Technical University of Varna
Annual Journal

https://doi.org/10.29114/ajtuv.vol7.iss1.267

Vol. 7 Issue 1 (2023)

Published: 2023-06-30

ISSN 2603-316X (Online)

System for assessment and forecast of air quality in populated areas

Milena Karova[2], *Tsvetelin* Petrov[1], *Kristian* Ivanov[2], *Nayden* Nikolov[1] and *Tony* Gadjev[1]

1 – Technical University of Varna, Department of Software and Internet Technologies, 9010, 1 Studentska street, Varna, Bulgaria
2 – Technical University of Varna, Department of Computer Science & Engineering, 9010, 1 Studentska street, Varna, Bulgaria

Corresponding author contact: mkarova@tu-varna.bg

Abstract. *The paper provides an account of a system for collecting data, forecasting and assessing the quality of ambient air in a given locality. The developed system allows for extremely sustainable analysis of the results and due consideration of the utilization of artificial intelligence algorithms and methods for the development of accurate forecasts. The obtained results are expected to detect the problems related to the quality of air before their actual occurrence.*

Keywords: measuring station, data base, forecasting, machine learning, neural network, servers

1 Introduction

One of the major problems facing modern society is air pollution. Air degradation in human settlements exerts a devastating effect on people’s health with a wide range of acute and chronic health problems ranging from irritating effects to death (Пулич В., 2014). The overcrowded population of Tehran and the increasing number of vehicles as well as the concentration of industries are the main causes of air pollution in the past two decades (Seyedeh Reyhaneh Shams, 2021). Over the years, there has been a trend for more and more diligent work on the design, construction and development of intelligent systems that monitor the state of the environment and provide accurate data and clear forecasts. Such systems help to anticipate and prevent environmental disasters in advance.

The need to obtain up-to-date information and prepare detailed and accurate forecasts for current and future trends in air pollution is crucial for modern society.

Many of these systems serve several environmental purposes. At a basic level, they inform us how clean or polluted the air is, help us monitor the progress in reducing air pollution, and inform the public about the air quality in their communities (Centers for Disease Control and Prevention, 2020). The problem, however, is that they do not provide air quality forecasts in certain regions.

The team's research suggests an approach that will make accurate predictions based on the collected data from the measurements and result processing through artificial intelligence algorithms specifically designed for the process of neural network training.

The system includes a hardware part consisting of networked measuring stations to measure the indicators for air pollution with each measuring station being capable to comprise:

- Controller
- Dust sensor for PM at 2.5 and 10 µm
- Gas sensor - to detect the presence of carbon monoxide – CO
- Gas sensor, Air Quality - TVOC, eCO2, which detects the presence of multiple VOC gases (organic volatile compounds) and allows to calculate the equivalent CO2 concentration (eCO2 - equivalent carbon dioxide).
- Sensor - Air Quality Measures pressure, temperature, humidity, altitude (Altimeter) and VOCs gases (organic volatile compounds).
- Optical Dust Sensor - an optical sensor for air pollution, designed to detect the amount of dust particles. The device has a diagonal infrared LED and a photo transistor enabling easy detection of light

reflected by dust in the air and proves to be extremely effective especially when finer particles need to be discerned.

The complete hardware system is placed in a special UV-protected mounting box, which protects the components (sensors and controller) from adverse weather factors.

The algorithmic structure and the neural network represents the software part of the system. It analyzes the collected data and generates detailed forecasts for a certain period in the future. This allows for early localization and elimination of problems related to air quality.

2 Creating a neural network training dataset

Conducted, in an effort to select the most appropriate machine learning algorithms, were a series of data set analyses generated by a given measuring station. The measurements were made for the period from October 1, 2021. until November 30, 2021. The analyses were performed on the report for fine dust particles with size ≥ 10 µm (PM10).

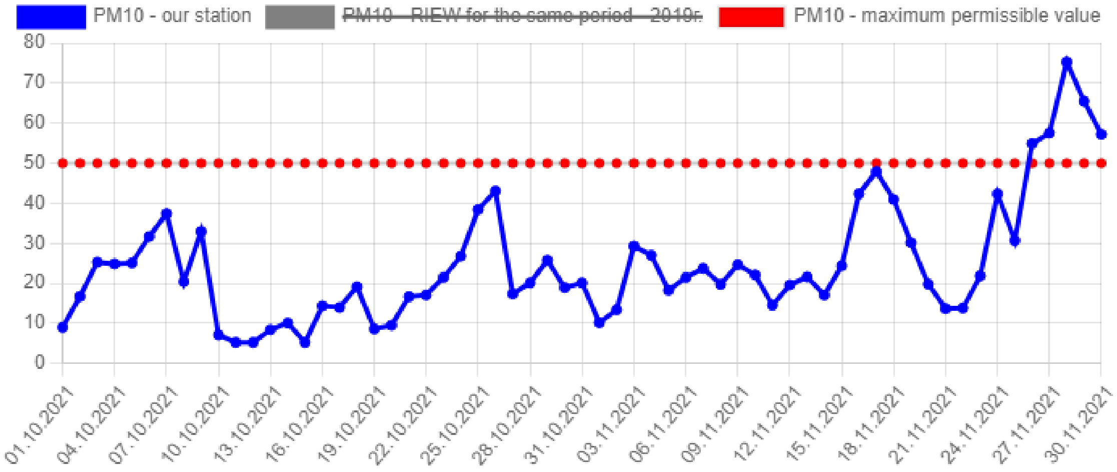

Fig. 1. Concentrated PM10 values.

To take the precise measurements, considerable advantage was taken from the measuring station built by the members of the student club “Creative Code” under the club project "Analysis and implementation of a model for assessment and forecasting of air quality" (Петров Ц., Иванов К., 2020). During the measurements, the station was positioned in the town of Beloslav, Varna, Bulgaria. The measurements were taken once a day at 12:00 o’clock for the period from October,1st ,2021 to November 30th , 2021 (Fig. 1).

The first stage of the analysis is to define the basic requirements that each of the values must meet:

1. Correctness of the data - the value must be zero or a positive number. This value may not correspond to this point. The reason for that may be an error in reading the data from the measuring station. In this case, the value must be created using the MICE method.
2. Data integrity - in case of an omission in the series of values, a new value is set for each missing value, which is calculated by the MICE method. For each variable with missing data, a condition is modeled and data from other variables is used to fill in the missing values. If the algorithm gives a value that does not coincide in any way with the previous ones, i.e. there are dramatic differences, it calculates the average value of the column, which is used as a substitute value for each missing value in that particular column. If the average value is not appropriate, we take the most common value in the column and use it as a substitute value (Динева К., 2020).

After accumulating a certain amount, the data are divided into test and training data sets and a series of machine learning algorithms are applied to train the data set (Fig. 2). The accuracy of the predictions depends on the amount of data recorded by each station.

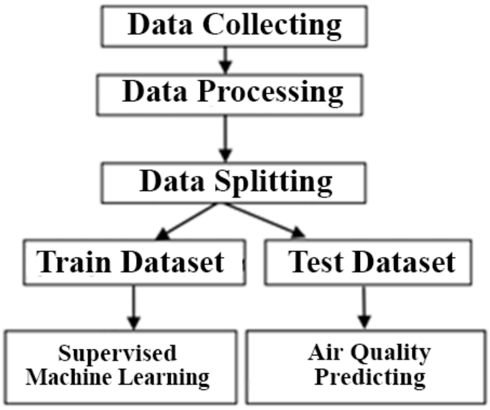

Fig. 2. Data processing

3 Implementation of algorithms for generating forecasts

The level of pollution is subject to various external factors such as car traffic, weather, etc. and therefore is likely to fluctuate wildly. Thus, it is even possible for values measured in the interval of one hour to differ significantly. It is this variability that can sometimes raise serious doubts as to the very accuracy of the predictions, given all the complex microphysical and chemical processes involved. Assessing the impact of chemical compounds on climate and air pollution in particular is quite a challenging task. The impact of climate change on chemical composition and, more specifically, on air quality, is even more difficult to evaluate (National Center for Atmospheric Research, Boulder, 2009).

To accurately predict the PM10 levels, the team reviewed and implemented the SVR model, which implies that the values according to PM10 are stored in variables X and Y. The variable Y contains the average values for the concentration of PM10 and X contains the indices for this value. Through the Numpy library the data is structured appropriately for use in Scikit library methods. The Numpy array is a data structure that efficiently stores and accesses multidimensional arrays (also known as tensors), and enables a wide range of scientific computations. It consists of a memory pointer, along with metadata that interprets the stored data, notably ‘data type’, ‘shape’ and ‘strides’ (Charles R, 2020).

Example code for storing value in variables:

```
x = []
y = []
for i in range(length(PM)):
x.append(i)
y.append(PM[i][3])
y[i] = PM[i][3]
yData = y
y = ['%.2f' yData]
x= numpy.array([X]).T
y = numpy.array(y).ravel()
y = numpy.array(y)
```

A method from the Scikit library is used to divide the data into training and test data sets. This allows selecting the test of train dataset size as a percentage. The result is stored in test and training variables (Marius Dobrea, 2020).

Example for dividing the data into training and test data sets:

```
XTest, XTrain, YTest, YTrain, YTest = train_test_split (x, y, testZize=0.6)
```

The method used to select the most appropriate gamma values is based on the value obtained from the correlation coefficient method (which should be as high as possible) and the method of determining the root mean square error. To obtain these values, it is necessary to conduct multiple tests for each parameter values via the fit() method. The method expects two arguments: XTrain and YTrain, which are real numbers in regression. To predict the data, the predict() method is applied to the object returned from the fit() method.

4 Results from the implementation of algorithms for generating forecasts

Once the prediction method is successfully implemented, the obtained data set (Fig. 3) is compared with the test data set using the correlation coefficient and calculating the root mean square error.

Result of the execution of the algorithm:

```
predicted: 21.7615478, expected: 23.2487545
predicted: 11.4568279, expected: 10.6541080
predicted: 9.8745135, expected: 9.13374424
predicted: 18.5438951, expected: 20.5214538
predicted: 13.4854621, expected: 13.11661625
predicted: 7.5468756, expected: 8.191203117
predicted: 12.2354879, expected: 11.90586853
predicted: 11.1025484, expected: 11.6021347
predicted: 13.4587954, expected: 14.17897511
predicted: 8.1235481, expected: 8.9668293
```

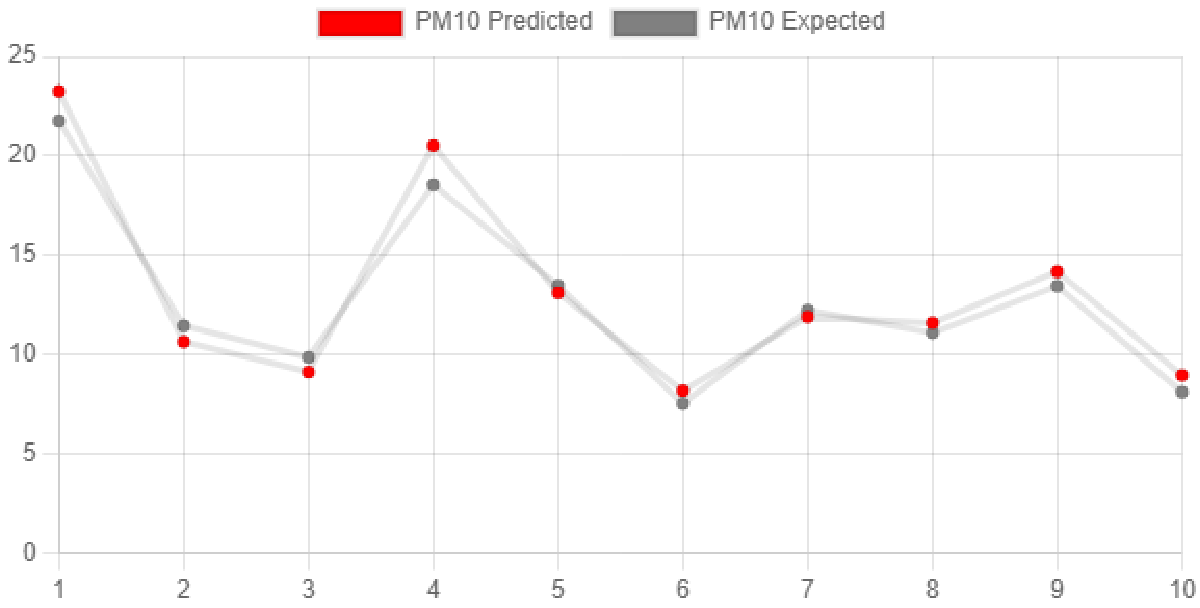

Fig. 3. Result of the implementation of the algorithm for PM10

The algorithms can be applied to the data set for each indicator (Fig. 4), measured and sent by the measuring stations.

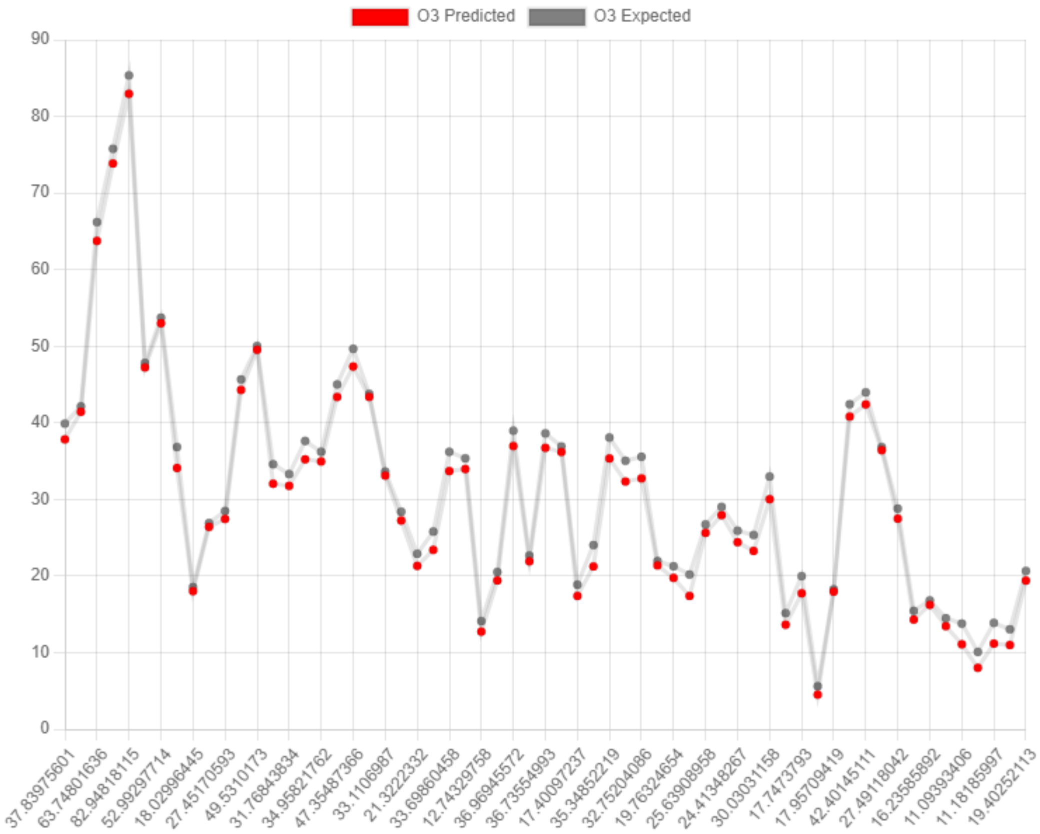

Fig. 4. Results of the implementation of the algorithm for Ozone (O3)

5 Algorithm for determining the total air quality factor

The indicators collected are grouped under a common index, which corresponds to the air quality for the respective area. To that effect, they are normalized to the eligibility threshold for each of them and the value used is averaged.

Ordinance №12 from 15 July 2010. (prom. SG, no. 58 of 30 July 2010) establishes the standards for the maximum permissible concentrations (MPC) for fine particulate matter (PM). The introduced MRLs aim to protect against their detrimental impact on human health and the environment (Министерството на околната среда и водите, Министерството на здравеопазването, 2010).

DA – Daily average
HA – Hourly Average
A8H – An average of 8 hours

Average annual, average daily and average hourly norms:
For PM10:
- (DA) Daily Average corresponds to 50 µg/m3
- (AAR) Average annual rate corresponds to 40 µg/m3

For ФПЧ2.5:
- (AAR) - 25 µg/m3

For SO2:
- (DA) - 350 µg/m3

For CO:
- An average of 8 hours - 10 µg/m3

For NO:
- The regulation document does not specify unambiguously clear norms, which is why the team considers this to be an irregularity!

https://doi.org/10.29114/ajtuv.vol7.iss1.267
Vol. 7 Issue 1 (2023)
Published: 2023-06-30
ISSN 2603-316X (Online)

For NO2:
-(HA) - 200 µg/m3
-(AAR) - 40 µg/m3
For (NO + NO2):
-(AAR) - 30 µg/m3
For O3:
- Population information threshold (PIN)
  -(HA) - 180 µg/m3
- Population information threshold (PIN)
  -(HA) - 240 µg/m3

The information on the impact of atmospheric pollutants on human health has been coordinated with the Ministry of Health (MH) and the National Center for Public Health and Analysis according to Art. 44, para 2 to Ordinance № 12 of 15 July 2010 and Order № RD-09-159 / 14.04.2003 of the Ministry of Health (Министерството на околната среда и водите, Министерството на здравеопазването , 2010г).

The form of the results of the measurements and the forecasts for the given indicators differs in large variations (for example: a normal value of PM10 can be 26,000 µg/m3, compared to a value of carbon monoxide CO, which can be 0.170 µg/m3). In view of this, they need to be normalized in order to achieve greater clarity. Hence the need to compile a formula that normalizes the data in the range from 0 to 2. The indicator zero is the lowest value (absence), and two is the highest value of pollution, designating that the norm has been exceeded twice. It follows that the object of scientific research is no longer the subject of research and is defined as malicious.

The normalization of the data for the respective indicators and the preparation of a general assessment of the air quality is defined as: The sum of the division of the value of an indicator and the norm (for each indicator), with the sum being divided by the total number of indicators (Fig. 5)

$$AQI = \frac{\sum_{k=1}^{N=5}(NNV_k/NCI_k)}{N} \tag{1}$$

Fig. 5. Formula for data normalization

AQI - Air Quality Index
NNV - Non-normalized value
NCI - Norm for a Current indicator
N – Number of indicators

Table. 1 shows the division, which is obtained by dividing the current NNV by NCI, as the answer is recorded in the column allocated for NV and the sum of the last column is divided by the number of indicators (N), the number to be presented AQI.

CN	CI	NI	Div	NNV	NV
DA	PM10	50	/	24	0.48
DA	SO2	350	/	13	0.037
A8H	CO	10	/	0.6	0.06
NO(Nitrogen Oxide) should appear here, again not clarified in the regulation					
DA	NO2	200	/	5	0.025
HA	O3	180	/	110	0.61
					0.2424
					AQI

Table 1. Table with sample non-normalized and normalized data

Where:
CN – Current norm
CI – Current indicator
NI – Norm of current indicator
Div – Division
NNV – Non - normalized value
NV – Normal value

Example code for algorithm implementation

```
var totalSum = 0;
var n_sum_pm10 = 0;
var n_sum_co = 0;
var n_sum_so2 = 0;
var n_sum_o3 = 0;
for(var i = 0; i < days_num; i++) {
n_sum_ pm10 += (pm10[i] / 50);
n_sum_co += (co[i] / 10);
n_sum_so2 += (so2[i] / 350);
n_sum_o3 += (o3[i] / 180);
}
n_sum_ pm10 = n_sum_ pm10 / pm10.length;
n_sum_co = n_sum_co / co.length;
n_sum_so2 = n_sum_so2 / so2.length;
n_sum_o3 = n_sum_o3 / o3.length;
var n_arr_indicator = [n_sum_ pm10, n_sum_co, n_sum_so2, n_sum_o3];
totalSum = (n_arr_indicator.reduce((a, b) => a + b, 0)) / n_arr_indicator.length;
```

The algorithm is applied on data sets measured in the period October 1st ,2021 to November 30th , 2021, corresponding to measurements of the following indicators: PM10, CO, SO2 and O3 (Петров Ц., Иванов К., 2020).

6 Experimental results

After applying the algorithms to conduct a general assessment of the air quality, as to the data obtained, the result produced is: 0.1810942116944959

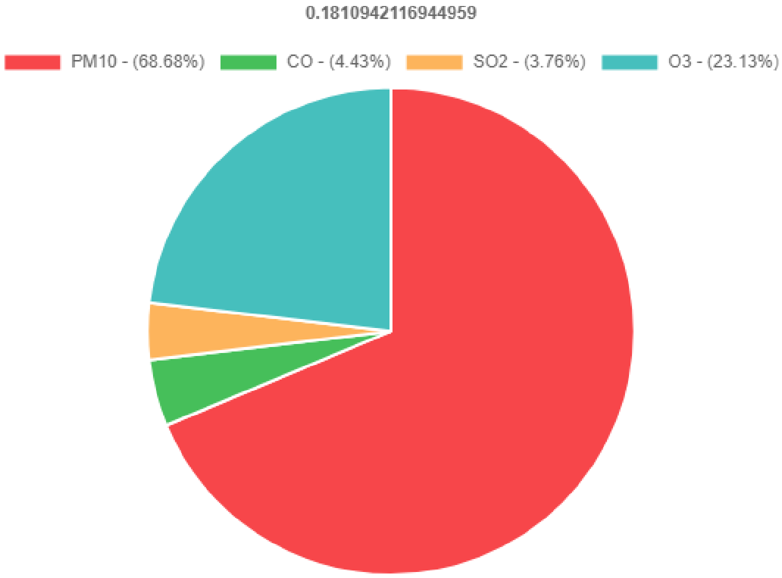

Fig. 6. Diagram of the obtained result

The diagram (Fig. 6) shows the overall assessment of the air quality and each section shows the relationship of the respective indicator to this assessment.

7 Conclusion

The air pollution problems of the future are expected to worsen with the increased use of fossil and nuclear fuels and the rapid growth of the population across the world (Daniel Vallero, 2008).

After conducting a detailed study and comparison of various algorithms in data analysis and machine learning, it was concluded that the use of SVR algorithms is one of the most effective methods for predicting air pollution. With its high accuracy and ability to work easily with a wide range of indicators, this is the algorithm that can predict data with a correlation coefficient of up to 88%, depending on the data set.

The team, however, recognizes that more in-depth work needs to be done in order to improve the algorithm’s ability to determine the overall air quality factor.

In the future, the team intends to expand the system by installing more measuring stations and developing a web system allowing for all the measured data and forecasts to be visualized for public use.

Acknowledgments

We would like to extend our sincere gratitude to Assoc. Prof. PhD Milena Karova, Department of CSE, Technical University of Varna for her generous grant. Without her support, the project at the “Creative Code” student club would not have been possible.

The scientific research, the results of which are hereto presented, was conducted as part of a larger project under the patronage of the Technical university of Varna to encourage sustained research activities, funded by the state budget.

https://doi.org/10.29114/ajtuv.vol7.iss1.267
Vol. 7 Issue 1 (2023)
Published: 2023-06-30
ISSN 2603-316X (Online)

References

Centers for Disease Control and Prevention (2020). Air Quality https://www.cdc.gov/nceh/tracking/topics/AirQuality.htm

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathanie J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Hal-dane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke & Travis E. Oliphant (2020). Array programming with NumPy. https://doi.org/10.1038/s41586-020-2649-2

Daniel Vallero (2008). Fundamentals of Air Pollution, Fourth Edition, 2 – IV

Dobrea, M., Bădicu, A., Barbu, M., Subea, O., Bălănescu, M., Suciu, G., ... & Dobre, C. (2020, October). Machine Learning algorithms for air pollutants forecasting. In *2020 IEEE 26th International Symposium for Design and Technology in Electronic Packaging (SIITME)* (pp. 109-113). IEEE. https://doi.org/10.1109/SIITME50350.2020.9292238

National Center for Atmospheric Research, Boulder (2009). Implications of Climate Change for Air Quality https://public.wmo.int/en/bulletin/implications-climate-change-air-quality

Seyedeh Reyhaneh Shams, Ali Jahani, Saba Kalantary, Mazaher Moeinaddini & Nema-tollah Khorasani (2021). Artifcial intelligence accuracy assessment in -NO2 concentration forecasting of metropolises air. https://doi.org/10.1038/s41598-021-81455-6

Динева К. (2020). ИНТЕГРИРАНЕ НА ХЕТЕРОГЕННИ ДАННИ ОТ РАЗПРЕДЕЛЕНИ IoT УСТРОЙСТВА https://www.iict.bas.bg/konkursi/2020/KrDineva/dissertation.pdf

Министерството на околната среда и водите, Министерството на здравеопазването (2010г). НАРЕДБА № 12 ОТ 15 ЮЛИ 2010 Г. ЗА НОРМИ ЗА СЕРЕН ДИОКСИД, АЗОТЕН ДИОКСИД, ФИНИ ПРАХОВИ ЧАСТИЦИ, ОЛОВО, БЕНЗЕН, ВЪГЛЕРОДЕН ОКСИД И ОЗОН В АТМОСФЕРНИЯ ВЪЗДУХ.

Петров Ц., Иванов К. (2020). Предложение за финансиране на клубен проект от СК “Cre-ative Code;” - ТУ Варна, ‘Анализ и реализация на модел за оценка и прогнозиране на качеството на въздуха.’

Пулич В. (2014). Замърсяването на въздуха и здравето в България Факти, данни и препоръки, Декември. Retrieved from https://env-health.org/IMG/pdf/heal_briefing_air_bulgaria_bgversion.pdf