Ensemble of Ensembles for Fine Particulate Matter Pollution Prediction using Big Data Analytics and IoT Emission Sensors


## Abstract

**Purpose** – The study seeks to develop a multilayer high effective ensemble of ensembles predictive model (stacking ensemble) using several hyperparameter optimized ensemble Machine Learning (ML) methods (bagging and boosting ensembles) trained with high volume data points retrieved from the Internet of Things (IoT) emission sensors, time-corresponding meteorology, and traffic data.

**Design/methodology/approach** – For a start, the study experimented with big data hypothesis theory by developing sample ensemble predictive models on different data sample sizes and compared their results. Secondly, it developed a standalone model and several bagging and boosting ensemble models and compared their results. Finally, it used the best-performing bagging and boosting predictive models as input estimators to develop a novel multilayer high-effective stacking ensemble predictive model.

**Findings** – Results proved data size to be one of the main determinants of ensemble ML predictive power. Secondly, it proved that, as compared to using a single algorithm, the cumulative result from ensemble ML algorithms is usually always better in terms of predicted accuracy. Finally, it proved the stacking ensemble to be a better model for predicting $PM_{2.5}$ concentration levels than bagging and boosting ensemble models.

**Research limitations/implications** – A limitation of this study is the trade-off between the performance of this novel model and the computational time required to train it. Whether this gap can be closed remains an open research question. As a result, future research should attempt to close this gap. Also, future studies can integrate this novel model into a personal air quality messaging system to inform the public of pollution levels and improve public access to air quality forecasts.

**Practical implications** – The outcome of this study will assist the public in proactively identifying highly polluted areas, thus potentially reducing pollution associated with COVID-19 (and other lung diseases) deaths, complications, and transmission by encouraging avoidance behaviour and supporting informed decisions to lock down by government bodies when integrated into an air pollution monitoring system.

**Originality/value** – This study fills a gap in the literature by justifying the selection of appropriate ensemble ML algorithms for $PM_{2.5}$ concentration level predictive modelling. Secondly, it contributes to the big data hypothesis theory, which suggests that data size is one of the most important factors in ML predictive capability. Thirdly, it supports the premise that when using ensemble ML algorithms, the cumulative output is usually better in terms of predicted accuracy than when using a single algorithm. Finally, it provides a novel multilayer high-performance hyperparameter optimised ensemble of ensembles predictive model that can accurately predict $PM_{2.5}$ concentration levels with improved model interpretability and enhanced generalizability, as well as the provision of a novel databank of historic pollution data from IoT emission sensors that can be purchased for research, consultancy, and policy making.



**Keywords:** Air Pollution Prediction, Big Data Analytics, Ensemble of Ensembles, IoT, Machine Learning

**Article Type**: Research paper

## 1 Introduction

Air pollution, which includes ozone and outdoor and indoor particulate matter, is a major public health risk factor for many of the main causes of death, such as heart disease, stroke, lower respiratory infections, lung cancer, diabetes, and chronic obstructive pulmonary disease, accounting for five million global premature deaths per year (N. Zhou *et al.*, 2022). Air pollution affects people at all stages of their lives; at the pregnancy stage where it causes low birth weight; at the child stage where it causes slower development of the lungs and more wheezing/cough; at the adult stage where it causes chronic obstructive pulmonary disease (as chronic bronchitis); and at the elderly stage where it causes dementia, diabetes, heart attack, heart failure, and strokes (Royal College of Physicians, 2016). Not only does air pollution affect public health, but it also has negative environmental impacts on the world. According to Lan Phuong Nguyen *et al.* (2021), air pollution has negative consequences for both soil and water because it creates acid rain, which harms trees and plantations as well as buildings, outdoor sculptures, structures, and statues. Also, it is the viewpoint of Manisalidis *et al.* (2020) that air pollution through gas emissions from industrial facilities, power plants, vehicles, and trucks, among other sources, causes haze, which lowers transparency in the atmosphere and causes significant obstacles in locomotion, transportation, and supply chain. Additionally, the stratospheric ozone layer, which shields us from the sun's harmful ultraviolet radiation, is gradually destroyed by air pollution caused by aerosols, chemicals, and pesticides. This results in significant impairment of the photosynthetic rhythm and metabolism in plants as well as malignant skin in humans. (Dong *et al.*, 2009).

In the United Kingdom (UK), nine out of ten individuals breathe air that exceeds World Health Organization (WHO) guidelines, resulting in between 28,000 and 36,000 fatalities per year from long-term air pollution exposure (Balogun, Alaka and Egwim, 2021). Ambient air pollution in the UK comprises Particulate Matter(PM), Oxides of Nitrogen($NO_x$), and Ozone($O_3$), among others, as air pollutants, which are mostly undetectable until they cause brown haze, which is becoming more frequent in other parts of the world (Royal College of Physicians, 2016). Among these air pollutants, fine PM particles ($PM_{2.5}$) are the most health-damaging pollutants, as they are significantly linked to excessive early death due to their tiny size (about 1/30th the width of a typical human hair) and ability to penetrate deep into lung passages (Cocârţă *et al.*, 2021). Notwithstanding that they are largely invisible, their impacts could be far-reaching, as the damage could occur over a lifetime. Unfortunately, neither the WHO's concentration limits nor those set by the UK government have been able to establish the level of exposure that could lead to various diseases and health-damaging effects (Public Health England, 2018). Furthermore, the health problems and diseases associated with air pollution exacerbated by the COVID-19 pandemic have a significant cost to the UK economy – representing £5.3 billion in the UK's National Health Service (NHS) per annum (Higham *et al.*, 2020). Therefore, urgent action is needed to tackle air pollution within society. This is especially as the benefits of any investment made in this direction will outweigh the cost as it has the potential to prevent life-long pain, reduce demands on the NHS, and enable people to live an active and productive life. Predicting the concentration level of $PM_{2.5}$ — the most health-damaging pollutant — is one of the important investments that would minimize air pollution in society, leading to informed avoidance of pollution hotspots for its inhabitants and a considerable reduction in air pollution-related fatalities.

Over the last decade, a vast body of literature (Ma *et al.*, 2020; Goyal and Routroy, 2021) through predictive modelling has shown that government policies aimed at reducing air pollution and moving toward net-zero carbon emissions have resulted in a drop in PM levels. There are countless variables, including meteorological conditions, the burning of fossil fuels, agricultural practices (such as the use of insecticides, pesticides, and fertilizers), exhaust from factories and industries, mining activities, and natural occurrences (such as volcanic eruptions, forest fires, and dust storms), among many others, have been identified by researchers (Reid *et al.*, 2021; Osman *et al.*, 2022) to be responsible for high concentration levels of $PM_{2.5}$ particles in society. Hence, these researchers have considered one or more factors as features (independent variables) when experimenting with predictive modelling with Machine Learning (ML) algorithms. This poses great impediments with regards to the generalization of their results, especially across different regions as the level of impact of these factors depends on the type of pollutants and choice of ML algorithm used (Chen *et al.*, 2021). Nonetheless, a thorough investigation of research (Balogun, Alaka and Egwim, 2021) showed that past studies have employed several standalone, hybrid, and ensemble ML algorithms for $PM_{2.5}$ predictive modelling. However, despite abundant evidence from the literature (Egwim and Alaka, 2021; Egwim *et al.*, 2021) that ensemble ML algorithms, which employ several algorithms whose combined results are nearly always more accurate predictors than the usage of a single ML technique, haven't been extensively utilized in $PM_{2.5}$ predictive modelling since they combine judgments from many algorithms to optimize overall performance. Unfortunately, a lot of the current studies (De Mattos Neto *et al.*, 2021; Sulaimon

*et al.*, 2022) have randomly adopted or simply used one or two ensemble methods from earlier research without justification, which has led to subpar performance, poor selection of models that perform well, or unenhanced generalizability of models created using these methods across other geographies of the world. Therefore, choosing an appropriate ensemble ML algorithm is a challenging and important choice for its effective implementation given the conundrum about the performance of ML algorithms. These studies have thereby created a knowledge gap that needs to be filled. Thus, a comparative study that will compile and assess the use of several ensemble ML algorithms for $PM_{2.5}$ concentration level predictive modelling is necessary. Consequently, with Big Data Analytics and the Internet of Things (IoT) emission sensors, this study seeks to develop a multilayer, high-performing ensemble of ensemble predictive models to predict $PM_{2.5}$ concentration levels. To accomplish this aim, the following objectives will be used:

1. To pre-process a relatively large data of $PM_{2.5}$ from IoT emission sensors with time-corresponding weather and traffic data to establish the most applicable factors causing high $PM_{2.5}$ concentration level.
2. To utilize established factors as independent variables for several ensemble ML algorithms (bagging and boosting ensembles) to develop hyperparameter optimized predictive models.
3. To build a multilayer, highly effective ensemble of ensembles (stacking) predictive model by combining the best predictive models.

This study makes significant theoretical, methodological, and practical contributions to the field of air pollution prediction. It addresses a gap in existing literature by justifying the selection of appropriate ensemble ML algorithms for $PM_{2.5}$ concentration level prediction modelling. It contributes to big data hypothesis theory by affirming that data size plays a crucial role in the predictive capability of ML models. The study also validates the premise that ensemble ML algorithms usually yield more accurate results than a single algorithm. Furthermore, the study introduces a novel predictive model, a multilayer, hyperparameter-optimized ensemble of ensembles, to improve the accuracy of $PM_{2.5}$ concentration level predictions. Moreso, it also provides a novel databank of historical pollution data from IoT emission sensors, which can be used in research, consulting, and policymaking to accurately predict PM2.5 levels. Finally, the study is the first to use robust ensemble ML approaches to forecast PM2.5 concentration levels in the UK. The findings can aid in selecting the best ensemble ML algorithms for preliminary predictive analysis and, when integrated into an air monitoring system, can help identify highly polluted areas. This can potentially reduce pollution-related deaths, complications, and transmission associated with COVID-19 and other lung diseases by encouraging avoidance behavior and supporting informed lockdown decisions by government bodies.

## 2 Literature Review

This section presents a comprehensive examination of pertinent scholarly sources related to two main themes: Fine Particulate Matter and Big Data Analytics in conjunction with Internet of Things (IoT) for Emissions. Through this review, we aim to highlight significant studies, findings, and theoretical developments in these areas, laying the groundwork for the subsequent discussion and analysis.

.

### *2.1 Fine Particulate Matter*

Fine particulate matter, classified as $PM_{2.5}$, is a salient constituent of air pollutants, encompassing particles of 2.5 micrometres in diameter or smaller. The diminutive scale of these particulates facilitates their intrusion into the respiratory system, and, in certain circumstances, allows entry into the bloodstream. Numerous pieces of evidence from an extensive wealth of scientific research have validated that this ability to penetrate deep into the human body's systems engenders a host of potential health complications. As an illustration, contemporary evidence provided by Manisalidis *et al.* (2020) suggest that exposure to $PM_{2.5}$ has been incontrovertibly linked to an array of health risks that are cause for significant concern. These include the premature death of individuals afflicted with heart or lung diseases, the occurrence of nonfatal heart attacks, the development of irregular heart rhythms, the exacerbation of asthma, a decrease in lung functionality, and an increase in respiratory symptoms. A broadly similar point has also recently been made by Jbaily *et al.*, (2022) who noted that certain demographics exhibit a greater susceptibility to these health effects. These groups include individuals with pre-existing heart or lung diseases, children, and older adults, placing them in a position of increased vulnerability. In addition to the substantial health risks posed by $PM_{2.5}$, there are also significant environmental repercussions to consider. It is the viewpoint of Hassan, Islam and Bhuiyan,

(2022) that these particulates can be transported over expansive distances by wind, eventually settling on terrestrial surfaces or aquatic bodies. The aftereffects of such deposition are contingent on the specific chemical composition of the particulates and can precipitate a multitude of environmental problems. Examples of these issues are the acidification of lakes and streams, an alteration in the nutrient equilibrium in coastal waters and large river basins, a depletion of nutrients in soil, damage to forests and crops that are sensitive to such pollutants, a disturbance to the diversity of ecosystems, and a contribution to the harmful effects of acid rain (Lan Phuong Nguyen *et al.,* 2021). Moreover, the impact of $PM_{2.5}$ extends to the realm of cultural heritage. According to Spezzano, (2021) these particulates have been observed to stain and damage materials such as stone, which are often used in culturally significant structures, including statues and monuments. This capacity for destruction adds an additional layer of concern to the existing health and environmental issues associated with $PM_{2.5}$ (Chen, Lin and Chiueh, 2023). The comprehensive examination of the numerous impacts of $PM_{2.5}$ underscores the urgency of addressing this pervasive component of air pollution and reaffirms the necessity of ongoing research and intervention strategies in mitigating these wide-ranging effects.

### *2.2 Big Data Analytics and IoT for Emissions*

In recent years, the dawn of the Big Data and Internet of Things (IoT) era has been heralded, ushering in innovative tools of considerable potency in navigating the intricate dilemma of emissions and climate change. The advent of these transformative technologies has revolutionized the systematization, processing, and evaluation of diverse data sources, a feat previously unachievable with conventional disciplinary analysis tools (Balogun, Alaka and Egwim, 2021). This paradigm shift, underpinned by the proliferation of Big Data and IoT, has thrust climate science into an era characterized by enhanced comprehension and the formulation of innovative mitigation strategies. According to the prevailing scholarly discourse (Sarker, 2022; Yang *et al.*, 2022), the value of these tools in the realm of climate science is immeasurable. They have expedited explorations into the nuances of climate change, simultaneously fostering the development and implementation of efficacious mitigation strategies. However, along the same lines, Beckage, Moore and Lacasse, (2022) argued that it is worth noting that the complexity of climate change as a global phenomenon necessitates the integration of socio-environmental factors into predictive models. In this regard, Big Data tools have proven to be an indispensable asset. Expanding upon the contributions from numerous scholars, the concept put forth by Marí-Dell'Olmo *et al.*, (2022) suggests that they facilitate the integration of heterogeneous data and models, enabling a comprehensive examination of the intricate interplay between environmental and social factors. Moreover, the application of these technologies extends beyond the confines of climate science, making noteworthy contributions to other sectors such as sustainability and social sciences (Visvizi, Troisi and Grimaldi, 2023). These areas are integral to the successful development and implementation of mitigation strategies, further emphasizing the universal applicability and importance of Big Data and IoT. Drawing on the work of a wide range of philosophers (Goyal and Routroy, 2021), the potency of Big Data tools and IoT becomes particularly salient in the context of predictive modelling of fine particulate matter ($PM_{2.5}$) concentration levels. This view is well supported by Cocârţă *et al.*, (2021), who argued that the deleterious health and environmental impacts of $PM_{2.5}$ are well-documented, underscoring the imperative to accurately predict $PM_{2.5}$ concentration levels for effective mitigation. On the basis of this findings, Li *et al.*, (2023) proposes that their unique capacity to integrate and analyze diverse data sources ushers in a new era of predictive modelling possibilities, such as predicting $PM_{2.5}$ concentration levels. A broadly similar point has also recently been made by researchers (Wong *et al.*, 2023) emphasizing the complexity of $PM_{2.5}$ concentration predictive modelling, necessitating the use of multiple ensemble ML algorithms. This approach harnesses the strengths of various algorithms to yield a more accurate and robust model. Ultimately, the convergence of Big Data analytics, IoT, and ML algorithms promises a future where prediction, and thus, mitigation of emissions and climate change impacts, is more precise and effective, fostering a healthier and more sustainable planet.

## 3 Research Methodology

In this section, we introduced study design, data collection and cleaning and the fundamental theories used for experimentation in this study, including bootstrap aggregating (bagging), hypothesis boosting (boosting), and stacked generalization (stacking). This methodology is underpinned by the theoretical foundations of ensemble learning methods which are powerful machine learning strategies known for their superior predictive performance and generalization capabilities.

### *3.1 Bagging Ensemble*

Bagging is an ensemble machine learning method that uses many models of the same algorithm with randomly chosen portions of data (Opitz and Maclin, 1999). Forests of randomized trees (random forest, extremely randomized trees, etc.) and bagging techniques are examples of bagging ensembles (bagging classifier or regressor meta-estimator). The underlying theory of bagging lies in generating multiple bootstrap samples from original data and training a separate model on each sample. The ensemble's final prediction, which is typically the average of the predictions from each model, will have lower variance than a single model trained on the original dataset. The following formula represents bagging mathematically:

$$f_{bag} = f_1(x) + f_2(x) + \ldots + f_b(x) \quad (1)$$

Where $\boldsymbol{f_{bag}}$ represents the final prediction of the bagging ensemble, $\boldsymbol{f_1(x) + f_2(x) + \ldots + f_b(x)}$ represent the individual base learners in the ensemble. Each $\boldsymbol{f_1(x)}$ is a function (i.e., a model) that takes an input $\boldsymbol{x}$ and produces a prediction. The subscript ***i*** is an index that ranges from ***1*** to ***b***, indicating each base learner in the ensemble, $\boldsymbol{x}$ represents the input to the model, ***b*** and represents the number of base learners in the ensemble. The intricacy involved in the training process of this model is denoted as a function **f (n, m)**, where n signifies the quantity of samples and m represents the number of features. In terms of computational complexity, it is expressed as **O (B * f (n, m))**. Here, **B** stands for the number of bootstrap samples.

### *3.2 Boosting Ensemble*

Boosting is a repeating strategy that adjusts the observation's weight based on the most recent grade. The weight of an observation would be increased if it had been incorrectly classified, and vice versa (Dietterich, 2000). The theoretical underpinning of boosting is based on the principle of adaptive reweighting of instances. After each iteration, the weights of the instances are adjusted based on the prediction performance of the previous model. In the process of implementing the boosting ensemble approach, three critical stages are undertaken. Initially, a base model, denoted as $\boldsymbol{f_0}$, is employed to generate predictions, with the residuals calculated as the difference between the actual target variable $\boldsymbol{y}$ and the predicted values from $\boldsymbol{f_0}$. Subsequently, a new model $\boldsymbol{h_1}$ is trained on these residuals. The improved version of $\boldsymbol{f_0}$, termed $\boldsymbol{f_1}$, is then constructed by integrating the predictions from $\boldsymbol{f_0}$ and $\boldsymbol{h_1}$, as represented by the equation 2.

$$f_1(x) < -f_0(x) + h_1(x) \quad (2)$$

To enhance the performance of $\boldsymbol{f_1}$, the procedure is reiterated for '*k*' iterations, each time generating a new model $\boldsymbol{f_k}$ based on the residuals of the previous model $\boldsymbol{f_{k-1}}$. This iterative process is encapsulated in the equation 3:

$$f_k(x) < -f_{k-1}(x) + h_k(x) \quad (3)$$

In each iteration, $\boldsymbol{h_k}$ is the model trained on the residuals of $\boldsymbol{f_{k-1}}$, and $\boldsymbol{f_k}$ is the updated model that incorporates the predictions of $\boldsymbol{f_{k-1}}$, and $\boldsymbol{h_k}$. The computational complexity associated with the boosting process is denoted as **O (T * f (n, m))**. Here, **T** corresponds to the total number of models or iterations involved in the process. The parameters **n** and **m** represent the number of samples and features respectively, while **f (n, m)** signifies the complexity inherent in training the base model.

### *3.3 Stacking Ensemble*

In contrast to bagging and boosting, **stacking**, further known as stacked generalization, takes into account heterogeneous week learners by merging the fundamental algorithms utilizing a meta model instead of various averaging procedures (Seni and Elder, 2010). The theoretical underpinning of stacking lies in its unique approach to model combination, which uses a meta-learner (or meta-model) to make the final prediction based on the predictions of individual base learners. Mathematically, stacking is represented as:

$$\min_f \sum_{i=1}^{n} l(f(x_i), y_i) + \lambda r(f) \quad (4)$$

Where the empirical risk, which is the first term in the equation above, is determined by a loss function S that evaluates how well the function $f$ performs. The second item, known as the regularisation term, measures the function's complexity and is often a norm of the function or one of its derivatives. For stacking, the computational complexity is represented as O (k * f (n, m) + g (k, n)). Here, k denotes the

quantity of base models, while n and m correspond to the number of samples and features, respectively. The function f (n, m) encapsulates the complexity associated with training the base model. Additionally, g (k, n) signifies the complexity involved in training the meta-model.

### *3.4 Performance Evaluation Metrics*

Typical performance metrics for evaluating regression-based problems like the one for this study (since the target $PM_{2.5}$ contains continuous data) are, Root Mean Square Error (RMSE), and Coefficient of Determination (R-Squared). A particular technique to predictive modelling called regression-based analysis looks at the relationship between a target and a feature(s) (Kuhn and Johnson, 2013). This is especially helpful since it may convey the level of effect that one or more attributes have during ML predictions on a target.

**RMSE** (see Equation 5) stands for the standard deviation of the deviations between the outcomes as predicted by the model and the actual values (training data). The better the model, the closer the RSME value is to zero.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (5)$$

**R-Squared** (see Equation 6) is expressed as the amount of the target's (dependent variable) variation that can be explained by the model's independent variables. Its values are in the range of 0 to 1, with 1 denoting the best model and 0 the worst.

$$\mathrm{R-Squared} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (6)$$

### *3.5 Study Design*

In this study fourteen low-cost IoT emission sensors were designed and strategically deployed across the urban landscape of Wolverhampton, a city located in the UK. The purpose of this deployment was to facilitate comprehensive monitoring of a diverse range of air quality parameters. The monitoring period extended from December 2019 through to April 2020. The mechanical design of these sensors (see Figure 1) is characterized by compactness and robustness, featuring an extruded aluminum body complemented by end moldings made of acrylonitrile styrene acrylate and polycarbonate, types of durable plastics. They operate within a temperature range of -20°C to +45°C ambient and a relative humidity range of 15 - 85% continuously, ensuring reliable performance under various environmental conditions. The sensor's electrical specifications include a power input range of 12-32V direct current, making it adaptable to different power sources. Also, they feature an internal Li-Ion battery with a capacity of approximately 55 Watt-hour, providing substantial operational time. Furthermore, these IoT sensors employs a high sensitivity global navigation satellite system and global positioning system module for location sensing and a 16-gigabyte security digital card for internal storage, capable of storing up to 32 million measurement sets. This allows for extensive data collection and precise location tracking. One of the key features of the sensors is their cartridge-based system that uses active sampling. These cartridges, available in various configurations, can measure a wide range of parameters, including gases, particulate matter, pressure, temperature, and relative humidity. This versatility allows for comprehensive air quality monitoring. The sensor's data handling capabilities are advanced, with its data infrastructure hosted in the Amazon web service cloud supporting various communication technologies for data transmission and provides data access via restful Application Programming Interface (API). This facilitates easy integration into existing traffic management, and meteorological systems in the UK, demonstrating their adaptability and integration capabilities.

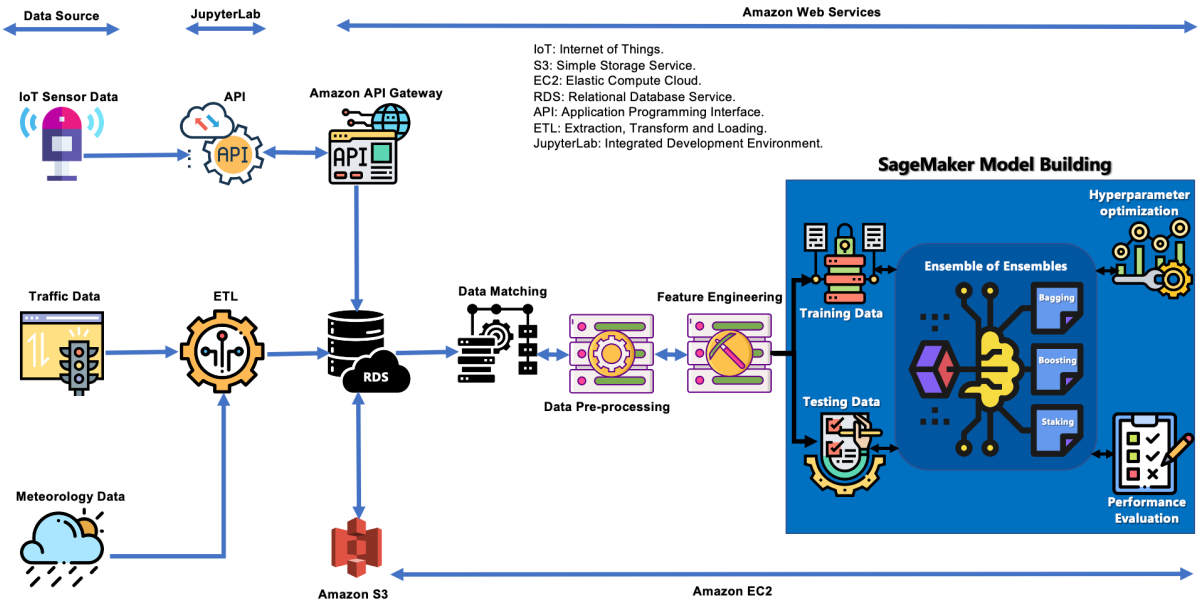


**Figure 1**: Pictorial representation of the research methods for development of high performant ensemble of ensembles predictive model. (Source: Authors own work)

### *3.6 Data Collection and Cleaning*

Most cities today use monitoring devices to measure traffic intensity, meteorological features, and environmental air quality to monitor and limit exposure to air pollution. Depending on the users' preferences, data is gathered at specific intervals (seconds, minutes, hours, days, and so on). To analyse its quantitative data, this study consolidated data from three different data sources: IoT emission sensors, traffic statistics, and meteorological features into a running instance of Amazon Relational Database Service (RDS). Specifically, there are 14 IoT emission sensors in all for $PM_{2.5}$ and other pollution concentrations installed around Wolverhampton City in the UK for this study (part of a deliverable for a funded project by Innovate UK), indicated by red pointers (see Figure 2). For five months, these sensors recorded $PM_{2.5}$ concentrations and other hazardous pollutants every 10 seconds (i.e., December 2019 and April 2020). Consequently, for this period, almost ten billion (i.e., 10 x 6 x 60 x 60 x 24 x 30 x 5 x 14) high volume data points were retrieved via restful JAVA 8 and Spring Boot Application Programming Interface (API) deployed using Amazon Elastic Beanstalk and accessible via Amazon API Gateway hosted on Elastic Compute Cloud (EC2) in the European (London) data centre to perform the big data analyses.

Figure 2: A map of the 14 IoT-installed $PM_{2.5}$ emission sensors in Wolverhampton City, UK.

More specifically, the vehicle counts broken down into several vehicle types made up the majority of the traffic statistic data, which came from the UK's Department for Traffic (DfT) (see Table 1). The sensor data and the traffic data are covered at the same time. Additionally, the UK Met Office provided the weather data for a comparable time. It contained a variety of meteorological factors, such as ambient temperature and pressure, among others (see Table 1). Hourly traffic and meteorological data are supplied, each with about 50,000 high-volume data points. Finally, the hourly average of pollutant concentration from the IoT emission sensors was utilised to match the associated hourly traffic and meteorological data, resulting in (24 hours x 30 days x 5 months x 14 IoT) data points.

**Table 1:** List of Features and Target.

| S/N | Features | Unit | Data source |
|---|---|---|---|
| 1 | Ambient humidity | RH | UK Met Office |
| 2 | Ambient pressure | Pa | |
| 3 | Ambient temperature | $^0$C | |
| 4 | Humidity | RH | |
| 5 | Temperature | $^0$C | |
| 6 | Road type | – | DfT |
| 7 | Link length in Km | – | |
| 8 | Link length in miles | – | |
| 9 | Pedal cycles | – | |
| 10 | Two wheeled motors | – | |
| 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 | Sensor_Id | – | IoT Emission Sensor |
| 23 | Date | – | |
| 24 | Holiday | – | |
| 25 | Day of the week | – | |
| 26 | X (3d coordinates) | – | |
| 27 | Y (3d coordinates) | – | |
| 28 | Z (3d coordinates) | – | |
| 29 | $PM_1$ | $\mu g/m^3$ | |
| 30 | $PM_{10}$ | $\mu g/m^3$ | |
| ***31*** | ***$PM_{2.5}$*** | ***$\mu g/m^3$*** | |

***Key:*** RH= Relative Humidity, Pa=Pascal, $^0$C =Celsius, DfT = Department for Traffic, Lgvs=Large Goods Vehicles, 3d= Three Dimensions, Hgvs=Heavy Goods Vehicles, XYZ= Absolute World Coordinates, $PM_1$= Ultrafine Particles, $PM_{2.5}$= Fine Particles, $PM_{10}$= Coarse Particles, $\mu g/m^3$=Micrograms Per Cubic Meter Air, Km= Kilometers

(Source: Authors own work)

### *3.7 Data Analysis*

The data is a two-dimensional array of 46282 rows and 35 columns, according to a statistical exploratory data analysis of the matching data implemented using Pandas - an open-source data manipulation and analysis library in Python (from sensors, hourly traffic, and weather data), where features/independent variables are the 1st to 30th columns (meteorology, traffic, and other pollutants, e.g., nitrogen dioxide, ozone data), while the target/dependent variable is the 31st column. Pandas was chosen due to its ability to handle diverse data types, missing data and provides robust functionality for data filtering and sub setting, further aiding in the process of data exploration and hypothesis generation (Sulaimon *et al.*, 2022). In the examined data (see Figure 3), some outliers were discovered, notably between December 2019 and January 2020.

One may argue that these outliers towards the end of the year are brought on by holiday parties, shopping, and Christmas. Also, interestingly, lots of missing data found mostly around March 2020 is debatable owing to the influence of the first nationwide lockdown enacted during the COVID19 outbreak in UK cities, thus well justifying the validity of classes of data obtained from the 14 IoT emission sensors installed. Consequently, these outliers and missing data were found and eliminated, leaving a final dataset of 34,370 rows and 31 columns.

Figure 3 shows the 14 IoT emission sensors' hourly mean $PM_{2.5}$ concentration.

## 4 Experimental Results and Analysis

In this section, the feature engineering and big data analytics output will be presented first, followed by a comprehensive experimental comparison to demonstrate the performance of our established high-performing ensemble of ensembles predictive model. All the experimental work on predictive modelling utilized Scikit-learn, a Python programming language package that includes a wide range of cutting-edge methods for supervised and unsupervised medium-scale issues (Pedregosa *et al.*, 2011).

### 4.1 Feature Engineering

To achieve this criterion, this study employed the standardized feature scaling technique, which assumes that the normal distribution of each given dataset has a mean of zero and a variance of one (Yao *et al.*, 2022). A multivariate filter-based feature selection technique called Spearman's rank correlation coefficient was used to evaluate the entire feature space, eliminate redundant, noisy, and out-of-date features, as well as to increase model accuracy, make the model easier to understand, make the computations simpler, and make the model more generalizable. This Spearman's correlation coefficient is a non-parametric test that shows if a relationship between two or more qualities is strengthening or weakening by examining the degree of correlation between them. The estimated strength between the characteristics using Spearman's correlation coefficient fluctuates between +1 and 1 when one feature is a perfect monotone function of the other. As a result, to create the ensemble of the ensemble prediction model, 24 significant characteristics (see figure 4) with coloured bars were employed.

Figure 4: Feature selection ranking according to Spearman. (Source: Authors own work)

### 4.2 Big Data Hypothesis Testing

Given the availability of the relatively large and high volume of data generated, this study briefly conducted hypothesis testing to examine the validity of the theory that "more data means more predictive ability" (Egwim, C.N., Alaka, H., Egunjobi, O. O., Gomes, A., 2022; X. Zhou *et al.*, 2022). With the resulting clean, pre-processed, and feature engineered dataset (34370 data points) split randomly into three in a ratio of 20%, 30%, and 50%, we developed, for each data ratio, 6 sample predictive models using Random Forest (RF), Bagging, Extremely Randomized Trees (Extra-Trees), Adaptive Boosting (AdaBoost), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost) ensemble ML algorithms respectively. To be more specific, each predictive sample model was built using these ensemble ML algorithms for each data ratio (20% of 34370, 30% of 34370, and 50% of 34370), with further random training and testing dataset split at 60:40 respectively. Table 2 details the outcome of this hypothesis based on RMSE, and R-Squared performance evaluation metrics after the predictive modelling experimentation. To determine whether a higher volume of data results in more predictive power, we displayed the area curves for each data ratio's test dataset as shown in figures 5-7. Interestingly, these area curves showed an increase in performance for all sample predictive models developed. Specifically, the sequence of experiments in which increasing quantities of data were sampled (at random) from the original dataset to mimic varying data sizes yielded a 13% increase in R-Squared with a successive 14% decrease in RMSE, thus an overall improvement for each of the sampled predictive models. Hence, the underlying hypothesis is well justified in this study. Therefore, final predictive models for this study will be built using the whole dataset (34370 data points) randomly divided in proportions of 70% to 30% for training and testing, respectively.

**Table 2:** Sample algorithms, models, and their corresponding performance metrics

| | | Performance Evaluation Metrics | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 20% of Total Dataset | | | | 30% of Total Dataset | | | | 50% of Total Dataset | | | |
| | | Training Dataset | | Test Dataset | | Training Dataset | | Test Dataset | | Training Dataset | | Test Dataset | |
| **Algorithm** | **Model** | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** |
| RF | RandomForestRegressor | 0.933 | 4.608 | 0.603 | 29.924 | 0.961 | 2.702 | 0.706 | 25.332 | 0.973 | 2.116 | 0.815 | 11.503 |
| Bagging | BaggingRegressor | 0.803 | 13.635 | 0.556 | 33.437 | 0.885 | 7.931 | 0.676 | 27.919 | 0.902 | 7.534 | 0.797 | 12.603 |
| Extra-Trees | ExtraTreesRegressor | 1.000 | 0.000 | 0.739 | 19.662 | 1.000 | 0.000 | 0.757 | 20.895 | 1.000 | 0.000 | 0.868 | 8.186 |
| AdaBoost | AdaBoostRegressor | 0.975 | 1.758 | 0.694 | 23.054 | 0.986 | 0.936 | 0.739 | 22.499 | 0.990 | 0.751 | 0.820 | 11.160 |
| GBM | GradientBoostingRegressor | 1.000 | 0.000 | 0.728 | 20.535 | 0.984 | 1.130 | 0.753 | 21.230 | 1.000 | 0.044 | 0.787 | 13.208 |
| XGBoost | XGBRegressor | 0.995 | 0.374 | 0.720 | 21.100 | 1.000 | 0.008 | 0.744 | 22.054 | 0.975 | 1.899 | 0.794 | 12.816 |

***Key:*** **RMSE** = Root Mean Square Error, **R-Squared** = Coefficient of Determination

(Source: Authors own work)

***Figure 5:*** *Sample Models' Prediction on 20% of Total Dataset (Source: Authors own work)*

***Figure 6:*** *Sample Models' Prediction on 20% of Total Dataset (Source: Authors own work)*

***Figure 7:*** *Sample Models' Prediction on 50% of Total Dataset (Source: Authors own work)*

### 4.3 Model Development Process and Performance Measures

When experimenting with regression analysis, a range of ensemble ML techniques may be employed to create prediction models. The number of features, the curvature of the regression line, and the type of target all influence which one to choose. Without regard to the above-mentioned requirements, we conducted tests using all the ensemble ML algorithms available in scikit-learn version 0.23.1 at the time of this study.

In concrete, six ensemble ML algorithms: three Bootstrap Aggregating (Bagging) and three Hypothesis Boosting (Boosting) ensemble ML algorithms were used with their default settings to create the individual ensemble models using the training dataset (70 per cent of the total dataset). Also, the Decision Tree (a typical unstable standalone algorithm) was used for benchmark and fair comparisons. This resulted in a total of seven developed models (six ensemble models and one standalone model). After that, the models' performance was evaluated using the unseen test dataset (30% of the entire dataset). To avoid individual model overfitting on the dataset, stratified k-fold, a version of k-fold that yields stratified folds having roughly the same proportion of target class as the initial dataset, was used for cross-validation, where k=10. The general processes involved in developing these ensemble predictive models are shown in figure 1. Finally, as indicated in Table 3, RMSE and R-Squared performance evaluation metrics were used to assess the performance of these prediction models, with figures 8 to 4 displaying their predicted values vs their actual values on the training and test data, respectively.

**Table 3:** Default parameter optimized algorithms, models, and their respective performance evaluation metrics

| Algorithm | Model | Performance Evaluation Metrics | | | |
|---|---|---|---|---|---|
| | | Training Dataset | | Test Dataset | |
| | | R-Squared | RMSE | R-Squared | RMSE |
| Decision Tree (DT) | DecisionTreeRegressor | 0.900 | 0.000 | 0.666 | 26.247 |
| Random Forest (Ensemble of DT) | RandomForestRegressor | 0.977 | 1.577 | 0.845 | 12.171 |
| Bagging Ensemble | BaggingRegressor | 0.9309 | 4.8259 | 0.815 | 14.519 |
| Extremely Randomized Trees (Extra-Trees) | ExtraTreesRegressor | 0.994 | 0.420 | 0.875 | 9.778 |
| Adaptive Boosting (AdaBoost) | AdaBoostRegressor | 1.000 | 1.000 | 0.889 | 8.762 |
| Gradient Boosting / Gradient Boosting Machine (GBM) | GradientBoostingRegressor | 0.998 | 0.165 | 0.866 | 10.527 |
| Extreme Gradient Boosting (XGBoost) | XGBRegressor | 0.962 | 2.628 | 0.858 | 11.135 |

***Key:*** **RMSE** = Root Mean Square Error, **R-Squared** = Coefficient of Determination

***Figure 8:*** *DT Prediction Plot (Source: Authors own work)*

***Figure 9:*** *RF Prediction Plot (Source: Authors own work)*

***Figure 10:*** *Bagging Prediction Plot (Source: Authors own work)*

***Figure 11:*** *Extra Tree Prediction Plot (Source: Authors own work)*

***Figure 12:*** *AdaBoost Prediction Plot (Source: Authors own work)*

***Figure 13:*** *GMB Prediction Plot (Source: Authors own work)*

***Figure 14:*** *XGBoost Prediction Plot (Source: Authors own work)*

To develop the multilayer high performant ensemble of ensembles (stacking) predictive model, we combined the best (based on performance evaluation metrics) predictive models from bagging and boosting ensemble ML algorithms. To begin, we compared the performance of each ensemble learning algorithm using a decision tree (a typically unreliable standalone algorithm) as their base estimator. Next, we carried out an experiment using three bagging ensemble algorithms, optimized RF (a natural ensemble of DT), Bagging, and Extremely Randomized Trees, to tune the hyperparameters and stabilize the base estimator. Alpha and lambda, with values of 100 and 10, are the principal variables utilized in hyperparameter optimization. The chosen parameters serve as regularization terms in the model, effectively managing the bias-variance trade-off to achieve low bias and low variance. The selection of these specific parameter values was not arbitrary but was determined through a systematic grid search. This process involved conducting a series of experiments with varying parameter values and subsequently selecting the values that yielded optimal model performance. This empirical approach ensures that the learning process is controlled and that the model is appropriately regularized. Interestingly, they all outperformed the basic estimator in terms of evaluation metrics (RMSE and R-Squared); the problem, though, was figuring out which bagging ensemble approach was optimal. We used scikit- learn's VotingRegressor to cast a vote with the bagging ensembles using the hard and soft voting criteria to reduce bias and improve generalizability. Amazingly, a successful model called ***Ensemble 1*** evolved that was more accurate and had lower variation (see Figure 15). The experiment was repeated using the same base estimator and hyperparameters this time, but instead of using the bagging ensembles, three boosting ensemble algorithms — Adaptive Boosting, Gradient Boosting Machine, and Extreme Gradient Boosting — were used. Not surprisingly, they all outperformed the base estimator (see Table 4). We again polled them, and the results produced yet another effective model (see Figure 16), which we will refer to as ***Ensemble 2*** in this study.

***Figure 15:*** *Bagging Decision Boundaries (Source: Authors own work)*

***Figure 16:*** *Boosting Decision Boundaries (Source: Authors own work)*

**Table 4:** Hyperparameter optimized algorithms, models, and their corresponding performance metrics

| | **Algorithms** | **Model** | **Performance Evaluation Metrics** | | | |
|---|---|---|---|---|---|---|
| | | | **Training Dataset** | | **Test Dataset** | |
| | | | **R-Squared** | **RMSE** | **R-Squared** | **RMSE** |
| BASE ESTIMATOR | DT | DecisionTreeRegressor | 0.816 | 12.873 | 0.777 | 17.366 |
| BAGGING | RF | RandomForestRegressor | 0.987 | 0.893 | 0.893 | 8.314 |
| | Bagging | BaggingRegressor | 0.957 | 3.046 | 0.872 | 10.012 |
| | Extra-Trees | ExtraTreesRegressor | 0.997 | 0.224 | 0.913 | 6.783 |
| BOOSTING | AdaBoost | AdaBoostRegressor | 1.000 | 0.000 | 0.924 | 5.958 |
| | GBM | GradientBoostingRegressor | 0.932 | 4.793 | 0.818 | 14.182 |
| | XGBoost | XGBRegressor | 1.000 | 0.017 | 0.896 | 8.077 |
| STACKING | **Ensemble of Ensembles** | StackingRegressor | 1.000 | 0.000 | 0.942 | 2.672 |

*Root Mean Square Error (**RMSE**), coefficient of determination (**R-Squared**), decision tree (**DT**), and random forest (**RF**) are the key terms. **AdaBoost**, **GBM**, and **XGBoost** are acronyms for adaptive, gradient, and extreme gradient boosting, respectively.*

(Source: Authors own work)

In conclusion, we used these combined *Ensemble 1 and 2* predictions to train and test a new model, called Ensemble of Ensembles in this work, using the stacked generalization (stacking) approach through Scikit-learn's StackingRegressor. To be more exact, each estimator's predictions from Ensemble 1, and Ensemble 2 are piled and fed into a final estimator, which computes the prediction to reduce their biases. Throughout the Ensemble of Ensembles' training, these estimators were fitted to the entire training dataset. Therefore, to generalize and avoid over-fitting, the final estimator was internally trained on out-samples through cross-validation as shown in figure 17. Ultimately, this yielded a highly performant hyperparameter optimized Ensemble of Ensembles predictive model that was better than *Ensemble 1 and Ensemble 2* based on their respective performance evaluation metrics as shown in Table 4.

Figure 17: A multilayer high-performance ensemble of ensembles predictive model's learning curve for hyperparameter optimization (Source: Authors own work)

## 5 Discussion

The high volume of data points received by the installed IoT emission sensors across the UK cities for this study, matched with the associated hourly traffic and meteorological data, gave the rationale to conduct a hypothesis to examine the validity of the theory that "more data means more predictive ability". We can therefore find that, based on the results of this study, large amounts of data can result in lower estimation variance and, as a result, greater predictive power (see Table 2). This fact is in line with the vast body of knowledge. For instance, it is the viewpoint of Liang and Liu, (2018) and Barba-González *et al.*, (2019) that more data means there's a better chance it'll include relevant information, which is beneficial as there is a natural desire to use these data assets by businesses to improve decision-making since the gathering and processing of "bigdata" have become more common owing to ubiquitous computing and tons of millions of petabytes of data storage in the cloud. One should note, however that although the predictive power of each sample predictive model used to test this hypothesis increasingly improved as the size of training dataset increased, looking through figures 5 – 7, the overall performance of some models (e.g., RF, GMB, etc.) when compared with other models was inconsistent. Also, recall as discussed briefly above that we treated missing values and outliers, and selected some features from the feature space before testing this hypothesis. All these and in line with researches (Zhang, Yang and Zhang, 2021; Zhao *et al.*, 2022) thus suggesting that simply having more data isn't a size-fits-all paradigm. More precisely, it's not only large data that helps us develop high-performing predictive ML models; it's also high-quality data. Therefore, it is a strong recommendation to perform exploratory data analysis to identify missing values and outliers, feature engineering, feature transformation, feature selection, and use multiple algorithms on the big data before it can be valuable in predicting the concentration level of $PM_{2.5}$ — the most health-damaging pollutant.

Furthermore, a close attention to the amount of data streams retrieved from the 14 installed IoT emission sensors every 10 seconds for a period of 5 months (i.e., 10 x 6 x 60 x 60 x 24 x 30 x 5 x 14) with the corresponding hourly traffic counts and metrological variables (24 hours x 30 days x 5months x 14 IoTs) to the actual number of data (34370) finally used for predictive modelling in this study and as argued by Icek (Ajzen, 1991) suggests that there is a high tendency for noise rate to increase as data size increase when data is collected from human actions (traffic congestion, burning of fuel, etc.) because of the restrictions imposed by behavioral inclinations. Hence, this implies that what is really very important is a collection of data points that describe the range of changes for each class that you want to train the ML models with, thus well justifying the need to consider the data velocity and variety when measuring bigdata (volume) for $PM_{2.5}$ concentration level predictive modelling. By comparing the performance of the six ensemble predictive models with the most used tree based standalone model named DT, it can be found from this study and in line with the findings of Egwim *et al.*, (2021) that the performance of ensemble predictive models with or without being hyperparameter optimized is always greater in terms of predictive accuracy relative to the use of a standalone model. Also, this assertion about ensemble predictive models not only holds true for standalone models, but it is also true for hybrid predictive models. This can be justified by comparing the results from this study with the results obtained from the research (Balogun, Alaka and Egwim, 2021) who developed a hybrid model using similar dataset. Consequently, researchers are strongly encouraged to employ ensemble methods when predicting the concentration level of $PM_{2.5}$.

However, considering the existence of several ensemble predictive models vis-a-vis their individual performances, a difficult and important choice is making an ensemble machine learning algorithm that is appropriate for predictive modelling. To mitigate this dilemma, using hard and soft voting rule, we developed *Ensemble 1* (an emergent predictive model from the aggregated list of bagging ensemble models) and *Ensemble 2* (an emergent predictive model from the aggregated list of boosting ensemble models) to serve as input estimators for a novel multilayer high performant hyperparameter optimized stacking ensemble model called Ensemble of Ensembles. Looking through table 4 we can see an outstanding performance made by this novel Ensemble of Ensembles predictive model over all ensemble predictive models that have been proven to be better than hybrid predictive models and standalone predictive model based on performance evaluation metrics. Looking through the learning curve of the multilayer high effective hyperparameter optimized ensemble of ensembles predictive model in figure 17, we can observe a note of caution as regards its evaluation time. More specifically, although this novel Ensemble of Ensembles predictive model have been found to have more predictive power over hybrid and standalone models for $PM_{2.5}$ concentration level predictive

modelling, there is a trade-off for its time complexity. Therefore, as was the case in this study, researchers are strongly encouraged to leverage the power of on-demand cloud computing platforms with a variety of sophisticated clustered computers distributed across several datacenters in the world to computational complexity when implementing this novel model to reduce its time complexity.

## 6 Conclusion and Recommendation

Due to the increased $PM_{2.5}$ concentration level – the most health-damaging pollutant in recent years, more air pollutant predictive models have been developed by past studies. Unfortunately, many past studies have utilized or simply adapted one or two ensemble ML methods from earlier research without rationale, resulting in poor performance, bad model selection, and poor model selection or unenhanced generalizability of models developed using these ensemble ML algorithms across other regions. In this study therefore, a multilayer high performant ensemble of ensembles predictive model developed with several hyperparameter optimized ensemble ML algorithms for $PM_{2.5}$ concentration level predictive modelling with bigdata analytics and IoT emission sensors was proposed. To demonstrate the advantage of this novel model firstly, we tested the bigdata hypothesis by developing sample predictive models on different data sample sizes and compared their results. Secondly, we developed a standalone model, and several bagging and boosting ensemble models and compared their results. Finally, we used the best performing bagging and boosting predictive models as input estimators to develop this specialize type of stacking predictive model. This novel model takes into account the properties of traffic statistics, and meteorological features and pollution concentrations from IoT emission sensors including ambient temperature, absolute world coordinates, ambient humidity, fine particles, traffic counts from cars, taxis and heavy goods vehicles among many others.

The findings of this study will aid in the initial selection of appropriate ensemble ML algorithms for future predictive analysis. Also, this novel model can be used to make decisions on forthcoming events such as pollution exposure evasive conduct, accurate policy making and can be used by air pollution consultants as well as academics thus reducing associated illness and their cost to economy. Furthermore, when this novel model is integrated into an air monitoring system can help the public to proactively identify high polluted areas thus potentially reduce pollution associated/ triggered Covid-19 (and other lung diseases) deaths/ complications/ transmission by encouraging avoidance behavior and support informed decision to lock down by government bodies. This is in accordance with the UK government's clean air plan, which calls for a personal air quality message system to alert the people to levels of pollution. Reduced pollution exposure inculcated by this system can decrease pollution related illness, reducing illness related productivity losses which can cost billions of pounds. This is especially useful as the benefits of any investment made in such direction will outweigh the cost as it has the potential of preventing life-long pain, reducing demands on the NHS, and enabling people to live an active and productive life.

This study suggests that first exploratory data analysis is necessary in order to identify missing values and outliers, feature engineering, feature transformation, feature selection, use multiple algorithms on bigdata and leverage the power of on-demand cloud computing platforms with a variety of sophisticated clustered computers distributed across several datacenters in the world for other pollutant concentration level predictive modelling. A limitation of this study is the tradeoff between performance of this novel model and the computational time required to train it. Whether this gap can be closed remains an open research question. As a result, future research should attempt to close this gap. Additionally, future research might incorporate this innovative approach into a personal message system for air quality to better notify the general population about pollution levels and provide access to air quality forecasts. Additionally, they can investigate using this cutting-edge algorithm to predict additional pollutants.

## References

Ajzen, I. (1991) ‘The theory of planned behavior’, *Organizational Behavior and Human Decision Processes*, 50(2), pp. 179–211.

Balogun, H., Alaka, H. and Egwim, C.N. (2021) ‘Boruta-grid-search least square support vector machine for NO2 pollution prediction using big data analytics and IoT emission sensors’, *Applied Computing and Informatics*,.

Barba-González, C. *et al.* (2019) ‘BIGOWL: Knowledge centered Big Data analytics’, *Expert Systems*

*with Applications*, 115, pp. 543–556.

Beckage, B., Moore, F.C. and Lacasse, K. (2022) 'Incorporating human behaviour into Earth system modelling', *Nature Human Behaviour 2022 6:11*, 6(11), pp. 1493–1502.

Chen, H.S., Lin, Y.C. and Chiueh, P. Te (2023) 'Nexus of ecosystem service-human health-natural resources: The nature-based solutions for urban PM2.5 pollution', *Sustainable Cities and Society*, 91, p. 104441.

Chen, J. *et al.* (2021) 'Artificial intelligence-based human-centric decision support framework: an application to predictive maintenance in asset management under pandemic environments', *Annals of Operations Research*, pp. 1–24.

Cocârţă, D.M. *et al.* (2021) 'Indoor Air Pollution with Fine Particles and Implications for Workers' Health in Dental Offices: A Brief Review', *Sustainability 2021, Vol. 13, Page 599*, 13(2), p. 599.

Dietterich, T.G. (2000) 'Ensemble methods in machine learning', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 1–15.

Dong, M. *et al.* (2009) 'PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining', *Expert Systems with Applications*, 36(5), pp. 9046–9055.

Egwim, C.N., Alaka, H., Egunjobi, O. O., Gomes, A., & M. (2022) 'Comparison of Machine Learning Algorithms for Evaluating Building Energy Efficiency Using Big Data Analytics', *Journal of Engineering, Design and Technology*.

Egwim, C.N. *et al.* (2021) 'Applied artificial intelligence for predicting construction projects delay', *Machine Learning with Applications*,.

Egwim, C.N. and Alaka, H. (2021) 'A Comparative Study on Machine Learning Algorithms for Predicting Construction Projects Delay', in *Environmental Design and Management International Conference, Bristol, United Kingdom.*

Goyal, S. and Routroy, S. (2021) 'Analyzing environmental sustainability enablers for an Indian steel manufacturing supply chain', *Journal of Engineering, Design and Technology*,.

Hassan, S., Islam, T. and Bhuiyan, M.A.H. (2022) 'Effects of Economic and Environmental Factors on Particulate Matter (PM2.5) in the Middle Parts of Bangladesh', *Water, Air, and Soil Pollution*, 233(8), pp. 1–20.

Higham, J.E. *et al.* (2020) 'UK COVID-19 lockdown: 100 days of air pollution reduction?', *Air Quality, Atmosphere & Health 2020 14:3*, 14(3), pp. 325–332.

Jbaily, A. *et al.* (2022) 'Air pollution exposure disparities across US population and income groups', *Nature 2022 601:7892*, 601(7892), pp. 228–233.

Kuhn, M. and Johnson, K. (2013) *Applied predictive modeling*, *Applied Predictive Modeling*.

Lan Phuong Nguyen, K. *et al.* (2021) 'Developing an ANN-based early warning model for airborne particulate matters in river banks areas', *Expert Systems with Applications*.

Li, Z. *et al.* (2023) 'Study on the influencing factors on indoor PM2.5 of office buildings in beijing based on statistical and machine learning methods', *Journal of Building Engineering*.

Liang, T.P. and Liu, Y.H. (2018) 'Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study', *Expert Systems with Applications*,.

Ma, J. *et al.* (2020) 'Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series', *Advanced Engineering Informatics*,.

Manisalidis, I. *et al.* (2020) 'Environmental and Health Impacts of Air Pollution: A Review', *Frontiers in Public Health*,.

Marí-Dell'Olmo, M. *et al.* (2022) 'Climate Change and Health in Urban Areas with a Mediterranean Climate: A Conceptual Framework with a Social and Climate Justice Approach', *International Journal*

*of Environmental Research and Public Health 2022*.

De Mattos Neto, P.S.G. *et al.* (2021) 'Neural-Based Ensembles for Particulate Matter Forecasting', *IEEE Access*,.

Opitz, D. and Maclin, R. (1999) 'Popular Ensemble Methods: An Empirical Study', *Journal of Artificial Intelligence Research*,.

Osman, N. *et al.* (2022) 'Real-Time and Predictive Analytics of Air Quality with IoT System: A Review', pp. 107–116.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12(85), pp. 2825–2830. Available at: http://scikit-learn.sourceforge.net. (Accessed: 7 January 2021).

Public Health England (2018) *Health matters: air pollution - GOV.UK*, *UK Government*. Available at: https://www.gov.uk/government/publications/health-matters-air-pollution/health-matters-air-pollution (Accessed: 22 August 2021).

Reid, C.E. *et al.* (2021) 'Daily PM2.5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018', *Scientific Data 2021 8:1*, 8(1), pp. 1–15.

Royal College of Physicians (2016) *Every breath we take: the lifelong impact of air pollution*, *Report of a working party*. Available at: https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution.

Sarker, I.H. (2022) 'Smart City Data Science: Towards data-driven smart cities with open research issues', *Internet of Things*,.

Seni, G. and Elder, J.F. (2010) 'Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions', *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1),.

Spezzano, P. (2021) 'Mapping the susceptibility of UNESCO World Cultural Heritage sites in Europe to ambient (outdoor) air pollution', *Science of the Total Environment*,.

Sulaimon, I.A. *et al.* (2022) 'Effect of traffic data set on various machine-learning algorithms when forecasting air quality', *Journal of Engineering, Design and Technology*,.

Visvizi, A., Troisi, O. and Grimaldi, M. (2023) 'Big data and Decision-making: How Big Data Is Relevant Across Fields and Domains', *Big Data and Decision-Making: Applications and Uses in the Public and Private Sector*,.

Wong, P.Y. *et al.* (2023) 'An ensemble mixed spatial model in estimating long-term and diurnal variations of PM2.5 in Taiwan', *Science of The Total Environment*,.

Yang, M. *et al.* (2022) 'Circular economy strategies for combating climate change and other environmental issues', *Environmental Chemistry Letters 2022*.

Yao, R. *et al.* (2022) 'A novel mathematical morphology spectrum entropy based on scale-adaptive techniques', *ISA Transactions*,.

Zhang, W., Yang, D. and Zhang, S. (2021) 'A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring', *Expert Systems with Applications*,.

Zhao, H. *et al.* (2022) 'Intelligent Diagnosis Using Continuous Wavelet Transform and Gauss Convolutional Deep Belief Network', *IEEE Transactions on Reliability*.

Zhou, N. *et al.* (2022) 'Prototyping an IoT-based system for monitoring building indoor environment', *Journal of Engineering, Design and Technology*,.

Zhou, X. *et al.* (2022) 'Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism', *Engineering Applications of Artificial Intelligence*,.