paper is to pre-process a relatively large data of NO_2 from Internet of Thing (IoT) sensors with time-corresponding weather and traffic data and to use the data to develop NO_2 prediction models using BA-GS-LSSVM and popular standalone algorithms to allow for a fair comparison.

Design/methodology/approach – This research installed and used data from 14 IoT emission sensors to develop machine learning predictive models for NO_2 pollution concentration. The authors used big data analytics infrastructure to retrieve the large volume of data collected in tens of seconds for over 5 months. Weather data from the UK meteorology department and traffic data from the department for transport were collected and merged for the corresponding time and location where the pollution sensors exist.

Findings – The results show that the hybrid BA-GS-LSSVM outperforms all other standalone machine learning predictive Model for NO_2 pollution.

Practical implications – This paper's hybrid model provides a basis for giving an informed decision on the NO_2 pollutant avoidance system.

Originality/value – This research installed and used data from 14 IoT emission sensors to develop machine learning predictive models for NO_2 pollution concentration.

Keywords IoT, Bigdata, Air pollution prediction, Hybrid machine learning

Paper type Research paper

1. Introduction

Air pollution, a release of pollutants into the air, remains one of the significant challenges in the UK and globally, with over 25,000 associated deaths recorded yearly in the UK [1] and around 8.8 million deaths recorded globally [2]. Apart from deaths, air pollution exposure can result in various short and long-term health challenges [3, 4]. Examples of short-term health challenges include eye pain, throat irritation, headaches, allergic reactions, and upper respiratory infections. While lung cancer, brain damage, liver damage, kidney damage, heart disease, respiratory disease, and suchlike are examples of long-term health challenges [5].

Aside from the severe impact of air pollutants on health, air pollution has significant consequences on the UK and the global economy. It costs the UK government approximately £40bn yearly [6] and around £3 trillion economic costs globally [7]. Recent studies by the centre for research on energy and clean air (CREA) links over 1.5 billion days of absence from

Applied Computing and Informatics
Vol. 21 No. 1/2, 2025
pp. 101-113
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-04-2021-0092

work, over 3.5million new cases of asthma and approximately 2million preterm births to air pollutants leading to an increase in health care cost and decrease in economic productivity.

Air pollutants are airborne substances, usually of two categories: particulate matter and gases. Of the gases, Nitrogen dioxide (NO_2) is arguably the most dangerous to human health [8]. NO_2 pollutant emanates from combustion processes such as vehicle emissions, and this was noted during the Covid-19 pandemic, with a 20% decrease in global NO_2 concentration [9]. However, a recent finding hypothesized that NO_2 will still exceed the (Air quality index) AQI limit by 2025 [3]. Therefore, the NO_2 AQI estimate poses a responsibility to stakeholders and researchers to devise strategic means to curb exposure to this UK's pollutant.

Arguably, predicting NO_2 concentration is among the most efficient and effective ways to save lives from exposure to this deadly pollutant in different geographical locations. Furthermore, this prediction can help people avoid such areas when they have high NO_2 concentration levels.

1.1 Related Work

Studies on NO_2 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.

A machine learning-built predictive model's performance is, among other factors, vastly dependent on the machine learning algorithm used [10, 11]. Several studies have thus compared some of the most popular algorithms (e.g. artificial neural network, support vector machine, and suchlike) in terms of their performance in predicting NO_2 [12– 14] with Random forest, support vector machine usually performing better. However, despite clear proof from the literature that hybrid algorithms have performed better than standalone, they have not been vastly employed in the comparison studies [15, 16].

One such hybrid is the optimal-hybrid artificial intelligent algorithm based on the Least squares support vector machine optimized by grid search, whose features were selected using the Boruta Algorithm (BA-GS-LSSVM). The Least square support vector machine (LSSVM) differs from the classical SVM due to improved objective function. LSSVM is widely used for classification and regression problems due to its high predictive ability compared to classical SVM. Findings from research like the prediction of gasoline's price [17], speed of wind's forecast [18] indicated that this model presents more operation speed and convergence accuracy. However, some shortcomings are associated with this algorithm's performance, including optimizing parameter and feature selection. The BA-GS-LSSVM solves these comings.

Thus, this paper seeks to assess the performance levels of BA-GS-LSSVM compared to popular standalone algorithms used to build NO_2 prediction models. The objectives are as follows:

(1) To pre-process a relatively large data of NO_2 from IoT sensors with time-corresponding weather and traffic data

(2) To use the data to develop NO_2 prediction models using BA-GS-LSSVM and popular standalone algorithms to allow for a fair comparison.

It is imperative to describe the symbols used in this research work. Table 1 defines most of the symbols and their description.

The rest of this paper organized as follows: section two presents a brief explanation of the source of data and data volume. Section three presents feature selection techniques used in selecting the valuable features for developing the hybrid model. Section four presents the

hybrid model detailing the theoretical/mathematical representation of the model and how it differs from classical SVM. Lastly, Section five describes the development of the BA-GS-LSSVM, other popular standalone machine learning algorithms for NO_2 prediction and their performance assessment for comparison. Finally, the conclusion and discussion form part of the fifth section.

2. Data description and big data analytics

Many UK cities, just like other cities of the world, suffer from air pollution. A significant contributor to air pollution is increasing traffic emissions [19]. Air pollution caused by traffic depends on the type of vehicle (diesel, gasoline, petrol, electric), level of congestion, time spent in the traffic jam, and the atmospheric/geographical features of the environment at a given time.

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. The data is collected at specified frequencies (seconds, minutes, hours, days, and suchlike) depending on the users' preference. For this project, a total of 14 Internet of Things (IoT) monitoring sensors for NO_2 and other pollutant concentrations represented as blue circles were deployed across Wolverhampton City in the UK (see Figure 1).

The sensors collected NO_2 concentration and other harmful pollutant's data every 10 s for five months (i.e. December 2019 and April 2020). Over ten billion (i.e. $10 \times 6 \times 60 \times 60 \times 24 \times 30 \times 5 \times 14$) data points were generated for this period which was massive. The data through the Middleware gateway deployed on elastic bean of amazon web service

Symbol	Description
y_I	Observed NO_2 concentration
y_i^*	Predicted value for NO_2 concentration
Log(n)	Depth of tree
n	Number of rows
d	Number of features
t	Number of trees
k	Number of k neigbour
γ	Regularisation parameter
σ^2	width of Kernel parameter

Table 1. Symbol and description

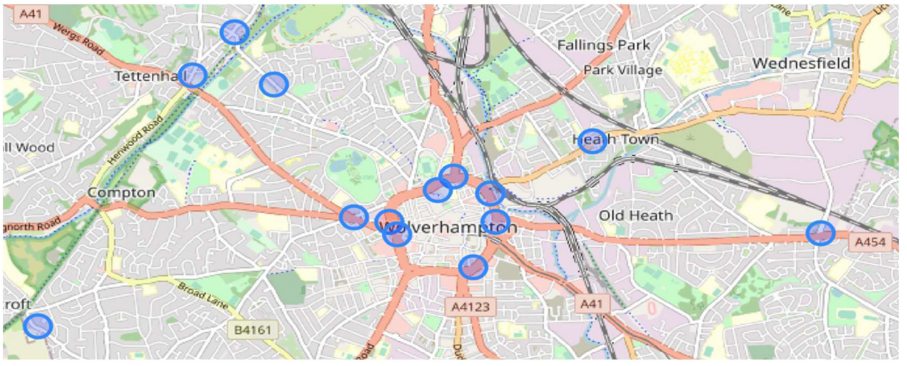

Figure 1. Map showing the 14 NO_2 IoT monitoring sensors deployed at Wolverhampton City, UK

(AWS) directly dump the data into an AWS Elastic Computing cloud two (EC2) Relational database. We used the AWS EC2 infrastructure to run the big data analytics required for this study due to the extensive data.

For the development of the BA-GS-LSSVM, the data from the 14 IoTs monitoring sensors for NO_2 pollutant concentration was the dependent variable. In contrast, weather, other pollutants, e.g. PMx and Ozone and traffic data, were the independent variables. The traffic data was sourced from the UK's Department for Traffic (DfT) and included mainly vehicle counts, split into various vehicle types (see Table 2). The traffic data, covering the same period as the data from the sensors, were retrieved. 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. To match the weather and traffic data with the pollutant data from the sensors, the hourly average of the pollutant concentration was used to match the corresponding hourly weather and traffic data leading to (24 hrs × 30 days × 5months × 14 IoTs) data points.

The concentration of NO_2 from December 2019 to April 2020 across the 14 installed sensors indicates some interesting trends, with some outliers around the end of 2019 (see Figure 2). These outliers at the end of the year are arguable due to the shopping, Christmas and other season celebration. In addition, the pollution concentration is arguably

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	–	IoT
23	Date	–	
24	Holiday	–	
25	Day of the week	–	
26	X (3d coordinates)	–	
27	Y (3d coordinates)	–	
28	Z (3d coordinates)	–	
29	Pm1	µg/m^3	
30	Pm10	µg/m^3	
31	Pm25	µg/m^3	

Table 2.
Independent features after matching the three data sources

influenced by the national lockdown imposed across the UK cities during the covid19 pandemic (see Figure 2).

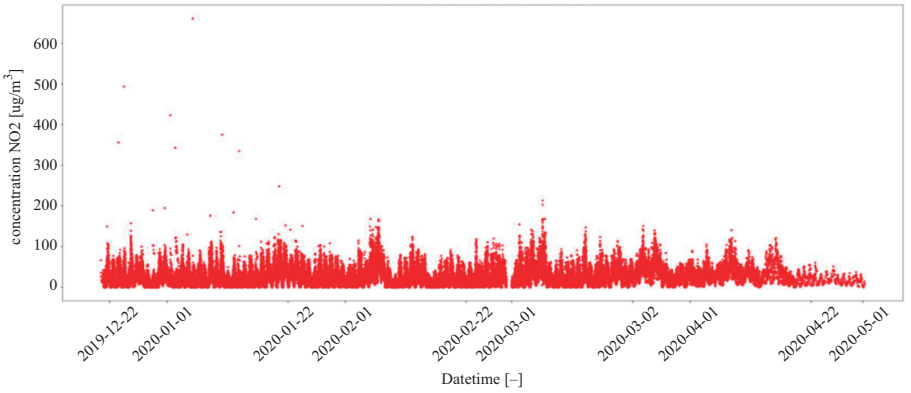

Figure 2.
The hourly averaged NO_2 concentration across the 14 IoT emission sensors

Another exciting exploration is the outliers discovered within some days of the week (see Figure 3). Looking at the boxplot, it is arguably the days with the highest amount of traffic as we can hypothetically say these are days many go out to bars, clubs and other gatherings at the end of the week.

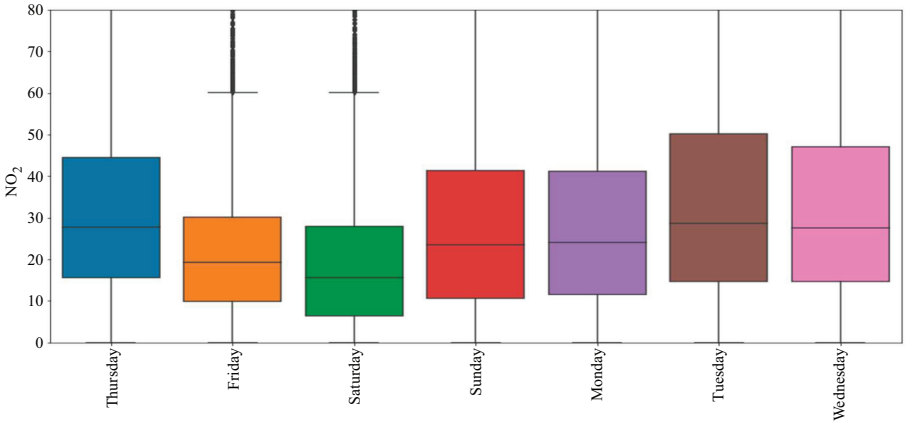

Figure 3.
Day of the week NO_2 Concentration collected across the 14 IoT sensors

After pre-processing was completed on the AWS Big data infrastructure, the complete data was split into data (60%) for training and (40%) for model testing at random to avoid biases and other shortcomings.

3. Feature selection

The predictive capability of various machine learning depends on the features' dimensionality; LSSVM is not an exception [20]. Not all features impact the prediction, making feature/variable selection critical in developing/building machine learning predictive models.

Dimensionality reduction has been proven to help make predictive models perform better [21]. Of the reduction techniques, feature selection is selecting the most impactful features from the original set of features as the new input features.

Since Random forest (RF) has consistently proven in past studies, e.g. [21–24], to be very good at selecting the most impactful features, the wrapper Boruta algorithm(BA) built around the RF was implemented for feature selection in this study. BA uses the same strategy as the classical RF classifier model introduced by [25]. The BA is implemented using the following steps:

(1) Replicate and add a copy of all input features, i.e. weather and traffic features, to form an information system (IS)

(2) Shuffle the IS and remove correlations among features in the IS

(3) Apply a random forest classifier on the comprehensive IS

(4) compute the Z scores represented as n for all the features and Identify the maximum n among shadow features (MnSF)

(5) Assign a value to every feature that scores more than MnSF.

(6) Carry out a test of equality and drop features lower than MnSF

(7) Eliminate all the shadow features

(8) Repeat the procedure

After applying the feature selection process, a total of 13 essential features were selected by the BA, namely, Timestamp, O_3, All motor vehicles, Humidity, Ambient pressure, Temperature, PM10, PM2.5, PM1, Day of the week, and x,y,z which is the 3d-geocentric-representation of the longitude and latitude. The timestamp arguably suggests some level of consistency in the pollutant levels at a specific time. For instance, the morning peak period or close of the day peak periods results in higher traffic. Afterwards, the 13 selected features will be used in developing an LSSVM predictive model, as discussed in the next section.

4. Least square support vector machine

LSSVM, improvement to SVM was proposed [15]. LSSVM provides a linear equation solution with an improvement in the objective function of classical SVM.

We use x_k *as the* 13 *feature selected with BA and* y_k *is the* NO_2 *concentration*

Then the improved SVM model can be mathematically written:

$$y(x) = \omega^T.\varnothing(x) \; + \; b \tag{1}$$

where, $\varnothing(x)$ = nonlinear mapping function, ω = weight, and b = bias.

The equation can be expressed:

$$\min_{\omega,b,e}(\omega, e) \; = \; \frac{1}{2}\,\omega^T\omega \; + \; \frac{1}{2}\gamma\sum_{k=1}^{n} e_k^2 \tag{2}$$

Subject to

$$y_k = \omega^T\varnothing(x_k) \; + \; b \; + \; e_k, \quad k \; = \; 1, 2, \; \ldots n \tag{3}$$

Where, γ = regularisation parameter and e_k = error term.

The model can be optimized using the LaGrange function as follows

$$L(\omega, b, e, \alpha) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{n} e_k^2 - \sum_{k=1}^{n}\{\alpha_k[\omega^T\varnothing(x_k) + b + e_k - y_k]\} \quad (4)$$

where

$\alpha_k \in R$ = Lagrange multiplier

From Karush-Kuhn-Tucker (KKT) equation given as.

$$\begin{cases} \omega = \sum_{k=1}^{n}\alpha_k\varphi(x_k) \\ \sum_{k=1}^{n}\alpha_k = 0 \\ \alpha_k = e_k\gamma,\ \omega^T\varphi(x_k) + b + e_k - y_k = 0 \end{cases} \quad (5)$$

The optimization equation can be transformed to the linear equation given as Eqn 6, after eliminating the variables ω and e_k.

$$\begin{bmatrix} 0 & 1 & \ldots & 1 \\ 1 & K(x_1, x_1) + 1/\gamma & \ldots & K(x_1, x_i) \\ 1 & K(x_j, x_1) & \ldots & K(x_j, x_i) + 1/\gamma \end{bmatrix}\begin{bmatrix} b \\ \alpha_1 \\ \alpha_j \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ y_j \end{bmatrix} \quad (6)$$

The final equation of the LSSVM model is.

$$f(x) = \sum_{k=1}^{j} \propto_k K(x, x_i) + b \quad (7)$$

where

$$K(x, x_i) = \varphi(x)^T * \varphi(x_j) \text{ is the kernel function.}$$

The finite response and the Radial basis function (RBF) kernel function was used in this research and mathematically expressed as,

$$K(x, x_i) = exp\left(-\gamma * |x - x_i|^2\right) \quad (8)$$

Where, $\gamma = \frac{1}{2\sigma^2}$

4.1 Grid search

After the development of the LSSVM using the features selected with the BA, the optimization of the LSSVM model parameter: regularisation parameter (γ) and kernel-parameter (σ^2) is another challenging area that should not be ignored as it could lead to poor prediction performance if not carefully chosen.

The choice is where the grid search algorithm comes in. It does this by pairing the all-possible values of regularisation and kernel parameter (γ, σ^2). It is applied to optimize these two parameters to have an improved prediction capability. Each pair of regularisation and kernel parameters is subjected to cross-validation and hence producing MSE value.

For this study, the LSSVM parameters were initialized, taking a search range [0.01, 1000] and [0.1, 1000] for γ and σ^2, respectively. Afterwards, cross-validation, with all possible values of γ, and σ^2, the pair with the minimum MSE is the best, and that was used to create GS-LSSVM.

5. Model development process and performance measures

Related works on the development of predictive models (e.g. [26– 29], identified Random forest (RF), Support vector machine (SVM), Decision tree (DT), XGboost (XGB), Adaboost, Artificial neural network (ANN) and Linear Regression (LR) as powerful machine learning algorithms for prediction. These popular algorithms were developed and compared with BA-GS-LSSVM.

Given that feature selection may not be entirely favourable to some algorithms [11, 20], 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 the feature selection processes. The results from this were recorded and compared (see section 5 on results). The second was to develop the models using the 13 features selected with the Boruta algorithm. The results from this were also recorded and compared (see section 5 on results). Finally, the best results for each algorithm (whether from the first or second development) were compared to determine the best algorithm overall.

The LSSVM Regression model has no specific python package, so we have implemented the Scikit learn package in python. Figure 4 presents the flow chart and overall procedures to build the hybrid GS-LSSVM predictive model to predict the concentration of NO_2.

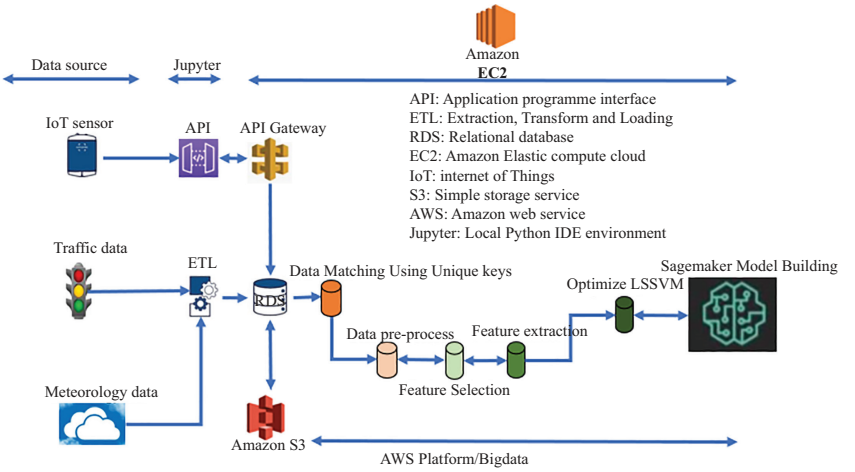

Figure 4.
Flowchart for the GS-LSSVM

To determine the predictive capability of a regression predictive machine learning model, various metrics measures loss and score models. Among these metrics, four, including the mean absolute error (MAE), mean square error (MSE), Explained variance score (EVS) and R Squared (R^2), were used in this paper because of their popularity and are briefly described below.

MAE is the average absolute variation(error) between each point in a scatter plot between the actual observation and the corresponding predicted value. It is a risk metric corresponding to the expected value of the absolute error loss. The best possible score is 0.00; the higher the MAE, the worse the predictive model's performance. The MAE can be mathematically written as

$$\mathrm{MAE} = 1/n \sum_{i}^{n} \left| y_i - y_i^* \right| \tag{9}$$

MSE is another risk metric that corresponds to the average of all the error squares between the predicted value and the actual value of the target variable. It is also referred to as mean squared deviation. MSE value is strictly positive, ranging between [0,1], and values closer to zero signifies a better predictive model. The mathematical definition of MSE is as follows.

$$\mathrm{MSE} = 1/n \sum_{i=0}^{n} \left(y_i - y_i^*\right)^2 \tag{10}$$

Unlike the risk metric functions (i.e. MAE, MSE), The Explained variance score and R-square score depicts a better regression model when the score is getting closer to 1.0 and not zero. 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 it is mathematically written as.

$$\mathrm{EVS} = 1 - \frac{y_i - y_i^*}{y_i} \tag{11}$$

Lastly, *R*-squared referred to as the coefficient of determination is the proportion of dispersion in the feature(s) and the target variable. It indicates the goodness of fit and aid the measurement of how well-unseen data are likely to be predicted by the model. The best possible value for *R*-square is 1.0. It is mathematically given as.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=0}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{12}$$

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

5.1 Discussion of result

In this study, an optimal-hybrid artificial intelligent algorithm based on the Least squares support vector machine was optimized by grid search, whose features were selected using the Boruta Algorithm (BA-GS-LSSVM) to predict NO_2 pollutant concentration were developed. We identified the most optimal values for the parametric functions of LSSVM to be $\gamma = 1000\, and\, \sigma^2 = 10$ using grid the search.

The model was all developed following the union of the three data sources, including the weather, traffic, and IoT data on a big data platform considering many data points recorded (i.e. 24 hrs × 30 days × 5 months × 14 IoTs). The data were merged and matched for the development of the predictive models. We then compared the performance capability of the proposed hybrid model and other powerful standalone machine learning in predicting NO_2, however, in two streams to achieve a fair comparison with no bias. The two streams of comparisons to avoid biases are; (1) All models implemented without feature selection (2) All Models implemented with feature selection. Table 3 and Figure 5 presents metrics for developed models with (WF) and without feature (WoF) selection for a fair comparison. The BA-GS-LSSVM and all other ML models developed were done on a big data platform due to the algorithms' time complexity. The time complexity of the models developed are GS-LSSVM:O(n^3), RF:O(d*log(n)), KNN:O(knd), ANN:O(n^4), DT:O(n*log(n)*d), SVR:O(n^3), XGB:O(n*d*log(n)), LR:O(nd), LSSVM:O(n^3), and ADB:O(nd^2).

As shown in Figure 5, the error measures, including the MAE, MSE for all the developed models, were presented in decreasing order. The order, in this case, shows the Adaboost (AB) to have the maximum error, followed by the LSSVM and linear Regression (LR) implemented without feature selection. At the same time, we can see GS-LSSVM with feature selection, i.e. BA-GS-LSSVM with the most negligible error score value. This explains the higher performance ability of the hybrid model over other standalone models.

Table 3.
Results of the overall model developed

Model/Metrics	MAE		MSE		*R*-square		EVS	
	WoF	WF	WoF	WF	WoF	WF	WoF	WF
GS-LSSVM	8.32	6.91	114.57	97.08	0.76	0.82	0.76	0.82
KNN	7.6	7.59	152.9	131.93	0.73	0.77	0.73	0.77
RF	5.8	7.66	110.8	132.87	0.79	0.77	0.79	0.77
ANN	8.3	8.26	164	134.3	0.71	0.77	0.71	0.77
LSSVM	11.6	8.32	285.5	114.57	0.51	0.76	0.51	0.76
XGB	9.4	9.01	219	189.1	0.62	0.74	0.62	0.74
SVR	9.5	9.01	219.5	189.1	0.62	0.67	0.62	0.67
LR	9.5	10.3	219.5	199.98	0.62	0.62	0.62	0.62
DT	8.1	10.49	192	267.48	0.67	0.54	0.67	0.54
ADB	18.4	14.5	523	332.67	0.1	0.42	0.1	0.42

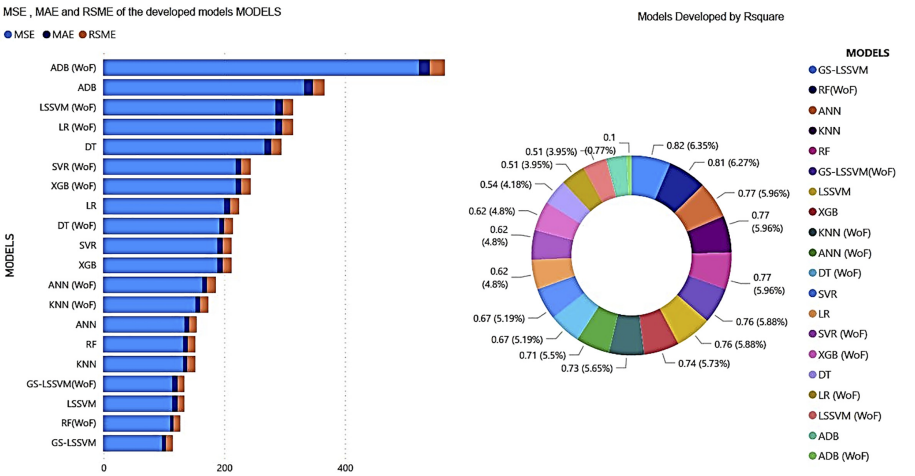

Figure 5.
The NO_2 concentration predictive models developed

In addition, the doughnut chart shows the *R*-squared score for all the models developed, and the maximum score was yielded in the development of BA-GS-LSSVM. i.e. 6.35%(0.82). The assertion of the bias caused by feature selection was proved in this paper; for instance, the first approach (i.e. model's implementation without feature selection) shows poor and woeful model performance.

In addition, the models developed with feature selection was subjected to the 10-fold cross-validation to ensure efficient/unbias evaluation. For this, the research present in Figure 6, a box and whisker plot showing the spread in different performance metrics across each cross-validation fold for each algorithm

From these results, BA-GS-LSSVM is identified the best considering its minimal error metric scores (i.e. lowest MAE and MSE score) compared to other ML developed. At the same time, BA-GS-LSSVM has scored the highest EVS and *R*-square score. Thus, our model Conclusively performs best compared to all other standard and powerful standalone ML models developed in this paper.

The use of the big-data platform reduced the computational complexity for most of the models implemented. Also, the lower computational complexity of the LSSVM over the SVM is another outstanding advantage recognized in this research.

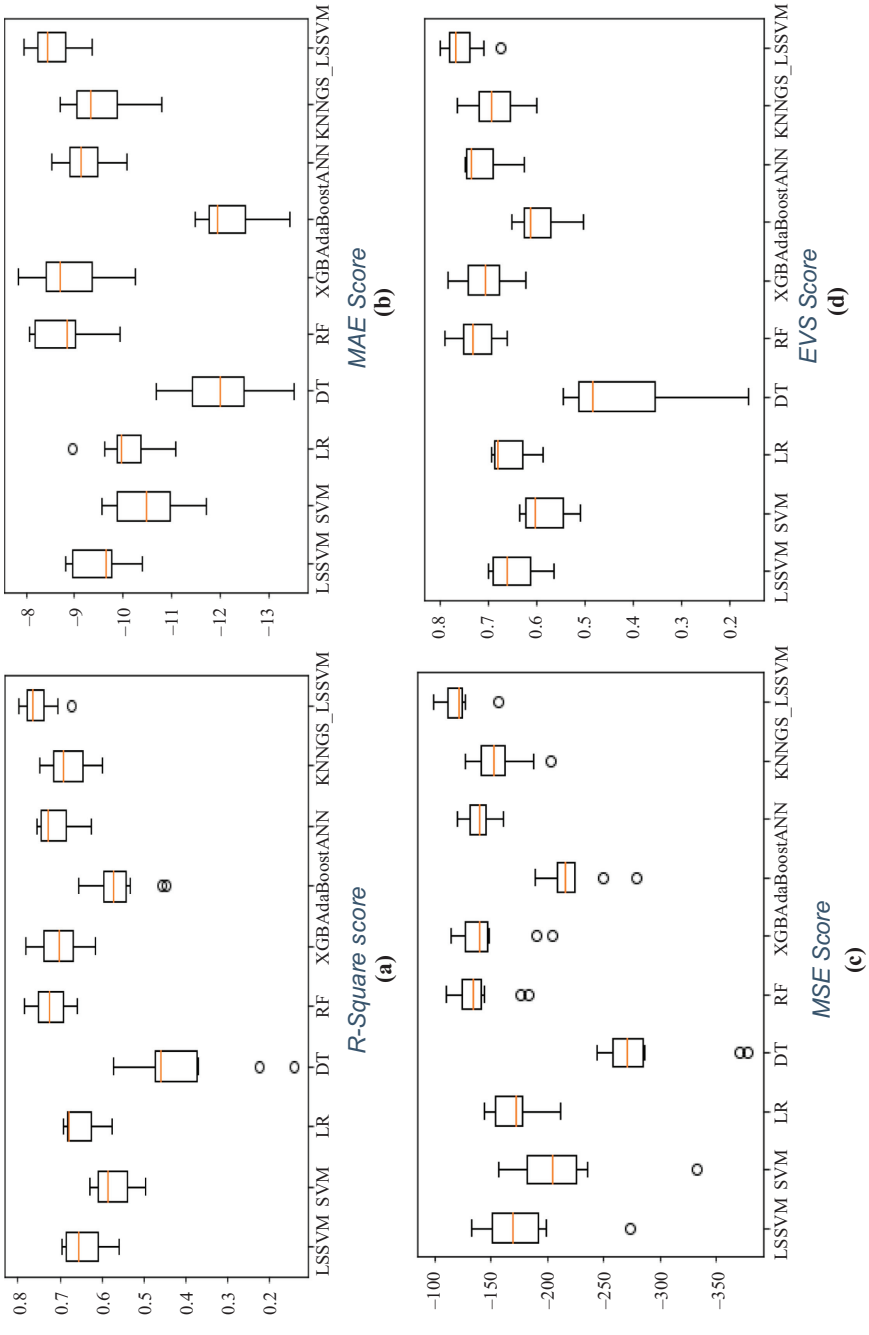

Figure 6.
Graphical illustrationss of the spread in (a) *R*-square, (b) MAE, (c) MSE, and (d) EVS score for each algorithm developed with feature selection

6. Conclusions

High-precision NO_2 prediction is critical to people's well-being, especially those that are vulnerable to air pollution. However, the BA-GS-LSSVM model in this paper happens to be appealing and proves to be better than popular algorithms. To demonstrate the advantages of the BA-GS-LSSVM model, nine different algorithms were compared. At the end of the study, the following list of the inferences can be reached, including:

(1) Boruta, a dimensionality selection technique, improves the performance of the ML model

(2) Compared with all other standalone models developed, our model, BA-GS-LSSVM, exhibits a better predictive ability in NO_2 concentration

(3) The BA-GS-LSSVM model provides a basis for delivering an informed decision on the NO_2 pollutant avoidance system.

Future studies should explore the use of BA-GS-LSSVM to predict other pollutants in the UK and other parts of the world experiencing this outburst in air pollution concentration.

References

1. Public Health England. Review of interventions to improve outdoor air quality and public health. 2019.
2. Nethery RC, Dominici F. Estimating pollution-attributable mortality at the regional and global scales: challenges in uncertainty estimation and causal inference. Eur Heart J. Oxford University Press. 2019; 40(20): 1597-1599.
3. DEFRA. Supplement to the UK plan for tackling roadside nitrogen dioxide concentrations. 2018(October), 1-54.
4. DfT and DEFRA. UK plan for tackling roadside nitrogen dioxide concentrations: detailed plan. Dep. Environ. Food Rural Aff. together with Dep Transp. 2017(July), 1-11.
5. Abdul Halim ND *et al.* The long-term assessment of air quality on an island in Malaysia. Heliyon. 2018; 4(12).
6. WHO Regional Office for Europe OECD. Economic cost of the health impact of air pollution in Europe: clean air, health and wealth. Eur Environ Heal Process. 2015, 1-54.
7. Myllyvirta L. Quantifying the economic costs of air pollution from fossil fuels key messages. 2020, 2-13.
8. Kopparapu R, Arney G., Haqq-Misra J., Lustig-Yaeger J., Villanueva G. Nitrogen dioxide pollution as a signature of extraterrestrial technology. Astrophys J. 2021; 908(2): 164.
9. Bauwens M *et al.* Impact of coronavirus outbreak on NO_2 pollution assessed using TROPOMI and OMI observations. Geophys Res Lett. 2020; 47(11): 1-9.
10. Alaka HA *et al.* Systematic review of bankruptcy prediction models: towards a framework for tool selection. Expert Syst Appl. 2018; 94: 164-184.
11. Alaka H, Oyedele L, Owolabi H, Akinade O, Bilal M, Ajayi S. Firms failure prediction models. IEEE Trans Eng Manag. 2018(4), 1-10.
12. Kamińska JA. A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions. Sci Total Environ. 2019; 651(2): 475-483.
13. Juhos I, Makra L, Tóth B. Forecasting of traffic origin NO and NO_2 concentrations by support vector machines and neural networks using principal component analysis. Simul Model Pract Theory. 2008; 16(9): 1488-1502.
14. Dou X *et al.* Estimates of daily ground-level NO_2 concentrations in China based on big data and machine learning approaches. 2020; arXiv(2).

15. Ardabili S, Mosavi A, Várkonyi-Kóczy AR. Advances in machine learning modeling reviewing hybrid and ensemble methods. Lect Notes Networks Syst. 2020; 101(August): 215-227.
16. Karballaeezadeh N, Mohammadzadeh SD, Shamshirband S, Hajikhodaverdikhan P, Mosavi A, wing Chau K. Prediction of remaining service life of pavement using an optimized support vector machine (case study of Semnan–Firuzkuh road). Eng Appl Comput Fluid Mechs. 2019; 13(1): 188-198.
17. Mustaffa Z, Yusof Y, Kamaruddin SS. Gasoline price forecasting: an application of LSSVM with improved ABC. Proced - Soc Behav Sci. 2014; 129: 601-609.
18. Tian Z. Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM. Eng Appl Artif Intell. 2020; 91(February): 103573.
19. Jia C, Li W, Wu T, He M. Road traffic and air pollution: evidence from a nationwide traffic control during coronavirus disease 2019 outbreak. Sci Total Environ. 2021.
20. Hafiz A, Lukumon O, Muhammad B, Olugbenga A, Hakeem O, Saheed A. Bankruptcy prediction of construction businesses: towards a big data analytics approach. Proc. - 2015 IEEE 1st Int Conf Big Data Comput Serv Appl BigDataService. 2015; 2015, 347-352.
21. Reddy GT *et al.* Analysis of dimensionality reduction techniques on big data. IEEE Access. 2020; 8: 54776-54788.
22. Deng H, Runger G. Gene selection with guided regularised random forest. Pattern Recognit. 2013; 46(12), 3483-3489.
23. Menze BH *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinf. 2009; 10: 1-16.
24. Dimitriadis SI, Liparas D, Tsolaki MN. Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: from the alzheimer's disease neuroimaging initiative (ADNI) data. J Neurosci Methods. 2018; 302: 14-23.
25. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw. 2010; 36(11): 1-13.
26. Choi J, Gu B, Chin S, Lee JS. Machine learning predictive model based on national data for fatal accidents of construction workers. Autom Constr. 2020; 110(May): 102974.
27. Purnus A, Bodea CN. A predictive model of contractor financial effort in transport infrastructure projects. Proced Eng. 2017; 196(June): 746-753.
28. Bilal M, yedele LO, Guidelines for applied machine learning in construction industry—a case of profit margins estimation, Adv Eng Informatics. 2020; 43(March) 2019, 101013.
29. Mehtab S, Sen J, Stock price prediction using convolutional neural networks on a multivariate timeseries. 2020.

Corresponding author
Hafiz Alaka can be contacted at: hafizalaka@outlook.com