Effect of traffic dataset on various machine-learning algorithms when forecasting air quality

Sulaimon, Ismail Abiodun

University of Hertfordshire

Alaka, Hafiz (Corresponding Author)

University of Hertfordshire

Olu-Ajayi, Razak

University of Hertfordshire

Ahmad, Mubashir

University of Hertfordshire

Ajayi, Saheed

Leeds Beckett University

Hye, Abdul
Rasuta Energy LTD

Abstract

Purpose

Road traffic emissions are generally believed to contribute immensely to air pollution, but the effect of road traffic datasets on air quality predictions has not been fully investigated. This research investigates the effects traffic dataset have on the performance of Machine Learning (ML) predictive models in air quality prediction.

Design/methodology/approach

To achieve this, we have set up an experiment with the control dataset having only the Air Quality (AQ) dataset and Meteorological (Met) dataset. While the experimental dataset is made up of the AQ dataset, Met dataset, and Traffic dataset. Several ML models (such as Extra Trees Regressor, eXtreme Gradient Boosting Regressor, Random Forest Regressor, K-Neighbors Regressor, and two others) were trained, tested, and compared on these individual combinations of datasets to predict the volume of $PM_{2.5}$, PM_{10}, NO_2, and O_3 in the atmosphere at various time of the day.

Findings

The result obtained showed that various ML algorithms react differently to the traffic dataset despite generally contributing to the performance improvement of all the ML algorithms considered in this study by at least 20% and an error reduction of at least 18.97%.

Research limitations/implications

This research is limited in terms of the study area and the result cannot be generalized outside of the UK as some of the inherent conditions may not be similar elsewhere. Additionally, only the ML algorithms commonly used in literature are considered in this research. Therefore, leaving out a few other ML algorithms.

Practical implications

This study reinforces the belief that the traffic dataset has a significant effect on improving the performance of air pollution ML prediction models. Hence, there is an indication that ML algorithms behave differently when trained with a form of traffic dataset in the development of an air quality prediction model. This implies that developers and researchers in air quality prediction need to identify the ML algorithms that behave in their best interest before implementation.

Originality/value

This will enable researchers to focus more on algorithms of benefit when using traffic datasets in air quality prediction.

Keywords: Air-Quality Prediction, Traffic Dataset, Big-Data, Machine Learning

1 Introduction

The urban population in the year 2050 is expected to have become 68% of the world population (UN DESA, 2018). The air quality of major cities around the world would be affected negatively if the forecast comes to pass and nothing is done to mitigate its impact on the environment. The World Health Organization (WHO) on their part have highlighted the effect of air pollution on various aspect of our lives as it is linked to about seven million annual deaths around the world. Additionally, 80% of the urban areas have air quality measures that are worse than the WHO guideline (WHO, 2014). The vulnerable group which includes children, the elderly, and people with respiratory and cardiovascular problems are the worse hit by the degrading air quality situation. Records have shown that in recent years, air pollution accounts for 1 out of 8 deaths globally (WHO, 2014). This situation emphasises the urgent need for a highly accurate air pollution Machine Learning (ML) prediction model.

The increase in the proportion of pollutants in the form of particles and inhabitable gasses in the atmosphere of an area leads to air pollution. Human activities as observed through transportation, industry, or domestic activities and sometimes biological or environmental activities such as the case of pollutants like ozone (O_3), pollen, dust contributes to the composition of the atmosphere that leads to air pollution. Even in developing countries, (Owusu-Manu et al., 2020) established air quality as one of the eight indicators for measuring green city development. Various air pollutants have been linked to critical health challenges such as cardiovascular diseases, pulmonary disease, acute respiratory infection and increased risk of lung cancer. The yield of crops has also been affected by the increasing concentration of O_3 in the atmosphere (Gul & Khan, 2020). Air pollutants such as Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Ozone are medically proven to irritate the airways of the lungs, increasing the symptoms of those suffering from lung diseases. Fine particles ($PM_{2.5}$ and PM_{10}) always find their way into the lungs to cause inflammation and a worse condition of heart and lung diseases. Carbon Monoxide (CO) is believed to inhibit the absorption of oxygen by the blood (DEFRA, 2021). These conditions can lead to a shortage of oxygen supply to the heart and most likely death.

(Rybarczyk & Zalakeviciute, 2018) reported research (such Lin et al., 2018; Martınez-Espana et al., 2018; Tao et al., 2019; Zhang & Ding, 2017; Zhao et al., 2016) that have reported significant performance gain in predicting many of the air pollutants. $PM_{2.5}$, PM_{10}, NO_2 and O_3 have been among the major pollutants of concern globally as they are linked to various health hazards. The particulate matters (PM) are worsening the situation recently and are getting into the limelight. An increase in exposure to air pollutants, (Bowatte et al., 2015; Power Melinda C. et al., 2011) have been able to link the air pollutant $PM_{2.5}$ to the risk of asthma across childhood up to twelve years of age and a decreased cognitive function in older men. The magnitude increases with age and the pattern is more prominent with $PM_{2.5}$. Many of these health hazards have been the driving force for governments around the world to redesign policies to reduce air pollution and its impact on the environment. Despite government efforts, pollution is yet to reduce to a significant low. Hence, the need to be aware of where and when pollution is high. So, individuals and organizations can be well equipped to make an informed decision about pollution in their respective environments.

As deterministic models struggle to capture the relationships between variables that affect air pollution, several implementations of ML algorithms ranging from the classical ML algorithms such as Support Vector Machine (SVM), Linear Regression (LR) and the sophisticated ML algorithms like Deep-Neural Network (DNN), and Extreme Machine Learning have been the trend (Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018). In the course of finding an optimally performing air quality ML prediction model, researchers have trained ML algorithms on varying combinations of datasets (Masih, 2019). Yet the selection of appropriate ML algorithms has been a challenge when using traffic datasets in developing air pollution ML prediction models.

This study aims to investigate the effect of traffic datasets on the performance of various ML algorithms used in the development of air pollution ML prediction models. To achieve this, we have selected air pollutants $PM_{2.5}$, PM_{10}, NO_2 and O_3; these have been reported to have some of the worse impacts on the human way of life globally. In the next section, recent literature in ML-based air quality prediction is reviewed. This is followed by the methodology section, where we explain the research methodology used in this study. The result analysis and discussion section present the outcome of the study. A conclusion of the findings of this research is recorded in the conclusion section.

2 Literature Review

Several recent research using ML (Alaka et al., 2018; Martınez-Espana et al., 2018; Tao et al., 2019; Zhang & Ding, 2017; Hellas et al., 2019; Jiang, 2019; Tu et al., 2020; Mane et al., 2020; Madeiros et al., 2019) have shown that deterministic models proved less efficient in the prediction problems in general and specifically in air pollution prediction, while ML algorithms are more promising in this domain as reported in the literature (Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018). As seen in our recent study (Sulaimon et al., 2021), improved data accessibility in recent years has enhanced the contribution of research in the domain of air pollution prediction. Hence, several research efforts with diverse approaches have been contributed. The state-of-the-art research in air pollution prediction using ML is summarized in Table 1. This explains a lot about the trend in the domain as there are several datasets used in the prediction of air pollution, researchers find Air Quality data and Meteorological data more useful. This has to some extent limit the performance of air pollution ML prediction models.

2.1 Systematic Literature Review Process

As with literature review studies (such as Alaka et al., 2018; Rybarczyk & Zalakeviciute, 2018; Egwim et al., 2021), we conducted a systematic review of the state-of-the-art research in this domain. The literature review is conducted to reveal the trends and assumptions in using traffic datasets in the development of air quality prediction models.

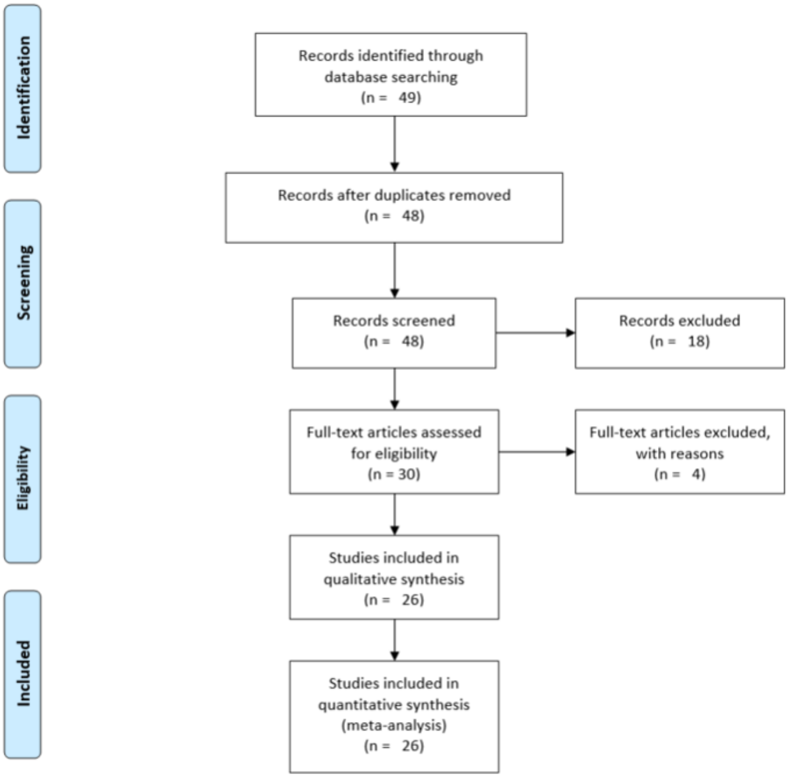
We searched the SCOPUS database using the search keyword- *(“air pollution” OR “air quality” OR “atmospheric pollution” OR “air pollutant”) AND (prediction OR forecast OR forecasting OR predict OR predicting) AND (“machine learning” OR ml OR “predictive model” OR modelling OR algorithm OR “big data” OR bigdata OR “artificial intelligence” OR ai)*. The keyword was developed through similar studies and common keywords in the domain, while the choice of the SCOPUS database is based on the benefits it offered by the unification of research globally and providing a comprehensive and wide range of scholarly information.

The search returned 49 relevant research articles between January 2007 and June 2021. The research articles were further screened using the following selection criterion:

- Publication must be published between the 1st of January 2000 and the 30th of May 2021.
- Publication must report on the use of ML models or algorithms in the prediction of air pollution.
- Publication must involve an empirical study to analyze the performance of the approach used and the result obtained.
- If some publications reported the same empirical studies, the most recent will be selected.

Research articles are excluded if:

- They do not meet the selection criterion, if
- They performed only a Systematic Literature Review, Systematic Mapping Study or General Literature Review.
- They do not conduct an empirical study.
- The full text is not available.
- They are written in languages other than English and without a translation to English.
- They are grey literature (such as technical papers, government reports, policy statements and issues papers, conference proceedings, pre-prints and post-prints of articles, theses and dissertations, research reports, geological and geophysical surveys, maps, newsletters and bulletins, fact sheets).


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 49)
Records after duplicates removed (n = 48)
Records screened (n = 48)
Records excluded (n = 18)
Full-text articles assessed for eligibility (n = 30)
Full-text articles excluded, with reasons (n = 4)
Studies included in qualitative synthesis (n = 26)
Studies included in quantitative synthesis (meta-analysis) (n = 26)

The 49 research articles discovered through our search were put to further scrutiny to ensure only quality research are included in the study. Our literature search flow diagram in Figure 3 shows that the literature identification stage ended up with 48 research articles as one of the articles is a duplicate. Hence, a copy of the article from the most reliable source is included for the study. At the screening stage, 18 articles are excluded after screening the abstract and ensuring that the selected articles were published within the defined period between 2017 and 2021. Despite the titles of some of the articles being related to our study, the abstract reflects that their objectives are very different from that of our study. While some other articles have no full text available for further study. Therefore, 30 research articles are remaining for the full-text analysis (eligibility) stage. The full-text analysis reveals that four of the articles deviate from the objective of our research, as one does not report the pollutants of focus, the other involved only a simulation study and the last two does not involve air pollution prediction.

Finally, only 26 of the research articles are selected for review, as they fulfil the selection criteria, and are the most recent in the domain as they were published between 2017 and 2021. The unselected articles either are out of scope, duplicates or have inaccessible full text. These primary study articles are further reviewed for data extraction, data synthesis and data analysis. A full report of the systematic review and result is found in (Sulaimon et al., 2021).

2.2 State-of-the-art in Air Quality Prediction

There are many pollutants of focus in air quality prediction using ML, but only six of them are the most common pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) among many research studies (Table 1, in appendix). This is partly due to the presence of these pollutants in almost everywhere there are human activities, while the rarely focused pollutants are found in the atmosphere of specific areas such as the insides of mines, tunnels and factories. Additionally, there are diverse combinations of datasets for the prediction of these air pollutants. The commonly used datasets are the Air Quality dataset and the Meteorological dataset. Most of the literature has a form of Air Quality dataset in their dataset collection used for air pollutant prediction. Despite the diversity observed in the choices for ML algorithms, Random Forest appears to be the most common choice of many researchers.

Not until recently have researchers started to use different forms of traffic datasets in air quality forecasts. This has been due to the recent availability of traffic datasets and the inherent belief induced through traffic emission modelling that simulate how traffic emission contributes to air pollution. Many studies like (Comert et al., 2020; Hatzopoulou et al., 2013; Pinto et al., 2020; Rossi et al., 2020) have supported the theory that traffic emission contributes to air pollution. This has caused several studies such as (Ashayeri et al., 2021; Rossi et al., 2019a; Rossi et al., 2020) in the domain of air quality forecast to use various forms of traffic datasets for prediction model development. Whereas, there is no significant research that proves the impact of traffic datasets on the performance of air quality ML prediction models. It is obvious in Table 1 (see appendix) that the performance measure reported in research that uses traffic datasets are not significantly better than the others, and it has not been verified if the traffic datasets are the contributors to the performance recorded. In addition, research reported does not use the same datasets, combination of datasets, methodologies nor ML algorithms. Hence, the results are not comparable to prove the assumption on how traffic datasets influence the predictive performance of air quality ML prediction models.

As traffic-related air pollution is a major contributor to air pollution as evident in traffic emission dispersion modelling, traffic data is believed to enable a dataset with high granularity and an ML predictive model with better performance. Hence, we verify through an experimental approach how traffic dataset influences the performance of ML algorithms in the development of air quality ML prediction models.

3 Methodology

Deterministic models have proved less efficient in the prediction of air pollutants (Iskandaryan et al., 2020; Liao et al., 2021), this research applies ML in the development of an air pollution ML prediction model. Hence, we implemented the full length of the ML pipeline, starting from Data preprocessing, Algorithm selection, Model training/validation/testing, Model evaluation.

As presented in Figure-1, the research begins with a systematic review of the state-of-the-art in air quality prediction. Simultaneously, data collection is ongoing to gather a sufficient amount of Air Quality dataset, Meteorological (Weather) dataset and Traffic dataset as needed. This is followed by the ML pipeline to develop the air quality prediction model. At the end of the air quality model development, the models are evaluated, results of the evaluations are analysed for comparison between models. The developed models are then prepared ready for use while a conclusion and recommendation are recorded.

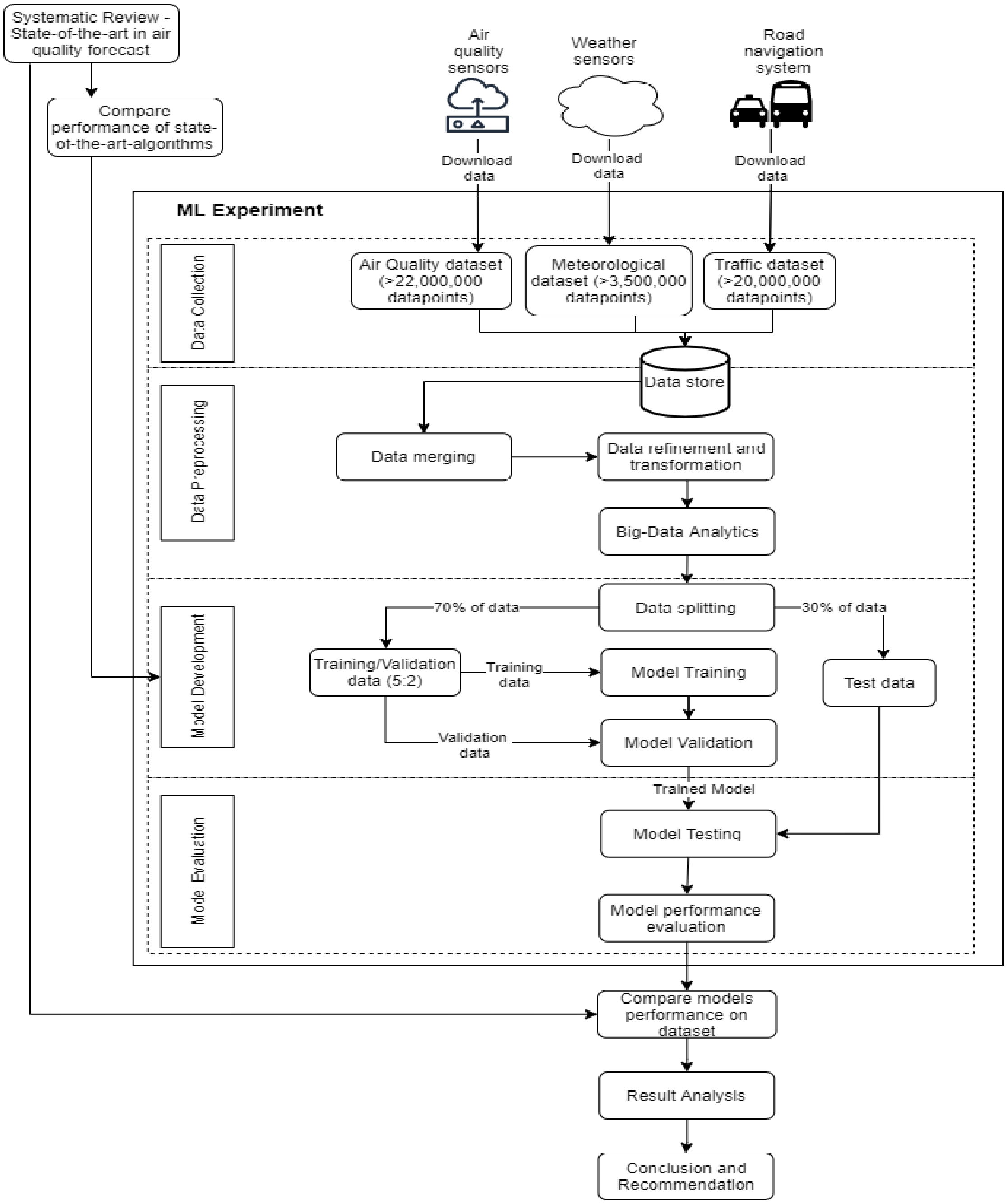

Systematic Review - State-of-the-art in air quality forecast
Compare performance of state-of-the-art-algorithms
Air quality sensors
Weather sensors
Road navigation system
Download data
Download data
Download data
ML Experiment
Data Collection
Air Quality dataset (>22,000,000 datapoints)
Meteorological dataset (>3,500,000 datapoints)
Traffic dataset (>20,000,000 datapoints)
Data store
Data Preprocessing
Data merging
Data refinement and transformation
Big-Data Analytics
Model Development
Data splitting
70% of data
30% of data
Training/Validation data (5:2)
Training data
Model Training
Test data
Validation data
Model Validation
Trained Model
Model Evaluation
Model Testing
Model performance evaluation
Compare models performance on dataset
Result Analysis
Conclusion and Recommendation

Figure 1: Research Methodology Block-Diagram

3.1 Study area

The study area for this research is the United Kingdom (UK), which is comprised of England, Wales, Scotland and Northern Ireland. The UK is the 21st most populated country in the world with a population of 68.2 million and a population density of 259 people per square kilometre. The predominantly urban and suburban England's southeast region houses about a third of the population of the UK (KS3 Geography Revision, 2021). This explains the higher density of dataset collection sites in the southeastern part of England as seen in Figure-2.

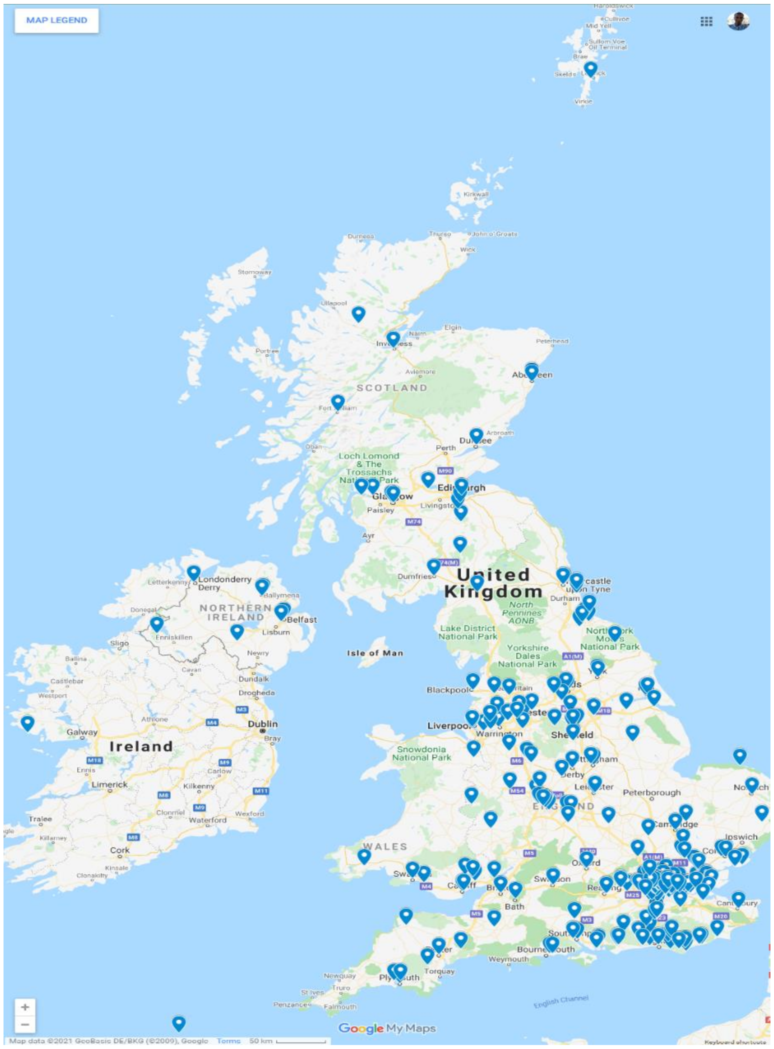


Figure 2: 338 Data Collection site across the United Kingdom

Data was collected from 338 different locations across the UK as shown in the identified location on the map in Figure-2. The data collected is made up of 11,322,240 data points of the Air Quality dataset, Meteorological dataset and Traffic dataset.

3.2 Data

In the process of applying ML to solving a problem, the availability of data, the volume and quality of the available data are important metrics to be considered. To ensure an optimal combination of high-quality dataset and selection of features, we documented and review all the stages of Data collection, Data merging, Data preprocessing, Data visualization and analysis, Feature selection.

3.2.1 Data collection

The Internet-of-Things (IoT) has simplified the process of data collection and monitoring in many fields (Arowoiya et al., 2020; Gamil et al., 2020). The data collection process for this research was conducted using IoT-based sensors within the 338 data collection sites selected within the UK. This was done with a focus on ensuring high granularity and accuracy of our model prediction. Therefore, the datasets collected include the meteorological (weather) dataset, air quality dataset and traffic dataset for 338 postcodes spread around the UK. For this research, we used a dataset for the 6 months starting from 01:00 am, 1st of December 2020, till 00:00 1st of June 2021. The period was selected as it falls within the period when the covid-19 lockdown was eased in the UK. Additionally, it enables capturing datasets for multiple seasons with variable weather conditions.

Air Quality Data: The advent of IoT-based technology has made the recording of Air Quality (AQ) data easier. The air quality data used in this research is sourced via the 338 IoT sensors stationed across the UK. The sensors record hourly data for several pollutants (such as NO_2, PM_{10}, SO_2, FINE, O_3, PM_1, $PM_{2.5}$, TSP, CO, NO). Each of the sensors does not record readings for all the pollutants. Hence, the recordings for various pollutants were collated from several sensors. The Air Quality Data is recorded hourly by each of the 338 sensors and stored using the AWS Document DB that supports MongoDB. A description of the air quality dataset is seen in Table 2 (see appendix)

Meteorological Data: The recording of the meteorological (Met) data was conducted for each of the data collection sites selected in the UK. This data was sourced via the Open Weather Map API (OpenWeatherMap, 2021). A JSON format object is used for the API response, while the data is extracted and stored on the Amazon Web Service (AWS) S3 bucket to ease accessibility. Table 3 (in appendix) contains a brief description of the meteorological (weather) dataset.

Traffic Data: The process of getting the traffic data has been challenging, as several traffic data sources do not have it readily available while many of the sources do not have the data recorded hourly nor at the level of granularity needed. Hence, for this research, we sourced traffic data from TomTom traffic flow API (Tom Tom, 2021). The traffic data is collected at 15 minutes intervals, but it is further analysed and processed to derive an hourly average of every day of the week. This process involves the elimination of the data recorded for the period immediately after the COVID-19 lockdown was eased in the UK (between the 8th of March 2021 and the 22nd of March 2021). The purpose of the data elimination was due to the irregular spike in the traffic situation across the UK roads during the stated period. The distortion introduced into the data is eliminated through the observed procedure. A summary of the traffic dataset is presented in Table 4 (in appendix).

3.2.2 Data preprocessing

The data preprocessing phase of the ML pipeline is a need-based process. Hence, we performed the preprocessing needed for the datasets available for the research. Due to the large size of the dataset used in this research, the dataset is stored and processed as big data. The dataset is stored using the Amazon Web Service (AWS) S3 bucket and the AWS DocumentDB. This allows for flexibility in processing the data via the AWS cloud.

Data merging: Using multiple datasets require the process of data merging to ensure that the required data can be sourced from a unified repository. The three datasets used for this research were processed and merged using the Pandas package in the Python programming language. The data were merged using their common feature of Location (Latitude, Longitude), Date and Time.

Data Refinement: The refinement applied to the data includes the process of cleaning the data of outliers and treating the missing data. The records with missing pollutants values are eliminated to avoid ambiguity in the data modelling phase and the model training and testing phase.

Data Transformation: All the features of the dataset are transformed into numerical and float data types. This is to enable easy computation.

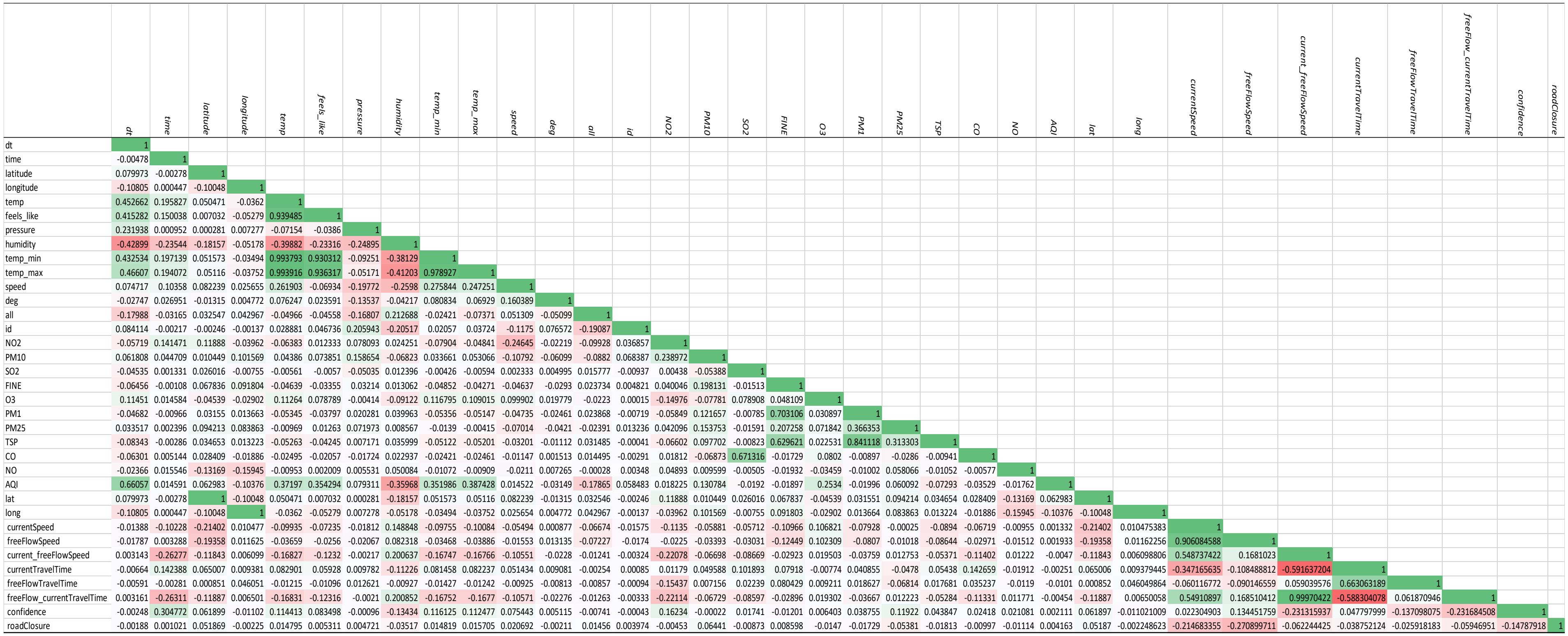
3.2.3 Feature selection

In selecting the features to be used for training the model, we ensured that only the important features of the dataset are selected. The Air Quality, Meteorological and Traffic datasets consists of numerous features that are not related to the study. Hence, these features are removed from the dataset for the training process. Table 5 describes the selected features of the dataset after it has been merged. The

Location (Latitude, Longitude), Date and Time features of the dataset are used only to merge the dataset. These features were removed before the dataset is used in the ML pipeline. The correlation analysis that guides the feature selection process of the dataset is presented in Figure-3.

Table 1: Selected variables for model development

Data	**Features**	**Unit**	**Type**
Location	Latitude		
	Longitude		
Date	Date		
Time	Time	Hour	
Meteorological	Absolute Temperature	^{o}c	Independent
	Feels-Like	^{o}c	
	Pressure	Hg	
	Humidity	%	
	Minimum Temperature	^{o}F	
	Maximum Temperature	^{o}F	
	Wind Speed	km/h	
	Wind Degree	degree	
	Cloud	okta	
	Rainfall	mm	
Traffic	Current Travel Speed	km/h	
	Free Flow Travel Speed	km/h	
	The ratio of Current Travel Speed and Free Flow Travel Speed		
	Current Travel Time	Seconds	
	Free Flow Travel Time	Seconds	
	The ratio of Current Travel Time and Free Flow Travel Time		
	Data Confidence		
	Road Closure	0,1,2,3,4	
Air Quality	PM$_{2.5}$	µg/m^3	Dependent
	PM$_{10}$	µg/m^3	
	NO$_2$	µg/m^3	
	O$_3$	µg/m^3	



	dt	time	latitude	longitude	temp	feels_like	pressure	humidity	temp_min	temp_max	speed	deg	all	id	NO2	PM10	SO2	FINE	O3	PM1	PM25	TSP	CO	NO	AQI	lat	long	currentSpeed	freeFlowSpeed	current_freeFlowSpeed	currentTravelTime	freeFlowTravelTime	freeFlow_currentTravelTime	confidence	roadClosure
dt	1																																		
time	-0.00478	1																																	
latitude	0.079973	-0.00278	1																																
longitude	-0.10805	0.000447	-0.10048	1																															
temp	0.452662	0.195827	0.050471	-0.0362	1																														
feels_like	0.415282	0.150038	0.007032	-0.05279	0.939485	1																													
pressure	0.231938	0.000952	0.000281	0.007277	-0.07154	-0.0386	1																												
humidity	-0.42899	-0.23544	-0.18157	-0.05178	-0.39882	-0.23316	-0.24895	1																											
temp_min	0.432534	0.197139	0.051573	-0.03494	0.993793	0.930312	-0.09251	-0.38129	1																										
temp_max	0.46607	0.194072	0.05116	-0.03752	0.993916	0.936317	-0.05171	-0.41203	0.978927	1																									
speed	0.074717	0.10358	0.082239	0.025655	0.261903	-0.06934	-0.19772	-0.2598	0.275844	0.247251	1																								
deg	-0.02747	0.026951	-0.01315	0.004772	0.076247	0.023591	-0.13537	-0.04217	0.080834	0.06929	0.160389	1																							
all	-0.17988	-0.03165	0.032547	0.042967	-0.04966	-0.04558	-0.16807	0.212688	-0.02421	-0.07371	0.051309	-0.05099	1																						
id	0.084114	-0.00217	-0.00246	-0.00137	0.028881	0.046736	0.205943	-0.20517	0.02057	0.03724	-0.1175	0.076572	-0.19087	1																					
NO2	-0.05719	0.141471	0.11888	-0.03962	-0.06383	0.012333	0.078093	0.024251	-0.07904	-0.04841	-0.24645	-0.02219	-0.09928	0.036857	1																				
PM10	0.061808	0.044709	0.010449	0.101569	0.04386	0.073851	0.158654	-0.06823	0.033661	0.053066	-0.10792	-0.06099	-0.0882	0.068387	0.238972	1																			
SO2	-0.04535	0.001331	0.026016	-0.00755	-0.00561	-0.0057	-0.05035	0.012396	-0.00426	-0.00594	0.002333	0.004995	0.015777	-0.00937	0.00438	-0.05388	1																		
FINE	-0.06456	-0.00108	0.067836	0.091804	-0.04639	-0.03355	0.03214	0.013062	-0.04852	-0.04271	-0.04637	-0.0293	0.023734	0.004821	0.040046	0.198131	-0.01513	1																	
O3	0.11451	0.014584	-0.04539	-0.02902	0.11264	0.078789	-0.00414	-0.09122	0.116795	0.109015	0.099902	0.019779	-0.0223	0.00015	-0.14976	-0.07781	0.078908	0.048109	1																
PM1	-0.04682	-0.00966	0.03155	0.013663	-0.05345	-0.03797	0.020281	0.039963	-0.05356	-0.05147	-0.04735	-0.02461	0.023868	-0.00719	-0.05849	0.121657	-0.00785	0.703106	0.030897	1															
PM25	0.033517	0.002396	0.094213	0.083863	-0.00969	0.01263	0.071973	0.008567	-0.0139	-0.00415	-0.07014	-0.0421	-0.02391	0.013236	0.042096	0.153753	-0.01591	0.207258	0.071842	0.366353	1														
TSP	-0.08343	-0.00286	0.034653	0.013223	-0.05263	-0.04245	0.007171	0.035999	-0.05122	-0.05201	-0.03201	-0.01112	0.031485	-0.00041	-0.06602	0.097702	-0.00823	0.629621	0.022531	0.841118	0.313303	1													
CO	-0.06301	0.005144	0.028409	-0.01886	-0.02495	-0.02057	-0.01724	0.022937	-0.02421	-0.02461	-0.01147	0.001513	0.014495	-0.00291	0.01812	-0.06873	0.671316	-0.01729	0.0802	-0.00897	-0.0286	-0.00941	1												
NO	-0.02366	0.015546	-0.13169	-0.15945	-0.00953	0.002009	0.005531	0.050084	-0.01072	-0.00909	-0.0211	0.007265	-0.00028	0.00348	0.04893	0.009599	-0.00505	-0.01932	-0.03459	-0.01002	0.058066	-0.01052	-0.00577	1											
AQI	0.66057	0.014591	0.062983	-0.10376	0.37197	0.354294	0.079311	-0.35968	0.351986	0.387428	0.014522	-0.03149	-0.17865	0.058483	0.018225	0.130784	-0.0192	-0.01897	0.2534	-0.01996	0.060092	-0.07293	-0.03529	-0.01762	1										
lat	0.079973	-0.00278	1	-0.10048	0.050471	0.007032	0.000281	-0.18157	0.051573	0.05116	0.082239	-0.01315	0.032546	-0.00246	0.11888	0.010449	0.026016	0.067837	-0.04539	0.031551	0.094214	0.034654	0.028409	-0.13169	0.062983	1									
long	-0.10805	0.000447	-0.10048	1	-0.0362	-0.05279	0.007278	-0.05178	-0.03494	-0.03752	0.025654	0.004772	0.042967	-0.00137	-0.03962	0.101569	-0.00755	0.091803	-0.02902	0.013664	0.083863	0.013224	-0.01886	-0.15945	-0.10376	-0.10048	1								
currentSpeed	-0.01388	-0.10228	-0.21402	0.010477	-0.09935	-0.07235	-0.01812	0.148848	-0.09755	-0.10084	-0.05494	0.000877	-0.06674	-0.01575	-0.1135	-0.05881	-0.05712	-0.10966	0.106821	-0.07928	-0.00025	-0.0894	-0.06719	-0.00955	0.001332	-0.21402	0.010475383	1							
freeFlowSpeed	-0.01787	0.003288	-0.19358	0.011625	-0.03659	-0.0256	-0.02067	0.082318	-0.03468	-0.03886	-0.01553	0.013135	-0.07227	-0.0174	-0.0225	-0.03393	-0.03031	-0.12449	0.102309	-0.0807	-0.01018	-0.08644	-0.02971	-0.01512	0.001933	-0.19358	0.01162256	0.906084588	1						
current_freeFlowSpeed	0.003143	-0.26277	-0.11843	0.006099	-0.16827	-0.1232	-0.00217	0.200637	-0.16747	-0.16766	-0.10551	-0.0228	-0.01241	-0.00324	-0.22078	-0.06698	-0.08669	-0.02923	0.019503	-0.03759	0.012753	-0.05371	-0.11402	0.01222	-0.0047	-0.11843	0.006098806	0.548737422	0.1681023	1					
currentTravelTime	-0.00664	0.142388	0.065007	0.009381	0.082901	0.05928	0.009782	-0.11226	0.081458	0.082237	0.051434	0.009081	-0.00254	0.00085	0.01179	0.049588	0.101893	0.07918	-0.00774	0.040855	-0.0478	0.05438	0.142659	-0.01912	-0.00251	0.065006	0.009379445	-0.347165635	-0.108488812	-0.591637204	1				
freeFlowTravelTime	-0.00591	-0.00281	0.000851	0.046051	-0.01215	-0.01096	0.012621	-0.00927	-0.01427	-0.01242	-0.00925	-0.00813	-0.00857	-0.00094	-0.15437	0.007156	0.02239	0.080429	0.009211	0.018627	-0.06814	0.017681	0.035237	-0.0119	-0.0101	0.000852	0.046049864	-0.060116772	-0.090146559	0.059039576	0.663063189	1			
freeFlow_currentTravelTime	0.003161	-0.26311	-0.11887	0.006501	-0.16831	-0.12316	-0.0021	0.200852	-0.16752	-0.1677	-0.10571	-0.02276	-0.01263	-0.00333	-0.22114	-0.06729	-0.08597	-0.02896	0.019302	-0.03667	0.012223	-0.05284	-0.11331	0.011771	-0.00454	-0.11887	0.00650058	0.54910897	0.168510412	0.99970422	-0.588304078	0.061870946	1		
confidence	-0.00248	0.304772	0.061899	-0.01102	0.114413	0.083498	-0.00096	-0.13434	0.116125	0.112477	0.075443	0.005115	-0.00741	-0.00043	0.16234	-0.00022	0.01741	-0.01201	0.006403	0.038755	0.11922	0.043847	0.02418	0.021081	0.002111	0.061897	-0.011021009	0.022304903	0.134451759	-0.231315937	0.047797999	-0.137098075	-0.231684508	1	
roadClosure	-0.00188	0.001021	0.051869	-0.00225	0.014795	0.005311	0.004721	-0.03517	0.014819	0.015705	0.020692	-0.00211	0.01456	0.003974	-0.00453	0.06441	-0.00873	0.008598	-0.0147	-0.01729	-0.05381	-0.01813	-0.00997	-0.01114	0.004163	0.05187	-0.002248623	-0.214683355	-0.270899711	-0.062244425	-0.038752124	-0.025918183	-0.05946951	-0.14787918	1

Figure 3: Correlation analysis of all features

This has informed the selection of features of the dataset that have the least correlations and are of most concern to this research as seen in Table 5 (of the appendix).

3.3 Algorithm selection

The approach to this study is to use ML regressor algorithms for the development of the ML prediction model. Hence, we trained 40 ML regressor algorithms and report the top six models based on their performance recorded using the performance metrics stated. The reported algorithms are Extra Trees Regressor, Histogram-Based Gradient Boosting Regressor, Light Gradient Boosted Machine (LGBM) Regressor, eXtreme Gradient Boosting (XGB) Regressor, Random Forest Regressor, Bagging Regressor, Gradient Boosting Regressor, K-Neighbors Regressor (KNN), Multi-Layer Perceptron (MLP) Regressor. These ML algorithms are trained on the merged datasets using their default hyper-parameters to ensure a fair comparison of the performance of the algorithms.

3.4 Performance metrics

The choice of performance metrics in evaluating the predictive models is based on the commonly used performance metrics found in related literature. This will enable easy comparison of our results with the state-of-the-art benchmarks. Many of the literature reported in reviews (such as Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018) evaluate their predictive models using R-Squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Hence, these two measures of performance are selected as the performance metrics for this study.

3.4.1 R-Squared (R^2) score

The performance score R^2 shows the degree to which the data fit the model. It is used to determine the proportion of the variance in the dependent variable(s) that is justifiable by the independent variables. The best R^2 score is 1.0, but the value of R^2 can range from negative values. When the value of R^2 is negative, it explains that the fit of the variance is worse than just fitting a horizontal line. The R^2 is defined in equation 1:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{data,i}\right)^2}{\sum_{i=1}^{n}\left(y_{data,i} - y_{data}\right)^2} \qquad (1)$$

Where

$$y_{pred,i} = Predicted\ Value$$
$$y_{data,i} = Actual\ Value y_{data} = Mean\ Actual\ Value n = Number\ of\ data\ objects$$

3.4.2 Root Mean Squared Error (RMSE)

The RMSE is the root of the Mean Squared Error (MSE). Hence, it is the root of the variances between the predicted value and the actual value. As a measure of error, the value of RMSE ranges between 0 and 1. While the lower the value, the better the performance of the model. The RMSE is formulated as seen in equation 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(AE_i - PE_i)^2} \qquad (2)$$

Where

$$AE_i = Actual\ Value$$
$$PE_i = Predicted\ Value$$
$$n = Number\ of\ data\ objects$$

3.4.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) computes the variation between the predicted values and true or actual values. MAE scores close to zero indicates better performance while a score greater than zero signifies a bad performance. Unlike RMSE, MAE is not sensitive to outliers (Hyndman & Koehler, 2006).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |AE_i - PE_i| \tag{3}$$

Where

$$AE_i = Actual\ Value$$
$$PE_i = Predicted\ Value$$
$$n = Number\ of\ data\ objects$$

3.5 Design of Experiments

We designed the experiment to have the control and the experiment. The control group ML models of the experiment are designed to be trained on the merged Air Quality dataset and Meteorological dataset. While the Experimental group of the ML models are developed using the merged dataset consisting of the Air Quality dataset, Meteorological dataset and Traffic dataset (Table 6, in appendix).

Table 2: Experiment design

Study	Dataset
Control	AQ, Met dataset
Experiment	AQ, Met, Traffic dataset

The design of the experiment as summarized in Figure-4 is made up of four phases- Data collection, Data preparation, Model development and Model evaluation. Aside from the first phase of Data collection, every other phase is similar for each of the control and the experiment.

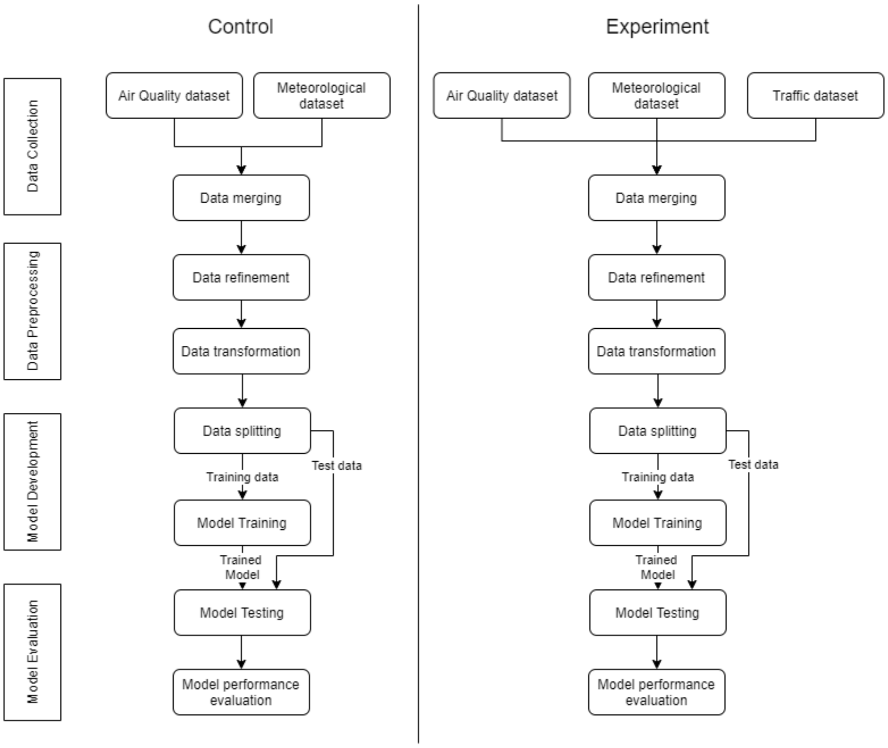


Figure 4: Design of the Experiment

3.5.1 Experimental setup

The experiment was conducted using Python programming language 3.6.8 (van Rossum & Drake, 1995) with the support of the Scikit-Learn Python Package (Pedregosa et al., 2012). The computation was performed using Apple MacBook Air with BigSur OS version 11.3 and an M1 processor with 16GB RAM. This configuration is in no way believed to impact the performance of the predictive models developed.

3.6 Model development

Following the experiment procedure, the model development stage is divided into two similar procedures for the control-dataset and the experimental-dataset respectively. The only difference between the two procedures is the dataset used in developing the predictive models (Figure 2). The models were developed to predict the independent variables ($PM_{2.5}$, PM_{10}, NO_2 and O_3) for each of the 338 locations at the various hour of the day.

For each of the experiment procedures, the dataset is randomly divided into a training dataset and a test dataset. 70% of the dataset is used for the training and validation, while the other 30% is used as the test dataset.

3.6.1 Training and validation

In the training phase of the ML pipeline, the selected ML algorithms are trained on the training dataset. Each of the algorithms is trained one after the other to ensure that the training process for each algorithm does not interfere in any way with the others.

The validation process is necessary while training the algorithm to ensure that the models improve in performance while undergoing training. Hence, the 10-fold cross-validation was done to ensure that the performance of the trained model is accurately recorded, and the model is well fitted.

3.6.2 Testing

For the performance evaluation purpose, testing the trained model is a vital step. The trained models are tested using the test dataset as provided from the main dataset. Here, the predictive models are used to predict the values of the dependent variable. These predictions are compared with the real value and the differences are recorded to be used for computing the evaluation metrics.

4 Result evaluation and discussion

Evaluating the model is based on the performance metrics used for the evaluation. The R^2 and RMSE have been selected to be used in this research due to their popularity in the research domain.

4.1 Evaluation

The evaluation process was conducted by using the predictive models to predict the dependent variables of the test data. The prediction is compared with the true values and used to derive the values of the performance metrics. This procedure is applied on each of the air pollutants of focus and recorded as presented below respectively.

4.1.1 $PM_{2.5}$

Table 7 presents the calculated R^2 score, RMSE and MAE for the predictive models trained on the control dataset and experimental dataset of $PM_{2.5}$. With a 20.81% increase in the performance recorded for the Bagging regressor when trained with the experiment dataset, the trend noticed for the performance boost of $PM_{2.5}$ prediction models appears interesting. The Bagging regressor is not the best performing algorithm in this category, but it displays the most performance boost (20.81%) when it is trained with the dataset that contains the traffic dataset. While the Random Forest Regressor had the largest percentage error reduction (18.97%) when trained experiment dataset.

Table 3: Performance of PM$_{2.5}$ Models

Model	Control			Experiment			% Difference		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Extra Trees Regressor	**0.6221**	**4.8191**	**3.0167**	0.7356	4.0311	**2.4445**	18.24	16.35	**18.97**
Random Forest Regressor	0.6214	4.8236	3.0580	**0.7370**	**4.0199**	2.5098	18.61	**16.66**	17.93
XGB Regressor	0.6048	4.9281	3.2270	0.7218	4.1349	2.6752	19.34	16.10	17.10
Histogram Gradient Boosting Regressor	0.5906	5.0158	3.3309	0.7005	4.2897	2.8068	18.62	14.48	15.73
LGBM Regressor	0.5898	5.0205	3.3367	0.6979	4.3086	2.8112	18.32	14.18	15.75
Bagging Regressor	0.5892	5.0243	3.1718	0.7118	4.2083	2.6199	**20.81**	16.24	17.40

4.1.2 PM$_{10}$

The calculated R^2 score, RMSE and MAE as recorded in Table 8 shows the performance for the predictive models trained on the control dataset and experimental dataset of PM$_{10}$. In the prediction of PM$_{10}$, the Extra Trees regressor records the best performance boost of 70% when measured using the R^2. Based on the MAE score, 27.77% is the largest error reduction recorded, and this is obtained by the Random Forest regressor. This is to reinforce the point that the best performing algorithm does not necessarily record the best improvement when the traffic dataset is added to the training dataset.

Table 4: Performance of PM$_{10}$ Models

Model	Control			Experiment			% Difference		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
LGBM Regressor	**0.4140**	**9.8203**	6.2937	0.6788	7.2707	4.9348	63.95	25.96	18.97
Histogram Gradient Boosting Regressor	0.4134	9.8255	6.2923	0.6728	7.3387	4.9527	62.74	25.31	21.29
XGB Regressor	0.4110	9.8455	**6.1002**	**0.6866**	**7.1815**	4.6899	67.06	**27.06**	23.12
Random Forest Regressor	0.4021	9.9199	6.1163	0.6629	7.4487	**4.4178**	64.86	24.91	**27.77**
Extra Trees Regressor	0.3846	10.0639	6.1727	0.6538	7.5483	4.4822	**70.00**	25.00	27.39
Bagging Regressor	0.3818	10.0863	6.2878	0.6336	7.7654	4.6895	20.81	16.24	17.40

4.1.3 NO$_2$

Like other pollutants reported in this study, the performance metrics for the predictive models recorded for NO$_2$ in the control and experimental dataset are slightly different. The calculated R^2 score and RMSE for the models trained on the control dataset and experimental dataset of NO$_2$ are recorded in Table 9 below. In a similar trend with other prediction models, the Extra Trees regressor reported the best performance in both the control dataset and the experiment dataset. Whereas the Extra Trees regressor which is an ensemble of many decision trees shows the largest performance increase of 46.82% and error reduction of 31.13% as recorded by the R^2 and MAE respectively.

Table 5: Performance of NO_2 Models

Model	Control			Experiment			% Difference		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
XGB Regressor	**0.5266**	**12.8396**	9.4102	**0.7407**	**9.5023**	6.6886	40.67	25.99	28.92
Random Forest Regressor	0.5242	12.8715	**9.3416**	0.7191	9.8906	6.6648	37.17	23.16	28.66
LGBM Regressor	0.5007	13.1855	9.8240	0.7053	10.1301	7.2517	40.86	23.17	26.18
Bagging Regressor	0.5002	13.1914	9.5231	0.6866	10.4458	7.0351	37.26	20.81	26.13
Histogram Gradient Boosting Regressor	0.4967	13.2375	9.8343	0.6961	10.2868	7.3742	40.13	22.29	25.02
Extra Trees Regressor	0.4937	13.2773	9.5058	0.7249	9.7881	**6.5467**	**46.82**	**26.28**	**31.13**

4.1.4 O_3

The performance metrics recorded for O_3 predictive models using R^2 score, RMSE and MAE are presented in Table 10. As with other predictive models in this study, the models are trained on the control dataset and experimental dataset of O_3, and the performance metrics are measured and recorded. The Extra Tree regressor consistently records the best performance boost and the best error reduction in the experiment.

Table 6: Performance of O_3 Models

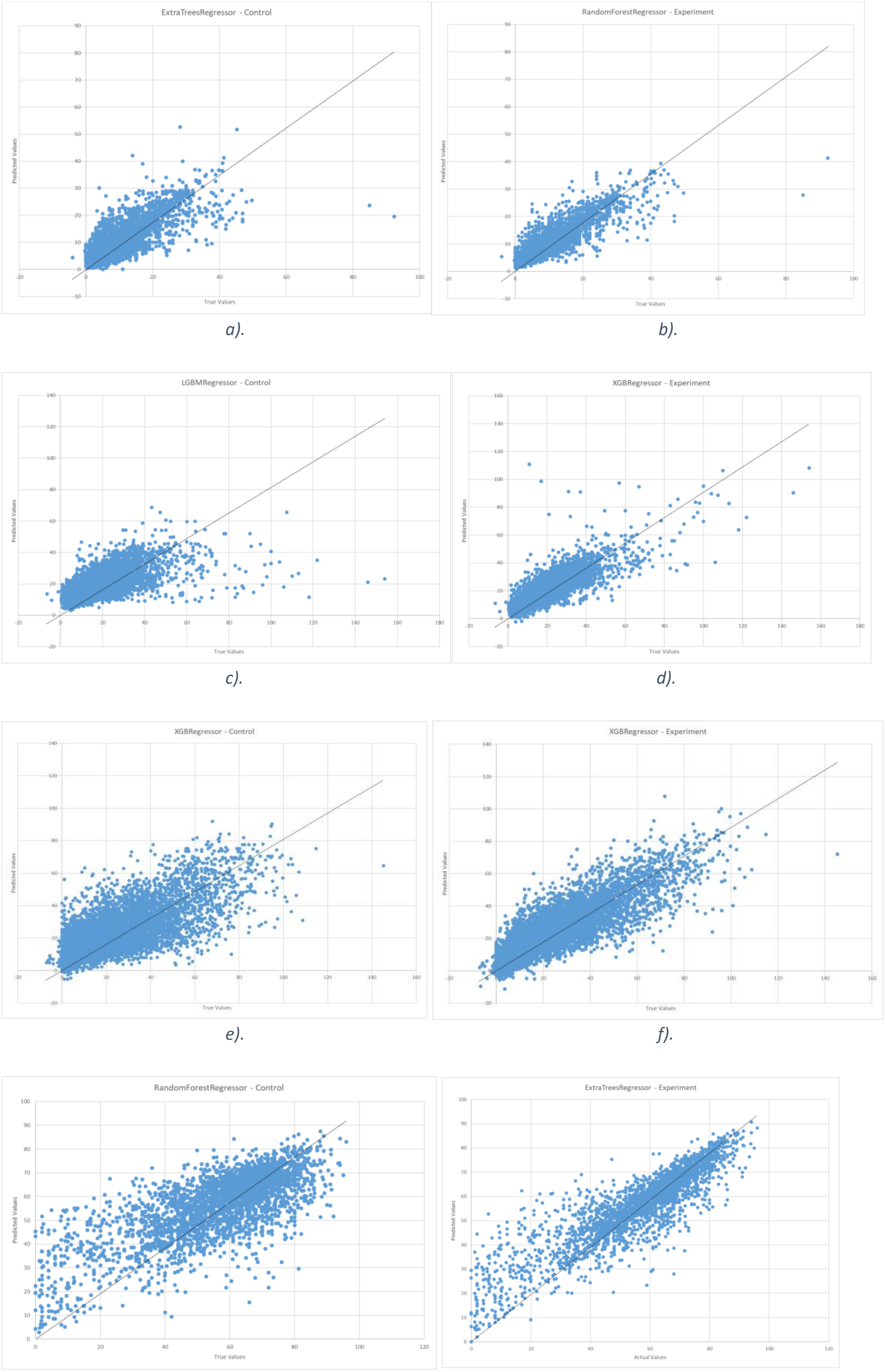
Model	Control			Experiment			% Difference		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest Regressor	**0.5305**	**12.8737**	9.4958	0.7415	9.5518	6.5092	39.78	25.80	31.45
LGBM Regressor	0.4981	13.3106	9.9812	0.6984	10.3189	7.4043	40.21	22.48	25.82
Extra Trees Regressor	0.4972	13.3225	**9.6159**	**0.7718**	**8.9760**	**5.8974**	**55.22**	**32.63**	**38.67**
Histogram Gradient Boosting Regressor	0.4942	13.3621	10.0266	0.6988	10.3111	7.4107	41.40	22.83	26.09
XGB Regressor	0.4942	13.3627	9.8844	0.7128	10.0690	7.0702	44.24	24.65	28.47
Bagging Regressor	0.4921	13.3906	9.8410	0.7034	10.2316	6.9975	42.96	23.59	28.90

4.2 Discussion

Traffic data has been in use in vehicle emission modelling for a long time as proven in (Comert et al., 2020; Hatzopoulou et al., 2013; Pinto et al., 2020; Rossi et al., 2020) and similar studies. The impact of traffic data on air quality ML prediction models has been assumed in several studies such as (Ashayeri et al., 2021; Rossi et al., 2019a; Rossi et al., 2020), where it is undoubtedly believed that traffic pollution contributes immensely to air pollution. This alone is not enough to explain how the traffic dataset influences the performance of air quality ML prediction models.

Observing the R^2 and RMSE performance metrics recorded for each of the pollutants $PM_{2.5}$, PM_{10}, NO_2 and O_3, the performance measured for each of the predictive models consistently indicates that the models trained on the control dataset perform worse than the models trained on the experimental datasets. As presented in Figure 5a, Extra Trees Regressor has the best performance recorded for $PM_{2.5}$ prediction models trained with the control dataset. Whereas it does not seem to perform much better with the experiment dataset. The Random Forest Regressor display a relatively outstanding performance boost when trained with the experimental dataset (Figure 5b). PM_{10} prediction models have similar behaviour to that of $PM_{2.5}$ prediction models, despite the variation in best-performing algorithms. The performance boost obtained in the PM_{10} experiment as presented in Figure 3d is significant when compared with the performance shown in Figure 3c which is obtained when the models are trained with the control dataset. Here, LGBM Regressor is best performing on the control dataset while the XGB Regressor obtained the best performance as measured for models trained with the experiment dataset. The prediction comparison presented in Figures 5e and 5f shows a consistent performance boost noticed for NO_2. XGB Regressor models obtained the best performance when trained with the control dataset, a relative boost in performance is also obtained when trained with the experiment dataset. As with $PM_{2.5}$ and PM_{10}, a significant boost in performance is obtained for O_3 models trained with the experiment dataset. Random forest Regressor obtained the best performance among models trained with the control dataset while the performance recorded for the experiment dataset is not significant enough to be the best performing model in this category (Figures 5g and 5h). The Random Forest Regressor is outperformed by the Extra Trees Regressor among the models trained with the experiment dataset.

Although the best-performing algorithms varies generally for each case, it is of immense benefit to have a performance gain in the air quality ML prediction models regardless of the algorithms used. This indicates that there is no one-size-fits-all situation to the best performing ML algorithms for building an air quality ML prediction model. There is limited research in the domain that compared the performance of ML algorithms developed with and without traffic datasets. Hence, there is no known research so far to compare the performance gain obtained in this work.



a).

b).

c).

d).

e).

f).

g).

h).

Figure 5. Model performance on Control and experimental datasets for various environmental parameters: a): PM2.5 - Control Dataset, b): PM2.5 - Experiment Dataset, c): PM10 - Control Dataset, d): PM10 - Experiment Dataset, e): NO2 - Control Dataset, f): NO2 - Experiment Dataset, g): O3 - Control Dataset, and h): O3 - Experiment Dataset

The patterns in the performance metrics recorded have reinforced the inherent belief that traffic datasets improve the performance of the Air Quality ML prediction model. This further validates the theory in (Comert et al., 2020; Hatzopoulou et al., 2013; Pinto et al., 2020; Rossi et al., 2020) and related literature that traffic emission has a major effect on air pollution modelling. Additionally, various ML algorithms react differently in the presence of traffic datasets. Hence, researchers in air quality prediction need to be selective in the process of choosing ML algorithms when their dataset contains a form of traffic dataset. Although the performance varied from one pollutant to the other, the overall performance record shows that models trained with the experimental data consisting of traffic dataset reported better performance when compared with the models trained on the control dataset with no traffic dataset.

This study implies that it reinforces the understanding that the traffic dataset has a significant effect on improving the performance of air pollution ML prediction models. Hence, there is an indication that ML algorithms behave differently when trained with a form of traffic dataset in the development of an air quality prediction model. This practically implies that developers and researchers in air quality prediction need to identify the ML algorithms that behave in their best interest before implementation. This research is limited in terms of the study area and the result cannot be generalized outside of the UK as many conditions may not be similar elsewhere. Additionally, only the ML algorithms commonly used in literature are considered in this research. Therefore, leaving out a few other ML algorithms.

5 Conclusion

In this research, we have investigated what impact traffic datasets have on the predictive performance of various air pollution ML prediction models. To achieve this, we have set up an experiment with the control dataset having only the AQ dataset and Met dataset. While the experimental dataset is made up of the AQ dataset, Met dataset and Traffic dataset. ML models were trained and tested on these individual combinations of datasets and the performance metrics were evaluated to show that the models trained on the experiment dataset consistently outperformed those trained on the controlled datasets. With a performance boost of at least 20% and an error reduction of at least 18.97% recorded for 98% of the ML algorithms when trained with a dataset containing a form of the traffic dataset, this study reinforces the belief that the traffic dataset has a significant effect on improving the performance of air pollution ML prediction models. Hence, it can be concluded that a performance boost is induced when ML algorithms are trained with a dataset containing a form of the traffic dataset for air quality prediction. Also, this study concludes that there is no single algorithm that has the best performance for all the pollutants and combinations of datasets. ML algorithms perform differently in various situations and with varying combinations of datasets. Future analysis of the result is envisaged with more ML algorithms and research datasets. Open research in this domain is to investigate the effect of the dataset granularity on the performance of the air pollution ML prediction model, and how the traffic dataset can contribute to the granularity of the dataset.

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.

Alpan, K. & Sekeroglu, B. 2020. Prediction of pollutant concentrations by meteorological data using machine learning algorithms. In *International Archives of the Photogrammetry, Remote Sensing* *and Spatial Information Sciences - ISPRS Archives*. International Society for Photogrammetry and Remote Sensing: 21–27.

Anurag, N.V., Burra, Y., Sharanya, S. & Gireeshan, M.G. 2019. Air quality index prediction using meteorological data using featured based weighted xgboost. *International Journal of Innovative Technology and Exploring Engineering*, 8(11 Special Issue): 1026–1029.

Arowoiya, V.A., Oke, A.E., Aigbavboa, C.O. & Aliu, J. 2020. An appraisal of the adoption internet of things (IoT) elements for sustainable construction. *Journal of Engineering, Design and Technology*, 18(5): 1193–1208.

Ashayeri, M., Abbasabadi, N., Heidarinejad, M. & Stephens, B. 2021. Predicting intraurban PM2.5 concentrations using enhanced machine learning approaches and incorporating human activity patterns. *Environmental Research*, 196. 10.1016/j.envres.2020.110423.

Babu, K.M. & Beulah, J.R. 2019. Air quality prediction based on supervised machine learning methods. *International Journal of Innovative Technology and Exploring Engineering*, 8(9 Special Issue 4): 206–212.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G. & Di Carlo, P. 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research*, 8(4): 652–659.

Bowatte, G., Lodge, C., Lowe, A.J., Erbas, B., Perret, J., Abramson, M.J., Matheson, M. & Dharmage, S.C. 2015. The influence of childhood traffic-related air pollution exposure on asthma, allergy and sensitization: a systematic review and a meta-analysis of birth cohort studies. *Allergy*, 70(3): 245–256.

Bozdağ, A., Dokuz, Y. & Gökçek, Ö.B. 2020. Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environmental Pollution*, 263.

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J. & Guo, Y. 2018. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. *Science of the Total Environment*, 636: 52–60.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L.D., Abramson, M.J. & Guo, Y. 2018. Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: A satellite-based estimation using the random forests approach. *Environmental Pollution*, 242: 605–613.

Cihan, P., Ozel, H. & Ozcan, H.K. 2021. Modeling of atmospheric particulate matters via artificial intelligence methods. *Environmental Monitoring and Assessment*, 193(5).

Comert, G., Darko, S., Huynh, N., Elijah, B. & Eloise, Q. 2020. Evaluating the impact of traffic volume on air quality in South Carolina. *International Journal of Transportation Science and Technology*, 9(1): 29–41.

Department for Environment, F. and R.A. (Defra) webmaster@defra gsi gov uk. Pollution forecast provided by the Met Office- Defra, UK. https://uk-air.defra.gov.uk/forecasting/ 1 March 2021.

Dobrea, M., Badicu, A., Barbu, M., Subea, O., Balanescu, M., Suciu, G., Birdici, A., Orza, O. & Dobre, C. 2020. Machine Learning algorithms for air pollutants forecasting. In *2020 IEEE 26th International*

Symposium for Design and Technology in Electronic Packaging, SIITME 2020 - Conference Proceedings. Institute of Electrical and Electronics Engineers Inc.: 109–113.

Du, Z., Heng, J., Niu, M. & Sun, S. 2021. An innovative ensemble learning air pollution early-warning system for China based on incremental extreme learning machine. *Atmospheric Pollution Research*, 12(9): 101153.

Egwim, C.N., Alaka, H., Toriola-Coker, L.O., Balogun, H., Ajayi, S. & Oseghale, R. 2021. Extraction of underlying factors causing construction projects delay in Nigeria. *Journal of Engineering, Design and Technology*, ahead-of-print(ahead-of-print). https://doi.org/10.1108/JEDT-04-2021-0211 22 October 2021.

Gamil, Y., A. Abdullah, M., Abd Rahman, I. & Asad, M.M. 2020. Internet of things in construction industry revolution 4.0: Recent trends and challenges in the Malaysian context. *Journal of Engineering, Design and Technology*, 18(5): 1091–1102.

Gan, K., Sun, S., Wang, S. & Wei, Y. 2018. A secondary-decomposition-ensemble learning paradigm for forecasting PM2.5 concentration. *Atmospheric Pollution Research*, 9(6): 989–999.

Gul, S. & Khan, G.M. 2020. Forecasting Hazard Level of Air Pollutants Using LSTM's. In I. Maglogiannis, L. Iliadis, & E. Pimenidis, eds. *Artificial Intelligence Applications and Innovations*. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing: 143–153.

Hatzopoulou, M., Weichenthal, S., Dugum, H., Pickett, G., Miranda-Moreno, L., Kulka, R., Andersen, R. & Goldberg, M. 2013. The impact of traffic volume, composition, and road geometry on personal air pollution exposures among cyclists in Montreal, Canada. *Journal of Exposure Science & Environmental Epidemiology*, 23(1): 46–51.

Hellas, M.S., Chaib, R. & Verzea, I. 2019. Artificial intelligence treating the problem of uncertainty in quantitative risk analysis (QRA). *Journal of Engineering, Design and Technology*, 18(1): 40–54.

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D. & Liu, Y. 2018. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. *Environmental Pollution*, 242: 675–683.

Hyndman, R.J. & Koehler, A.B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4): 679–688.

Iskandaryan, D., Ramos, F. & Trilles, S. 2020. Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. *Applied Sciences*, 10(7): 2401.

Jiang, Q. 2019. Estimation of construction project building cost by back-propagation neural network. *Journal of Engineering, Design and Technology*, 18(3): 601–609.

Jo, B.W. & Khan, R.M.A. 2018. An internet of things system for underground mine air quality pollutant prediction based on azure machine learning. *Sensors (Switzerland)*, 18(4).

KS3 Geography Revision. 2021. Population distribution - Population and migration - KS3 Geography Revision. *BBC Bitesize*. https://www.bbc.co.uk/bitesize/guides/zkg82hv/revision/1 25 July 2021.

Kumari, N.S.A., Kumar, K.S.A., Raju, S.H.V., Vasuki, H.R. & Nikesh, M.P. 2020. Prediction of Air Quality in Industrial Area. In *Proceedings - 5th IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2020*. Institute of Electrical and Electronics Engineers Inc.: 193–198.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S. & Duan, C. 2021. Statistical Approaches for Forecasting Primary Air Pollutants: A Review. *Atmosphere*, 12(6): 686.

Lin, Y., Zhao, L., Li, H. & Sun, Y. 2018. Air quality forecasting based on cloud model granulation. *Eurasip Journal on Wireless Communications and Networking*, 2018(1).

Liu, H. & Chen, C. 2020. Prediction of outdoor PM2.5 concentrations based on a three-stage hybrid neural network model. *Atmospheric Pollution Research*, 11(3): 469–481.

Madeiros, M., Vasconcelos, G., Veiga, Á. & Zilberman, E. 2019. *Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods*. Central Bank of Chile. https://ideas.repec.org/p/chb/bcchwp/834.html 19 January 2022.

Mane, K.M., Kulkarni, D.K. & Prakash, K.B. 2020. Prediction of shear strength of concrete produced by using pozzolanic materials and partly replacing NFA by MS using ANN. *Journal of Engineering, Design and Technology*, 19(2): 578–587.

Martınez-Espana, R., Bueno-Crespo, A., Timon, I., Soto, J., Munoz, A. & Cecilia, J.M. 2018. Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain. : 16.

Masih, A. 2019. Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management*, 5(4). https://doi.org/10.22034/GJESM.2019.04.10 9 November 2020.

Mo, X., Zhang, L., Li, H. & Qu, Z. 2019. A novel air quality early-warning system based on artificial intelligence. *International Journal of Environmental Research and Public Health*, 16(19).

Ni, X.Y., Huang, H. & Du, W.P. 2017. Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. *Atmospheric Environment*, 150: 146–161.

OpenWeatherMap. 2021. Weather API - OpenWeatherMap. https://openweathermap.org/api 17 July 2021.

Owusu-Manu, D.-G., Debrah, C., Oduro-Ofori, E., Edwards, D.J. & Antwi-Afari, P. 2020. Attributable indicators for measuring the level of greenness of cities in developing countries: lessons from Ghana. *Journal of Engineering, Design and Technology*, 19(3): 625–646.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. & Cournapeau, D. 2011. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*: 6.

Pinto, J.A., Kumar, P., Alonso, M.F., Andreão, W.L., Pedruzzi, R., dos Santos, F.S., Moreira, D.M. & Albuquerque, T.T. de A. 2020. Traffic data in air quality modeling: A review of key variables, improvements in results, open problems and challenges in current research. *Atmospheric Pollution Research*, 11(3): 454–468.

Power Melinda C., Weisskopf Marc G., Alexeeff Stacey E., Coull Brent A., Spiro Avron, & Schwartz Joel. 2011. Traffic-Related Air Pollution and Cognitive Function in a Cohort of Older Men. *Environmental Health Perspectives*, 119(5): 682–687.

Ray, R., Haldar, S., Biswas, S., Mukherjee, R., Banerjee, S. & Chatterjee, S. 2019. Prediction of benzene concentration of air in urban area using deep neural network. In *Advances in Intelligent Systems and Computing*. Springer Verlag: 465–475.

Rossi, C., Farasin, A., Falcone, G. & Castelluccio, C. 2019a. A Machine Learning Approach to Monitor Air Quality from Traffic and Weather data. In *CEUR Workshop Proceedings*. CEUR-WS: 66–74.

Rossi, C., Farasin, A., Falcone, G. & Castelluccio, C. 2019b. UAQE: Urban air quality evaluator. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer: 337–343.

Rossi, R., Ceccato, R. & Gastaldi, M. 2020. Effect of Road Traffic on Air Pollution. Experimental Evidence from COVID-19 Lockdown. *Sustainability*, 12(21): 8984.

van Rossum, G. & Drake, F.L. 1995. Python Reference Manual. : 196.

Rybarczyk, Y. & Zalakeviciute, R. 2018. Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied Sciences*, 8(12): 2570.

Selvi, S. & Chandrasekaran, M. 2020. Framework to forecast environment changes by optimized predictive modelling based on rough set and Elman neural network. *Soft Computing*, 24(14): 10467–10480.

Shah, S.K., Tariq, Z., Lee, J. & Lee, Y. 2020. Real-Time Machine Learning for Air Quality and Environmental Noise Detection. In *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*. Institute of Electrical and Electronics Engineers Inc.: 3506–3515.

Srivastava, C., Singh, S. & Singh, A.P. 2019. Estimation of air pollution in Delhi using machine learning techniques. In *2018 International Conference on Computing, Power and Communication Technologies, GUCON 2018*. Institute of Electrical and Electronics Engineers Inc.: 304–309.

Su, X., An, J., Zhang, Y., Zhu, P. & Zhu, B. 2020. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. *Atmospheric Pollution Research*, 11(6): 51–60.

Sulaimon, I.A. & Alaka, H.A. 2021. Traffic-Related Air Pollutant (TRAP) Prediction using Big Data and Machine Learning. In Environmental Design and Management International Conference. Bristol, United Kingdom.

Sulaimon, I.A., Alaka, H.A., Olu-Ajayi, R., Mubashir, A., Sunmola, F., Ajayi, S. & Hye, A. 2021. Air Pollution Prediction using Machine Learning – A Review. In Environmental Design and Management International Conference. Bristol, United Kingdom.

Tao, Q., Liu, F., Li, Y. & Sidorov, D. 2019. Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU. *IEEE Access*, 7: 76690–76698.

Tom Tom. 2021. TomTom Developer Portal | Maps APIs and SDKs for Location Applications. *TomTom Developer Portal*. https://developer.tomtom.com/ 17 July 2021.

Tu, J., Liu, Y., Zhou, M. & Li, R. 2020. Prediction and analysis of compressive strength of recycled aggregate thermal insulation concrete based on GA-BP optimization network. *Journal of Engineering, Design and Technology*, 19(2): 412–422.

UN DESA. 2018. 68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html 15 March 2021.

WHO. 2014. WHO | 7 million premature deaths annually linked to air pollution. *WHO*. https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/#.WqBfue47NRQ.mendeley 15 March 2021.

Wu, Q. & Lin, H. 2019. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of the Total Environment*, 683: 808–821.

Xu, Y., Ho, H.C., Wong, M.S., Deng, C., Shi, Y., Chan, T.C. & Knudby, A. 2018. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environmental Pollution*, 242: 1417–1426.

Yafouz, A., Ahmed, A.N., Zaini, N., Sherif, M., Sefelnasr, A. & El-Shafie, A. 2021. Hybrid deep learning model for ozone concentration prediction: comprehensive evaluation and comparison with various machine and deep learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, 15(1): 902–933.

Yarragunta, S., Nabi, M.A., P, J. & S, R. 2021. Prediction of Air Pollutants Using Supervised Machine Learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE: 1633–1640. https://ieeexplore.ieee.org/document/9432078/.

Zhang, J. & Ding, W. 2017. Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. *International Journal of Environmental Research and Public Health*, 14(2): 114.

Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W. & Liu, J. 2021. Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmospheric Pollution Research*, 12(1): 328–339.

Zhang, S., Li, X., Li, Y. & Mei, J. 2018. Prediction of Urban PM2.5 Concentration Based on Wavelet Neural Network. In *Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018*. Institute of Electrical and Electronics Engineers Inc.: 5514–5519.

Zhao, C., Heeswijk, M. van & Karhunen, J. 2016. Air quality forecasting using neural networks. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). 1–7.

Appendix

Table 7: Summary of related studies in the literature

SN	AUTHORS	POLLUTANTS	DATASETS	ML ALGORITHM	COMPARED WITH	BEST PERFORMANCE
1	(Zhang et al., 2021)	$PM_{2.5}$	AQ	EMD-BiLSTM	BiLSTM, LSTM	R^2=0.989, RMSE=6.86
2	(Du et al., 2021)	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	AQ	LCIELM	LCIELM, ELM, WNN, Elman, TVFEMD-ELM(T-E), TVFEMD-WNN(T-W), TVFEMD-Elman (T-El), VMD-LCIELM(V-L) and EMD-LCIELM(E-L)	RMSE=0.0482
3	(Sulaimon & Alaka, 2021)	$PM_{2.5}$, NO	AQ, M, T	XTR, HGBR	LGBM, XGB, RF, BR, NuSVR, GBR, KNN, SVR	R^2=0.789, RMSE=0.099
4	(Yarragunta et al., 2021)	PM_{10}, $PM_{2.5}$, SO_2, CO, NO_2, O_3	AQ	DT	LR , SVM, RFT, NBT, KNN	Accuracy= 99.8%
5	(Cihan et al., 2021)	PM_{10}, $PM_{2.5}$	AQ, M	ANFIS	SVR, CART, RF, KNN, ELM	R^2=0.97, RMSE=3.05)
6	(Ashayeri et al., 2021)	$PM_{2.5}$	T, BOP	SVR		R^2=0.842, RMSE=0.074
7	(Yafouz et al., 2021)	O_3	AQ, M	CNN, LSTM		R^2=0.9348, RMSE=0.0041
8	(Shah et al., 2020)	$PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO	AQ	SVM	RF, DT	Accuracy= 95%
9	(Alpan & Sekeroglu, 2020)	$PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO	M	RF	DT, SVR	R^2=0.74 - 0.86
10	(Kumari et al., 2020)	SO_2, O_3, NO_2	AQ	RF		
11	(Dobrea et al., 2020)	PM_{10}, $PM_{2.5}$	AQ, M	ARIMA	SVR, LSTM,	Corr-Coeff=0.935
12	(Bozdağ et al., 2020)	PM_{10}	AQ, GIS	ANN	ANN, LASSO, SVR, RF, KNN, XGB	R^2=0.58, RMSE=20.8
13	(Selvi & Chandrasekaran, 2020)	O_3, PM_{10}, NO_2	AQ, M	EINN		RMSE=0.0878
14	(Su et al., 2020)	O_3	AQ, M	KELM-WT-PLS	SVR, KELM, BPNN and SR	R^2=0.78
15	(Liu & Chen, 2020)	$PM_{2.5}$	AQ	HI-EWT-NNA-WRELM-IEWT	EWT-WRELM, EWT-WRELM-IEWT, HI-EWT-WRELM, HI-EWT-WRELM-IEWT, EWT-NNA-WRELM, HI-EWT-NNA-WRELM, EWT-NNA-WRELM-IEWT. CEEMD-GWO-SVR, WD-ENN, WPD-PSO-BPNN-AdaBoost, WPD-CEEMD-PSR-PSOGSA-LSSVR, FEEMD-CS-ELM-VMD-CS-ELM	RMSE=5
16	(Wu & Lin, 2019)	$PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3	AQ	LSSVM-BA		RMSE=4.4396
17	(Anurag et al., 2019)	CO, C_6H_6, $C_6H_5CH_2CH_3$, NO, $C_6H_4(CH_3)_2$, NO_x, O_3, $PM_{2.5}$, SO_2, C_7H_8	M	XGB	ANN, DT, MLR	RMSE=15.97
18	(Babu & Beulah, 2019)	$PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3, NH_3	AQ, M	DT	LR, RF, KNN, SVM	
19	(Srivastava et al., 2019)	$PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3	AQ, M	SVR	ANN	R^2=0.02534
20	(Rossi et al., 2019a)	NO_2, PM_{10}, O_3	T, M	BRNN	GLM, RF, SVM, ANN	Accuracy=0.8
21	(Rossi et al., 2019b)	NO_2, PM_{10}, O_3	T, M	BRNN	GLM, RF, SVM, ANN	Accuracy=0.8049
22	(Mo et al., 2019)	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	AQ	ICEEMDAN-WOA-ELM	ARMA, GRNN, ELM, GA-ELM, WOA-ELM, EEMD-WOA-ELM.	RMSE=0.0606
23	(Ray et al., 2019)	C_6H_6	AQ	DNN		RMSE=0.405181022925
24	(Gan et al., 2018)	$PM_{2.5}$	AQ	SD-LSSVR-CPSOGSA	CEEMD-PSOGSA, CEEMD-CPSOGSA, WPD-PSOGSA, WPD-CPSOGSA and SD-PSOGSA	RMSE= 3.7760
25	(Lin et al., 2018)	NO_2, PM_{10}, O_3, $PM_{2.5}$	AQ	CMG	SVR, NAR	Accuracy=71.43%

26	(Chen, Wang, et al., 2018)	PM_{10}	M, LU, AOD	RF	GAM, NLELRM	R^2=0.86, RMSE=34.2
27	(Huang et al., 2018)	$PM_{2.5}$	AOD, M, LU	RF		R^2=0.78
28	(Xu et al., 2018)	$PM_{2.5}$	AOD, LST	Cubist	RF, XGB	R^2=0.48, RMSE=2.64
29	(Chen, Li, et al., 2018)	$PM_{2.5}$	M, LU, AOD	RF	GAM, NLELRM	R^2=0.86, RMSE=10.7 and 6.9
30	(Zhang et al., 2018)	$PM_{2.5}$	AQ, M	WNN	ELM, LSSVM, FNN	R^2=0.9975, RMSE=1.5124
31	(Jo & Khan, 2018)	CH_4, CO, SO_2, H_2S	AQ	ANN	MLR, PCA-LR, ANN	R^2=0.6654, RMSE=0.2104
32	(Ni et al., 2017)	$PM_{2.5}$	M	BPNN	ARIMA	RMSE=6.76
33	(Biancofiore et al., 2017)	$PM_{2.5}$, PM_{10}	AQ, M	RNN	ANN, MLR	R^2=0.8855
34	**This study**	**$PM_{2.5}$, PM_{10}, NO_2, O_3**	**AQ, M, T**	**XTR, RF, LGBM, XGB**	**HGBR, BR, NuSVR, GBR, KNN, MLPR, DT, SVR**	**R^2=0.737023, RMSE=4.019938**

Note: **ANFIS**=Adaptive Neuro-fuzzy Inference System; **ANN**=Artificial Neural Network; **AOD**=Satellite-based Aerosol Optical Depth data; **AQ**=Air Quality data; **ARIMA**=Autoregressive Integrated Moving Average; **BA**= Bat Algorithm; **BiLSTM**=Bidirectional Long Short-Term Memory neural networks; **BOP**=Building Occupancy Pattern; **BPNN**=Back Propagation Neural Network; **BR**=Bagging Regressor; **BRNN**=Bayesian Regularization of Neural Networks; **CART**=Classification Regression Trees; **CEEMD**=Complementary Ensemble Empirical Mode Decomposition algorithm; **CH_4**=Methane; **CMG**=Cloud Model Granulation; **CO**=Carbon Monoxide; **CPSOGSA**=Chaotic Particle Swarm Optimization Method combined with the Gravitation Search Algorithm (CPSOGSA); **$C_6H_4(CH_3)_2$**=meta-Xylene; **$C_6H_5CH_2CH_3$**=Ethylbenzene; **C_6H_6**= Benzene; **C_7H_8**=Toluene; **DNN**=Deep Neural Network; **DT**=Decision Tree; **EEMD**=Ensemble Empirical Mode Decomposition; **ELM**=Extreme Learning Machine; **ELM**=Extreme Learning Machine; **ElNN**=Elman Neural Network; **EMD**=Empirical Mode Decomposition; **EWT**=Empirical Wavelet Transform; **FNN**=Fuzzy Neural Network; **GAM**=Generalized additive models; **GBR**=Gradient Boosting Regressor; **GIS**=Geographic Information System data; **GLM**=Generalized Linear Model; **GRNN**=Generalized Regression Neural Network; **HGBR**=Histogram-based Gradient Boosting Regressor; **HI**=Hampel Identifier; **H_2S**=Hydrogen Sulfide Carbonyl Sulfide; **ICEEMDAN**=Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise; **IEWT**=Inverse Empirical Wavelet Transform; **IMF**=Intrinsic Mode Function; **KELM**=Kernel Extreme Learning Machine; **KNN**=k-Nearest Neighbor; **LCIELM**=Length-Changeable Incremental Extreme Learning Machine; **LGBM**=Light Gradient Boosted Machine Regressor; **LR**=Logistic Regression; **LSSVM**=Least Squares Support Vector Machine; **LST**=Land Surface Temperature data; **LU**=Land Use data; **M**=Meteorological data; **MLPR**=Multi-Layer Perceptron Regressor; **MLR**=Multiple Linear Regressor; **NAR**=Non-linear Autoregressive neural network; **NBT**=Nave Bayes Theorem; **NH_3**=Ammonia; **NLELRM**=Non-linear Exposure-lag-response Model; **NNA**=Neural Network Algorithm; **NO_2**=Nitrogen Dioxide; **NuSVR**=Nu Support Vector Regression; **O_3**=Ozone; **PCA**=Principal Component Analysis; **PLS**=Partial Least Squares; **$PM_{2.5}$**=Particulate Matter(diameter=2.5µm); **PM_{10}**=Particulate Matter(diameter=10µm); **PSR**=Phase Space Reconstruction; **RF**=Random Forest; **RFT**=Random Forest Tree; **SD**=Secondary Decomposition; **SO_2**=Sulfur Dioxide; **SVM**=Support Vector Machine; **SVR**=Support Vector Regressor; **T**=Traffic data; **WNN**=Wavelet Neural Network; **WOA**=Whale Optimization Algorithm; **WRELM**=Weighted Regularized Extreme Learning Machine; **WT**=Wavelet Transformation; **XGB**=Feature-Based Weighted Xgboost; **XTR**= Extra Trees Regressor

Table 8: Description of Air Quality Dataset

	DATE_TIME	LATITUDE	LONGITUDE	NO2	PM10	SO2	FINE	O3	PM1	PM25	TSP	CO	SO2	NO2	NO	AQI
COUNT	1730587	1730587	1730587	1485522	992010	183928	48179	632308	20511	690523	18978	32854	52	51	609070	516492
MEAN	1614298670	58.71559	4.915295	23.2387	16.47164	4.088476	9.357272	50.46037	8.97026	9.019211	17.75867	3.872447	6.365385	20.70588	10.83223	1.680098
STD	4437418.76	305.251	306.3504	19.63155	12.51445	7.085552	7.879215	23.41542	9.690547	7.660558	15.14442	40.88974	0.65765	13.91732	24.97005	0.848755
MIN	1606780800	49.76681	-9.90392	-32.3	-14	-12.3	-6.4	-5.4	0.1	-13	0.5	-0.5	5	2	-0.68602	0
25%	1610449200	51.45797	-2.29377	8.415	9	1.6	5	34	3	4.151	9.4	0.11642	6	6	0.62365	1
50%	1614304800	51.5579	-0.78029	17.78625	13.8	3	7.2	53	5.3	7	14.6	0.34926	6	21	2.36987	2
75%	1617879600	53.40495	-0.12019	33	20.9	5	10.9	68	11.5	11.132	21.2	0.69852	7	31	9.35475	2
MAX	1622570400	15000	15000	282	1361.6	1210	141.9	201.7653	121.3	713	403.1	1100	7	54	766.3411	10

Table 9: Description of Weather Dataset

	DT	LATITUDE	LONGITUDE	TEMP	FEELS_LIKE	PRESSURE	HUMIDITY	TEMP_MIN	TEMP_MAX	SPEED	DEG	ALL	ID	1H
COUNT	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	1373712	0
MEAN	1614343976	52.27683	-1.09944	279.7678	275.7696	1011.972	79.94682	278.9453	280.5779	3.732348	184.7709	61.84165	761.3301	
STD	4567323.989	1.544163	1.682915	4.394833	4.904104	13.65112	15.01623	4.412379	4.436787	2.399483	102.2342	35.11993	101.2075	
MIN	1606780800	49.76681	-9.90392	259.15	241.94	952	1	259.15	259.15	0.02	0	0	211	
25%	1610420400	51.45636	-1.8774	276.51	272.31	1003	71	275.93	277.15	2.06	90	30	800	
50%	1614056400	51.53085	-0.2775	279.88	275.74	1013	83	279.15	280.37	3.13	210	75	802	
75%	1618005600	52.95473	-0.05077	282.81	279.1	1023	93	282.04	283.71	5.14	260	90	804	
MAX	1622595600	60.13922	1.463497	298.23	298.81	1060	100	298.15	304.15	50.93	360	100	804	

Table 10: Description of Traffic Dataset

	CURRENT SPEED	FREEFLOW SPEED	CURRENT_FREEFLOW SPEED	CURRENT TRAVELTIME	FREEFLOW TRAVELTIME	FREEFLOW_CURRENT TRAVELTIME	CONFIDENCE	ROADCLOSURE	DT	LONGITUDE	LATITUDE
COUNT	485955	485955	482904	485955	485955	482904	485955	485955	485955	479818	479818
MEAN	24.59236	26.78424	0.902648	73.63438	61.04907	0.902127	0.925651	0.026194	1618276937	-1.12234	52.33627
STD	11.39737	10.23113	0.163828	47.50204	27.75625	0.164403	0.142133	0.319594	1908591.839	1.484598	1.578814
MIN	0	0	0.053587	0	0	0.053699	0.5	0	1614909600	-7.3311	50.37167
25%	16.75	20.5	0.863971	46.75	41	0.863479	0.95	0	1616619600	-1.9808	51.4549
50%	23	25	1	66	60	1	0.9975	0	1618297200	-0.3456	51.53085
75%	30	31	1	89.5	78	1	1	0	1619892000	-0.0967	53.22137
MAX	71	71	1	1734.25	181	1	1	4	1621710000	1.4634	60.13922