IJIRIS: International Journal of Innovative Research in Information Security
Volume 10, Issue 04, May 2024
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives

Prediction of Air Quality Data with Feature Selection and Machine Learning Algorithms

Pandey Nivedita
Assistant Professor, Department of CSE
East Point College of Engineering and Technology,
Bangalore, INDIA
niveditapandey.cse@eastpoint.ac.in
Yashwanth N
Department of CSE
East Point College of Engineering and Technology,
Bangalore, INDIA
Yashwanthn1612@gmail.com
Tejas Kumar K N
Department of CSE
East Point College of Engineering and Technology,
Bangalore, INDIA
tejaskumarcs20@gmail.com
Deepu R M
Department of CSE
East Point College of Engineering and Technology,
Bangalore, INDIA
deepurm189@gmail.com

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.10/Issue04/MYIS10109
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IJIRIS/RS/Vol.10/Issue04/MYIS10109
Received: 07, April 2024 | Revised:16, April 2024 | Accepted: 27, April 2024 Published Online: 07, May 2024 Volume2024
Article ID MYIS10109 http://www.ijiris.com/volumes/Vol10/iss-04/30.MYIS10109.pdf
Article Citation: Pandey,Yashwanth,Tejas,Deepu(2024).Prediction of Air Quality Data with Feature Selection and Machine Learning Algorithms. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 268-274 **doi:>** https://doi.org/10.26562/ijiris.2024.v1004.30
BibTex key: Pandey@2024Prediction

Copyright: ©2024 This is an open access article distributed under the terms of the Creative Commons Attribution License; which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Predicting air quality is essential for managing public health and the environment. This study looks into how well machine learning (ML) algorithms forecast air quality data and incorporates feature selection strategies for better model performance. Air pollution is one of the biggest issues facing humanity today. Every year, air pollution causes millions of deaths due to direct or indirect effects. These days, there is a critical need for effective measures to mitigate the negative impacts of air pollution. Usually, the answers to the issues with air pollution are hasty decisions that don't end up helping. It is vital to concentrate efforts on the pollutants that cause the majority of air pollution in order to create an efficient counter-strategy. In many industrial and urban regions today, monitoring and safeguarding air quality has emerged as one of the government's top priorities. The combustion of fossil fuels, traffic and weather conditions, and industrial parameters

Keywords: Air Quality Prediction, Feature Selection, Machine Learning, Random Forest, Support Vector machine, LSTM.

INTRODUCTION

One of the main issues facing the globe today is air pollution. While there have been notable improvements in air quality over the past year due to the widespread lockdowns and decreased activity levels caused by COVID-19, as the world's population begins to return to normal, air pollution levels have once again begun to rise to dangerous levels. 21 of the world's 30 most polluted cities are in India, and perhaps more concerningly, air pollution is thought to be a contributing factor in the deaths of approximately 2 million Indians each year, according to data released by the World Health Organization and State of Global Air. Due to a variety of industrial, agricultural, socioeconomic, and other variables, the various states of India experience differing amounts of pollution throughout the year. The different elements that affect air pollution in India are depicted, along with the different pollutants that are taken into account while determining the country's Air Quality Index (AQI). The surface-mounted sensors that are positioned at different points in cities around the nation report the concentrations of the pollutants. The AQI, which measures the amount of air pollution or air quality, is then calculated using the data. The Indian Air Quality Index (AQI) is a numerical value ranging from 0 to 500. Based on the concentrations of different pollutants and the impact of air quality on health, it is further classified into six levels: good, satisfactory, moderate, poor, very bad, and severe. Rapid population growth and urban economic expansion in emerging nations like India have resulted in a host of environmental issues, including noise, water, and air pollution.

Human health is directly impacted by air pollution. In our nation, people are becoming more aware of the same things. Acid showers, asthma cases rising, and global warming are a few of the long-term effects of air pollution. Accurate air quality forecasts can lessen the impact of extreme pollution on both individuals and the environment. Therefore, improving air quality forecasting is one of society's top priorities. A gas is sulfur dioxide. one of the main contaminants in the air. It smells awful and harsh, and it's colorless. It mixes readily with other substances to create dangerous compounds such as sulfurous acid and sulphonic acid. Inhaling sulfur dioxide has an impact on one's health. It irritates the throat, nose, and airways, leading to wheezing, coughing, dyspnea, and constriction in the chest area. The amount of sulfur dioxide present in the atmosphere has an impact on the suitability of habitats for both animal and plant groups. Predicting air quality data with feature selection and machine learning techniques is crucial for managing the negative impacts of air pollution[1][2][3][4][5]. Various models like Gaussian Process Regression (GPR), Seasonal Autoregressive Integrated Moving Average (SARIMA), linear regression, lasso regression, random forest regression, and K-nearest neighbor regression have been employed for this purpose.

These models help in capturing complex interactions among input variables and output variables, seasonal and trend patterns in air quality data, and predicting pollutant levels accurately. Feature selection techniques combined with machine learning algorithms enhance the accuracy of air quality predictions, aiding in effective air quality monitoring and management. The use of advanced data mining techniques like neural networks and intelligent optimization algorithms further improves the prediction accuracy, enabling the forecasting of air quality with over 90% accuracy.

LITERATURE SURVEY

[1] Proactive air pollution mitigation requires accurate long-term air quality trend prediction. The complicated non-linearity and numerous frequency components seen in air quality data sometimes pose challenges for traditional approaches. This review of the literature examines new developments in deep learning techniques to tackle these issues, with a particular emphasis on the strategy described in "Deep Hybrid Model Based on EMD with Classification by Frequency Characteristics for Long-Term Air Quality Prediction"Owing to their capacity to recognize intricate patterns and non-linear correlations, deep learning models—in particular, recurrent neural networks (RNNs) and convolutional neural networks (CNNs)—have demonstrated encouraging outcomes in the prediction of air quality. This study proposes a deep hybrid model that leverages Empirical Mode Decomposition (EMD) to address the challenge of multiple frequency components. EMD decomposes the air quality data into Intrinsic Mode Functions (IMFs) representing different frequency bands.

[2] Predicting PM2.5 (fine particulate matter) concentrations with accuracy is essential for controlling air quality and safeguarding human health. Neural networks with Long Short-Term Memory (LSTM) have become an effective method for capturing the intricate spatiotemporal dynamics of PM2.5 data. The use of LSTM-based models for PM2.5 forecasting is investigated in this survey.2.By preprocessing the data with a CNN layer, spatial information can be extracted from the positions of the monitoring stations. GCNs effectively capture spatial dependencies and are well-suited for representing relationships between stations on a geographical network: During predicting, attention mechanisms in the LSTM design can concentrate on pertinent data from nearby stations. A technique that shows promise for PM2.5 spatial temporal forecasting is LSTM neural networks. Through a variety of methods, LSTM-based models can incorporate spatial information and attain excellent prediction accuracy. Improved forecasting and mitigation techniques for air quality will result from ongoing research in this field.

[3] Predicting air quality is essential for protecting public health and the environment. However, due to the complicated nature of air quality data—which is frequently high-dimensional and contains duplicate or unnecessary information—careful feature selection and the use of potent machine learning techniques are required. These two crucial areas are examined in this literature review in order to enhance air quality prediction models. These techniques entail the training and assessment of several machine learning models using various feature subsets. The model's performance metric is used to determine which feature subset is optimal. These techniques incorporate feature selection into the training of the model. A well-known example is LASSO regression, which implicitly does selection by penalizing coefficients and essentially driving unnecessary features to zero ideal for managing nonlinear relationships and high-dimensional data. useful for forecasting particulate air pollutants, such as PM2.5 Reliable ensemble technique that generates several decision trees using arbitrary feature subsets. enhances generalization and manages missing-value data.

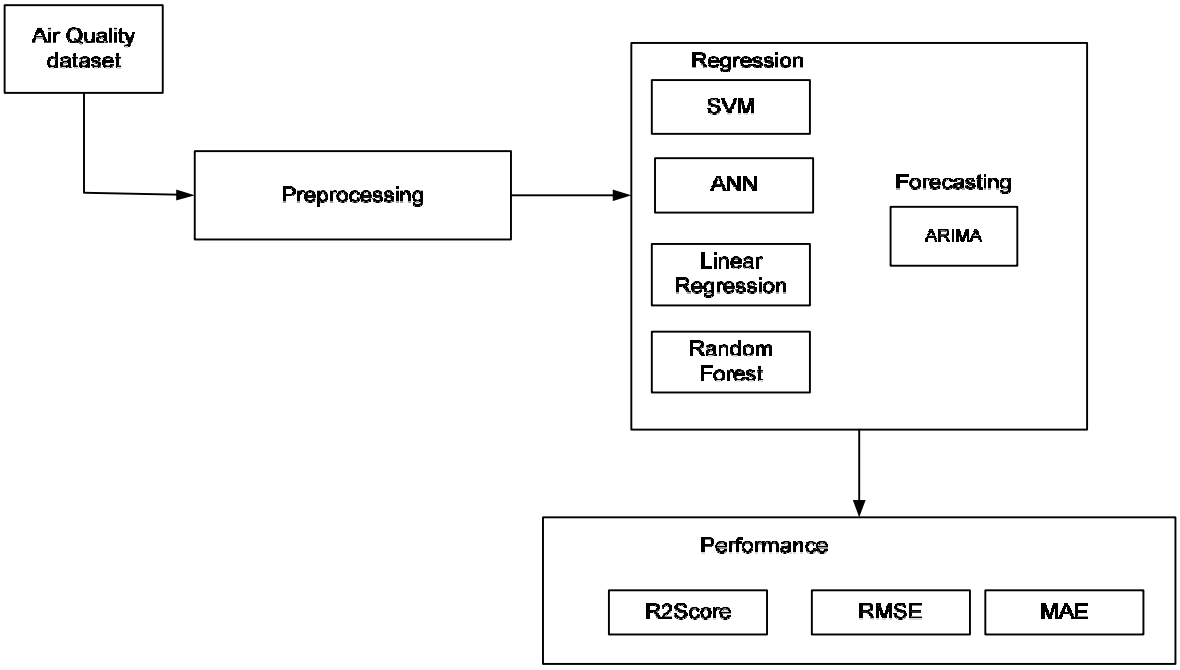
PROPOSED SYSTEM

This project suggests a novel feature selection and machine learning algorithm-based air quality prediction system. Our approach seeks to achieve high prediction accuracy for the Air Quality Index (AQI) and different types of air contaminants. When compared to other methods, the usage of Random Forest and Linear Regression can accurately determine if a data sample is contaminated or not. as well as machine learning methods for forecasting and evaluating air quality. In addition, we compute the predicted values for the next few days using the ARIMA model.Examine how deep learning architectures can be used to handle intricate spatiotemporal patterns in data on air quality. For more thorough forecasting, look at incorporating real-time data sources like satellite imaging or traffic data. Create comprehensible machine learning models to learn more about the variables influencing changes in air quality.

Advantages:

- Improved Prediction Accuracy
- Enhanced Generalizability
- Data-Driven Insights
- User-Friendly Interface

SYSTEM ARCHITECTURE



SYSTEM MODULES

Several essential modules can be separated out of the proposed air quality forecast system for effective development and execution.

1. Module for Data Acquisition:

The collection of air quality data from diverse sources is the main goal of this module.

Sub-modules:

Integration of Data Sources: retrieves information on pollutant concentrations (PM2.5, PM10, O3, etc.) and meteorological conditions (temperature, humidity, wind speed, etc.) by connecting to databases or APIs of air quality monitoring stations.

Data preprocessing: handles missing values, outliers, and ensures consistency (e.g., unit conversions) to clean and prepare the data.

2. Module for Feature Engineering and Selection:

Finding the most pertinent features for air quality prediction is the main goal of this module.

Sub-modules:

Feature engineering (optional): combines preexisting features to generate new ones in order to identify possible non-linear interactions affecting the quality of the air.

Feature selection uses a number of methods, such as correlation analysis to find redundant characteristics.

Acquire information to evaluate the characteristics' role in air quality prediction.

LASSO regression is used to choose features for the model's training.

3. Training and Evaluation Module for Machine Learning Models:

The training and assessment of machine learning models for prediction is the main topic of this session.

Sub-modules:

1. Model Training: This process teaches various machine learning algorithms, such as Support Vector Regression (SVR), to handle non-linear connections and high-dimensional data.
2. To handle missing values and ensure robustness, use Random Forest Regression.
3. Boost to achieve high accuracy and capture intricate interactions.
4. Model evaluation: To choose the best model, compare the performance of the model's using metrics like mean squared error (MSE) and R-squared (R^2).
5. Hyperparameter tuning: Enhances performance by fine-tuning the model's hyperparameters (such as learning rate and number of trees).

4. Module for Prediction and Visualization:

This module makes use of the trained model for forecasting and provides an understandable presentation of the outcomes.

Sub-modules:

Prediction: Forecasts future air quality values for different contaminants and the AQI using the selected machine learning model.

Visualization: Creates an intuitive user interface to show:

- Predicted long-term trends in air quality.
- Comparisons between the actual and anticipated values.
- Maps of air quality for geographical examination (if available data is geographical).

5. Optional User Interface (UI) Module:

An easy-to-use interface for communicating with the system is provided by this module (web application, mobile app).

Sub-modules:

Data Input: Enables users to provide desired contaminants for prediction, time periods, or places.
Interactive Visualization: Based on user input, this visualization dynamically presents forecasted air quality statistics.
Alert System (Optional): Notifies users when expected air quality falls below safe thresholds.
A well-organized foundation for creating the air quality forecast system is offered by these modules. Each module's individual features can be further tailored in accordance with project specifications and data accessibility.

SYSTEM IMPLEMENTATION

1.The SVM Regression Model Implementation Steps

1. Kernel Selection: The model's performance is greatly impacted by the kernel function selected. Try different RBF and linear kernels to see which one best fits your data.
2. Hyperparameter tuning: It's critical to determine the best hyperparameters for SVM regression. For effective hyperparameter space exploration, apply grid search or randomized search.
3. Scalability: When dealing with huge datasets, SVM regression can become computationally costly. If scalability starts to become an issue, take into account data sampling strategies or investigate substitute algorithms such as Boost.
4. You can put your air quality prediction system into practice by following these steps and concentrating on SVM regression within the larger feature selection and prediction/visualization framework. Don't forget to modify the implementation details according to the needs of the project and the tools you've selected.

2. The implementation steps for Multi-Layer Perceptron (MLP) Regression

To determine the most pertinent characteristics for air quality prediction, apply the same feature selection approaches, such as correlation analysis, information gain, and LASSO regression (if applicable with MLP). Try varying the MLP model's hyperparameters, including the batch size, learning rate, number of neurons per layer, and number of hidden layers. This may have a big effect on how well the model works. Employ techniques such as grid search or randomized search in order to identify the ideal hyperparameter setup. As with the earlier solution, optimizing performance requires careful consideration of data quality and hyperparameter adjustment. Known as "black-box" models, MLPs might be more difficult to understand than other algorithms, such as decision trees. To understand which features, contribute, think about methods such as feature importance analysis within the MLP framework.

3. The implementation steps for Linear Regression

Because of the intricate and frequently non-linear interactions between air pollutants and influencing factors, Linear Regression is generally not the most recommended model for air quality prediction, even though it can be used in your project for comparison purposes. Here's how your project pipeline might incorporate linear regression: Divided the preprocessed data into sets for testing and training. Utilizing the training set, train the linear regression model. Measures like MSE, R2 and MAE can be used to assess the model's performance on the testing set. To determine whether linear regression is appropriate for your data, compare its performance with other models you intend to use, such as SVR and Random Forest. Despite being a useful beginning point for comprehending the connections between air quality and influencing factors, linear regression may not be the most accurate model for prediction because of possible non-linearities. To investigate more sophisticated prediction capabilities, think about including other machine learning algorithms into your project, such as SVR, Random Forest, or Boost.

4. The implementation steps for Random Forest Regression

Employ feature selection strategies to find pertinent features for air quality prediction, such as information gain and correlation analysis. These methods can be applied to produce a subset of features that are intended only for the Random Forest Regression model's training. You may wish to compare the Random Forest Regression model's performance with that of other machine learning models, such as Support Vector Regression (SVR) or Boost, as detailed in the system implementation as a whole, after training and assessing it. Based on the assessment metrics, pick the model that performs the best overall on your particular dataset. A key component of Random Forest Regression is featuring selection. A well-chosen subset of features can greatly shorten training times and increase model performance. Try varying the hyperparameter settings until you discover the setup that works best for your data. Depending on the dataset and the particular air quality prediction task, other hyperparameters may work better.

SCOPE OF FUTURE APPLICATION

There are numerous potential future applications for the feature selection and machine learning algorithms-based air quality prediction system that can be advantageous to different stakeholders.

Environmental Organizations:

Improved Air Quality Monitoring: To give real-time forecasts and pinpoint locations that might have air quality problems, the system can be connected with already-existing networks for air quality monitoring.
Targeted Interventions: Agencies can take preemptive steps like issuing air quality advisories or putting in place temporary emission control measures by identifying regions with high expected pollution levels.

Identification of Emission Sources: By highlighting important variables affecting air quality, feature selection approaches help authorities pinpoint the main sources of pollution and create focused legislation.

Public health officials:

Protection of Vulnerable Populations: Timely issuance of air quality alerts, particularly for susceptible groups such as children, the elderly, and those suffering from respiratory ailments.

Public Health Risk Assessment: Forecasting changes in air quality can help public health professionals take preventative action by alerting them to possible health concerns linked to air pollution.

Individuals:

Personal Air Quality Awareness: The system's mobile applications can give users access to real-time air quality forecasts for their area, empowering them to plan outside activities with knowledge.

Personal Health Management: By using the forecasts, people with respiratory disorders can schedule their activities to reduce their exposure to air contaminants.

Extra Uses:

Urban Planning: Including air quality forecasts in the planning process can aid in creating cities that support sustainable growth and healthier air.

Smart Buildings: By utilizing air quality forecasts, buildings may enhance indoor air quality and maximize ventilation systems.

Precision Agriculture: By adjusting irrigation and pesticide applications in accordance with weather patterns and possible pollution hazards, air quality forecasts can be utilized to guide agricultural activities.

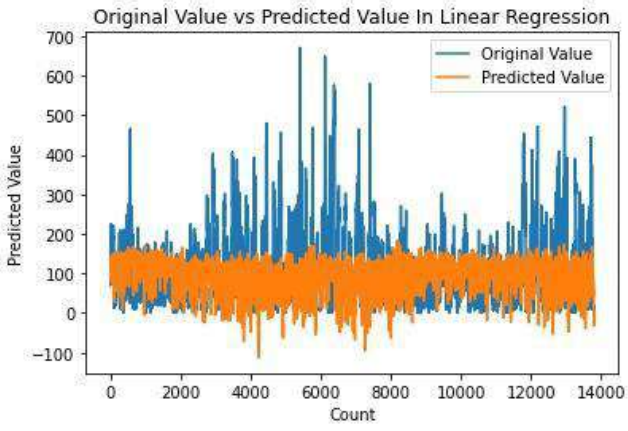
Upcoming Developments:

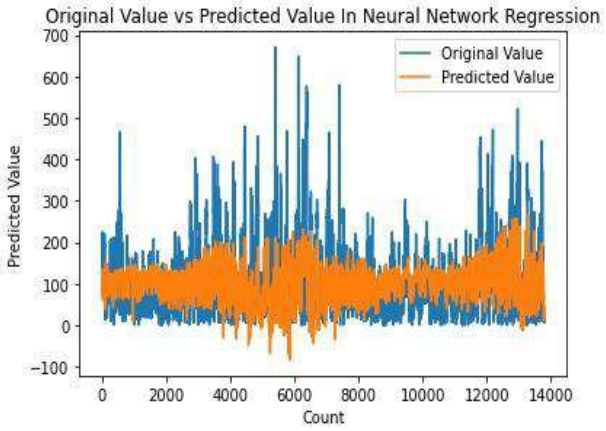
Real-time Integration: Using real-time data sources for more dynamic and thorough air quality predictions, such as satellite images or traffic data.

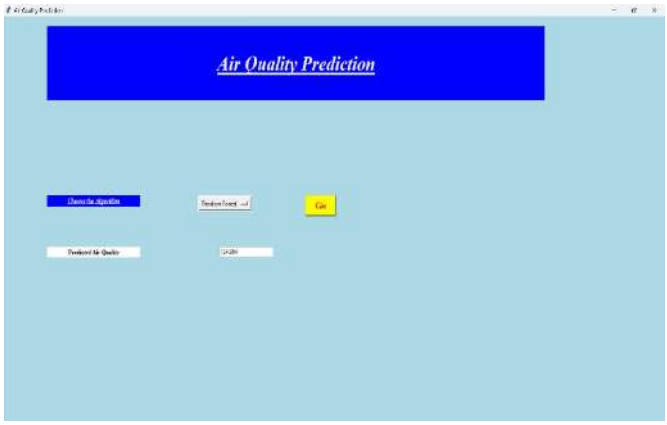
Spatiotemporal Modeling: Convolutional neural networks (CNNs), an advanced deep learning architecture, are capable of handling complicated spatiotemporal patterns to produce predictions that are particular to a certain region.

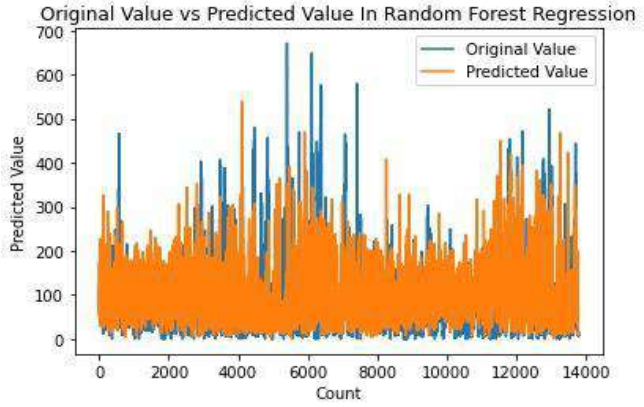
Citizen Science Integration: By including data from citizen science projects (such personal air quality monitors), the data set can grow.

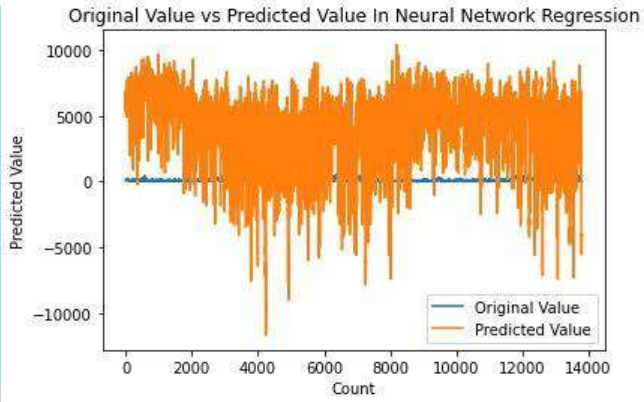
RESULTS











CONCLUSION

This work outlines a comprehensive approach to addressing air pollution concerns, particularly focusing on SO2 concentrations in cities like Pune and Mumbai. The use of various machine learning models, including ANN, Linear Regression, and Random Forest, along with ARIMA time series forecasting, demonstrates a multi-faceted strategy to predict Air Quality levels. Additionally, the exclusion of irrelevant features like locomotion monitoring station or station code highlights the importance of effective feature selection for model accuracy. The mention of future work suggests a forward-thinking perspective, identifying potential challenges and research needs related to other pollutants such as pm2.5 and AQI. This indicates an intention to expand the scope of the project and address broader air quality concerns. Furthermore, the proposal to integrate real-time sensors like MQ series sensor, Dust sensor, and DHT sensor using IoT concepts is a progressive step. This approach would enable the incorporation of up-to-the-minute data for more accurate and timely air quality predictions. The future work suggestions show a commitment to ongoing improvements, addressing additional pollutants, and incorporating emerging technologies like IoT for more dynamic and responsive air quality monitoring.

REFERENCES

[1]. https://www.scmp.com/news/world/europe/article/3149735/air-pollution-kills-7-million-year-says-who-it-tightens
[2]. https://www.dataversity.net/a-brief-history-of-machine-learning/
[3]. https://doi.org/10.1016/j.jclepro.2020.125341
[4]. https://doi.org/10.1016/j.scitotenv.2020.139454
[5]. https://www.iqair.com/
[6]. https://doi.org/10.1145/3369555.3369557
[7]. https://datadrivenlab.org/air-quality-2/chinas-new-air-quality-index-how-does-it-measure-up/
[8]. https://doi.org/10.1155/2019/6780379
[9]. https://web.archive.org/web/20111115230340/http://www.comeap.org.uk/images/stories/Documents/Reports/comeap%20review%20of%20the%20uk%20air%20quality%20index.pdf
[10]. https://doi.org/10.1109/ACCESS.2019.2925082
[11]. https://www.who.int/
[12]. https://www.comeap.org.uk