ANNALS OF “DUNAREA DE JOS” UNIVERSITY OF GALATI
MATHEMATICS, PHYSICS, THEORETICAL MECHANICS
FASCICLE II, YEAR XVI (XLVII) 2024, No. 1
DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2024.1.05

Algorithms for pollutants forecasting

Corina Dima [a,*]

[a] *”Dunarea de Jos” University of Galati, Faculty of Sciences and Environment, Department of Mathematics and Computer Science, 47 Domneasca St., RO-800008, Galati, Romania*

**Corresponding author: cbocaneala@ugal.ro*

Abstract
The impact of pollution on human health is very well known and affects more and more economic and social activities. Since the authorities do not always have the necessary equipment and trained personnel for pollution monitoring, we observed the need to develop algorithms for approximating different polluting values. This paper is a review of the main algorithmic approaches to estimate pollutant values. Most of the proposed approaches are based on machine learning, neural networks and decision trees.

Keywords: Algorithms, pollution, estimation models, forecast

1. INTRODUCTION

The analysis of air quality and water properties are very important in the current period, which is characterized by the accentuated industrial development and the exponential increase of the number of means of transport. The forecast of pollution is necessary so that the authorities are not taken by surprise and can take the necessary measures to avoid environmental hazards, or at least to reduce the impact of pollution. The level of air pollution is quantified by the air quality index which is a linear function determined by the concentrations of some pollutants such as: carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), particles ($PM_{2.5}$, PM_{10}). The air quality index differs according to the standards imposed by different countries and regions. Preventing water pollution is essential for preserving people's health by protecting drinking water resources, but also for preserving the balance of various aquatic ecosystems. Water pollution consists in its contamination with various chemical substances, for example with hydrocarbons, but also with various animal droppings, even with drug residues.

Machine learning algorithms are often used in order to estimate the level of pollution as accurately as possible. In the specialized literature, there are numerous papers that propose various methods of forecasting the air quality index or the water quality index. There are also studies that synthesize some algorithms for predicting the air quality index [1-6, 9, 11-14, 16] or water quality index [7, 15].

Machine Learning is a branch of artificial intelligence that proposes algorithms that, based on existing models and a large volume of data, are able to learn and make decisions and predictions when new information is available. Deep learning can be interpreted as a machine learning technique that uses a structure called an artificial neural network. Deep learning algorithms are a step forward in decision automation, as desired properties are automatically extracted. To work properly, deep learning algorithms need very large amounts of data.

Practical applications of machine learning and deep learning algorithms have appeared in recent years due to the increase of computing power and the possibility of storing a large volume of data. However, the first learning program was written in 1952 by Arthur Samuel and the first neural network was made in 1957 by Frank Rossenblatt [9].

This paper proposes a review of the main types of algorithms for predicting the level of air or water pollution.

2. ALGORITHMS FOR POLLUANTS FORECAST

In this section we present the most popular algorithms for air or water pollution forecast [9, 15].

Multiple linear regression Consider y a dependent variable, and x_1,x_2, …, x_n some independent variables. The linear regression method defines a linear function $f(x_1, x_2, \ldots, x_n)$ such that we determine $min[(y - f\,(x_1, x_2, \ldots, x_n))^2]$ (the square mean error). Simple linear regression supposes that $n = 1$.

Auto-Regressive Integrated Moving Average ARIMA(n, d, m) model forecast future data points using past data organized as time series:

$$y'_t=I+\alpha_1 y'_{t-1}+\alpha_2 y'_{t-2}+\cdots+\alpha_\mathrm{n} \mathrm{y}'_{\mathrm{t-n}}+e_t+\theta_1 e_{t-1}+\theta_2 e_{t-2}+\cdots+\theta_m e_{t-m}$$

The previous relation shows that the predictors are n data values for the autoregressive part, and m errors for the moving average. I means that the data points have been replaced with differenced values of d. [8]. When the time series exhibits seasonality, the model is called SARIMA.

Support vector regression [9] Consider y a dependent variable and x_1,x_2, …, x_n some independent variables. We define a linear regression function, $f(x) = w_1x_1 + w_2x_2 +\ldots+w_nx_n + b$ and choose the tolerance ε, assuming that all points will be at a distance less than ε from f. Our objective is to minimize $\frac{1}{2}||w||^2$ with the restrictions $y_i - w \cdot x_i - b \leq \varepsilon$ and $w \cdot x_i + b - y_i \leq \varepsilon$.

Decision trees This type of algorithm aims to predict a quantitative variable using a set of independent variables. We consider a structure of trees containing decision nodes and leaf nodes. In regression trees, the algorithmreduces standard deviation in order to determine node splits, choosing the variable with the smaller sum of squared errors. The root is a decision node, which branches based on the most relevant independent variable. The process continues until a termination criterion is met. The final nodes represent the prediction of the dependent variable and it is the mean of the values associated with the leaves.

Random Forest This method consists in the generation of some decision trees. The output of the algorithm is the average of the predictions obtained from the generated trees. The data needed to build the trees comes from the training data sets. Independent variables used for the split nodes are randomly selected. Additional data are needed to determine the error of the decision tree.

K-nearest neighbours regression This algorithm is specific to classification problems, but it can also be used in the case of regression problems. For a value k, the algorithm calculates the distance (Euclidean distance, for example) between a particular point and the points of the training data set. The goal is to select the first k points that are the closest, and the prediction will be their average. A variant of this algorithm is the **k-nearest weighted neighbours algorithm** when the prediction is a weighted arithmetic mean.

Deep learning algorithms Deep learning algorithms use neural networks that are structures inspired by the neural connections of the human brain. A neural network is composed of three layers. The first layer consists of neurons (nodes) that receive the input data set in the form of input information. The last layer contains nodes that will provide the output data. Between these two layers there are other layers that contain computational nodes. Each node in one layer is connected to the nodes from the next layer. A connection can have a weight that is used in calculations. We want to minimize an error function (a quadratic function). The neural network is based on the back-propagation algorithm that uses the gradient descent method to find the best weight of a node.

Multi-layer perceptron neural networks These are classical neural networks that contain one or more hidden layers.

Recurrent neural networks This type of neural networks processes time series, i.e. ordered and linked series of data. In the case of these networks, a neuron can receive as input data the output data that it previously calculated. We thus obtain a kind of short-term memory that can be useful in forecasting time series.

Long-short term memory neural networks These neural networks have a memory that helps them remember information from different time intervals. Network information is processed using three gates. The input gate determines which data will be used to update the memory. The forgetting gate determines that part of the previous data set to be forgotten. The output gate uses an activation function and determines which part of the cell state will be used as output. **Gated recurrent unit** is a simplified version combining the input and the forget gates.

Encoder-Decoder neural networks The architecture of the encoder-decoder model consists of three elements: encoder, decoder (composed of recurrent structures) and an intermediate vector. Each structure in the encoder processes an input element and encapsulates the results in the intermediate vector to increase the accuracy of the prediction. The decoder provides the prediction using the information from the intermediate vector.

Convolutional neural network is used in image recognition, having a feature extraction function and a great spatial information processing power .

In figure [17] a hybrid MTD-CNN-GRU model is proposed for $PM_{2.5}$ which combines convolutional neural network, gated recurrent unit (a kind of recurrent neural networks), and multi-task learning.

Some algorithms can generate a long-term forecast [10].

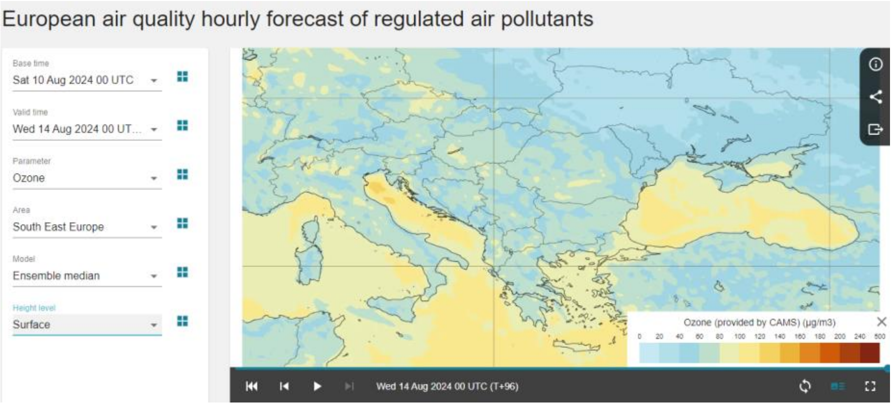
Algorithms based on neural networks [14, 18] or decision trees [14] for predicting the air quality index have been verified through numerous case studies.

3. APPLICATIONS FOR POLLUANTS FORECAST

There are several applications for pollution forecasting. Some of these are available online for free, being of great help to state authorities, but also to the population.

The Copernicus Atmospheric Monitoring Service (CAMS) is available to the public for free, accurate, and provides controlled information on air pollution. CAMS applications are based on a scientific approach to the study of the composition of the atmosphere and related processes. CAMS systems carry out forecasts and analyses of gases and aerosols on a global and also European level, but also estimates emissions. CAMS also proposes 4-day model forecasts for the main air pollutant concentrations in European capitals and other big cities cities.

CAMS wants to help the population understand the effects of reducing emissions on air quality and wants to show the area where the pollution originates. CAMS hopes that its reports and tools will be used by local and European authorities for immediate decisions, but also for long-term decisions related to environmental problems.



European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

Parameter: Ozone

Area: South East Europe

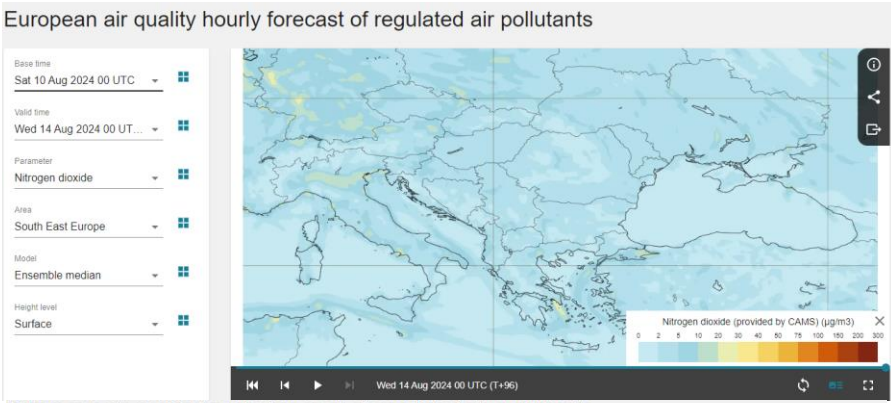
Model: Ensemble median

Height level: Surface

Ozone (provided by CAMS) (µg/m3)

Wed 14 Aug 2024 00 UTC (T+96)

Fig.1. Forecasting the main air pollutants in the south-eastern area of Europe in August 2024 using CAMS services (https://atmosphere.copernicus.eu/european-air-quality-forecast-plots)



European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

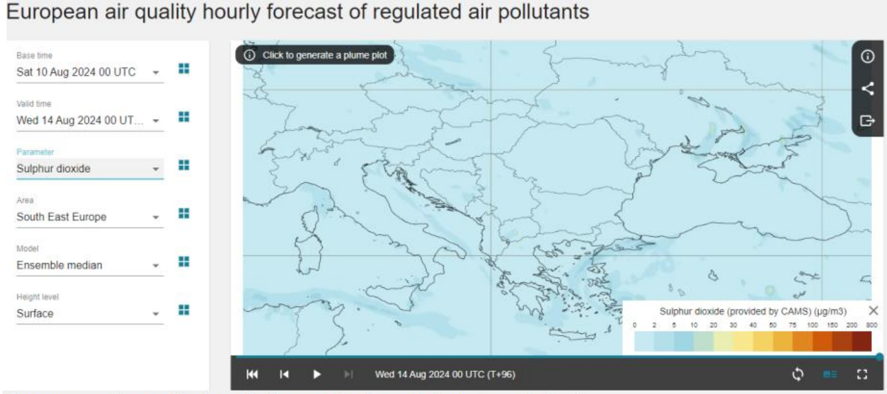
Parameter: Nitrogen dioxide

Area: South East Europe

Model: Ensemble median

Height level: Surface

Nitrogen dioxide (provided by CAMS) (µg/m3)
0 2 5 10 20 30 40 50 75 100 150 200 300
Wed 14 Aug 2024 00 UTC (T+96)



European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

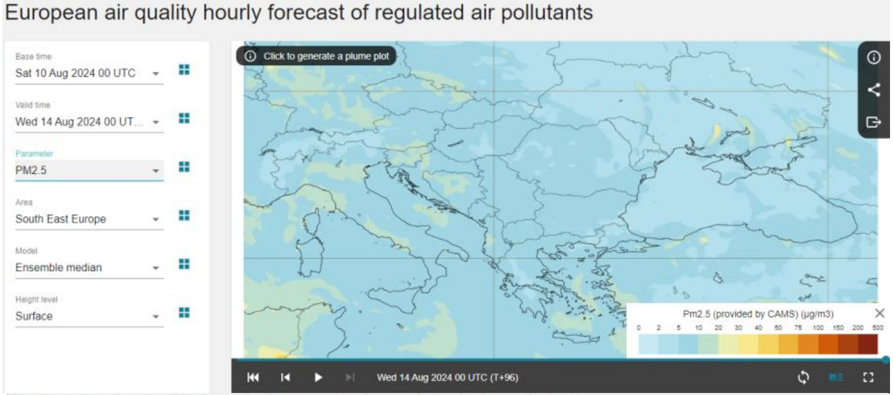
Parameter: Sulphur dioxide

Area: South East Europe

Model: Ensemble median

Height level: Surface

Click to generate a plume plot
Sulphur dioxide (provided by CAMS) (µg/m3)
0 2 5 10 20 30 40 50 75 100 150 200 800
Wed 14 Aug 2024 00 UTC (T+96)



European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

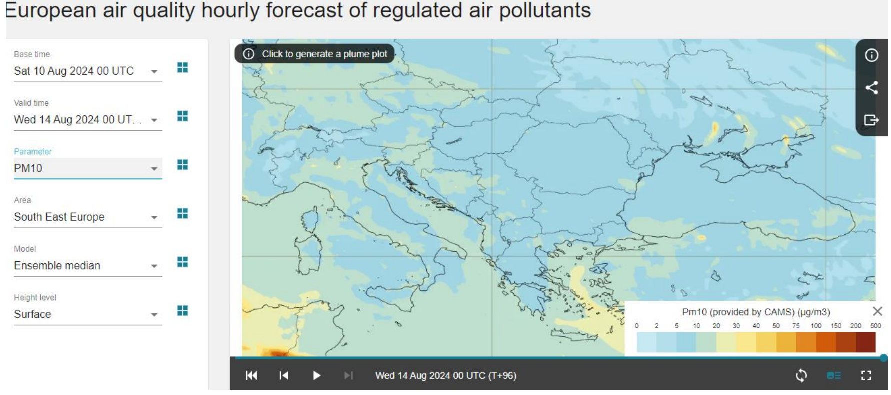
Parameter: PM2.5

Area: South East Europe

Model: Ensemble median

Height level: Surface

Click to generate a plume plot
Pm2.5 (provided by CAMS) (µg/m3)
0 2 5 10 20 30 40 50 75 100 150 200 500
Wed 14 Aug 2024 00 UTC (T+96)



European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

Parameter: PM10

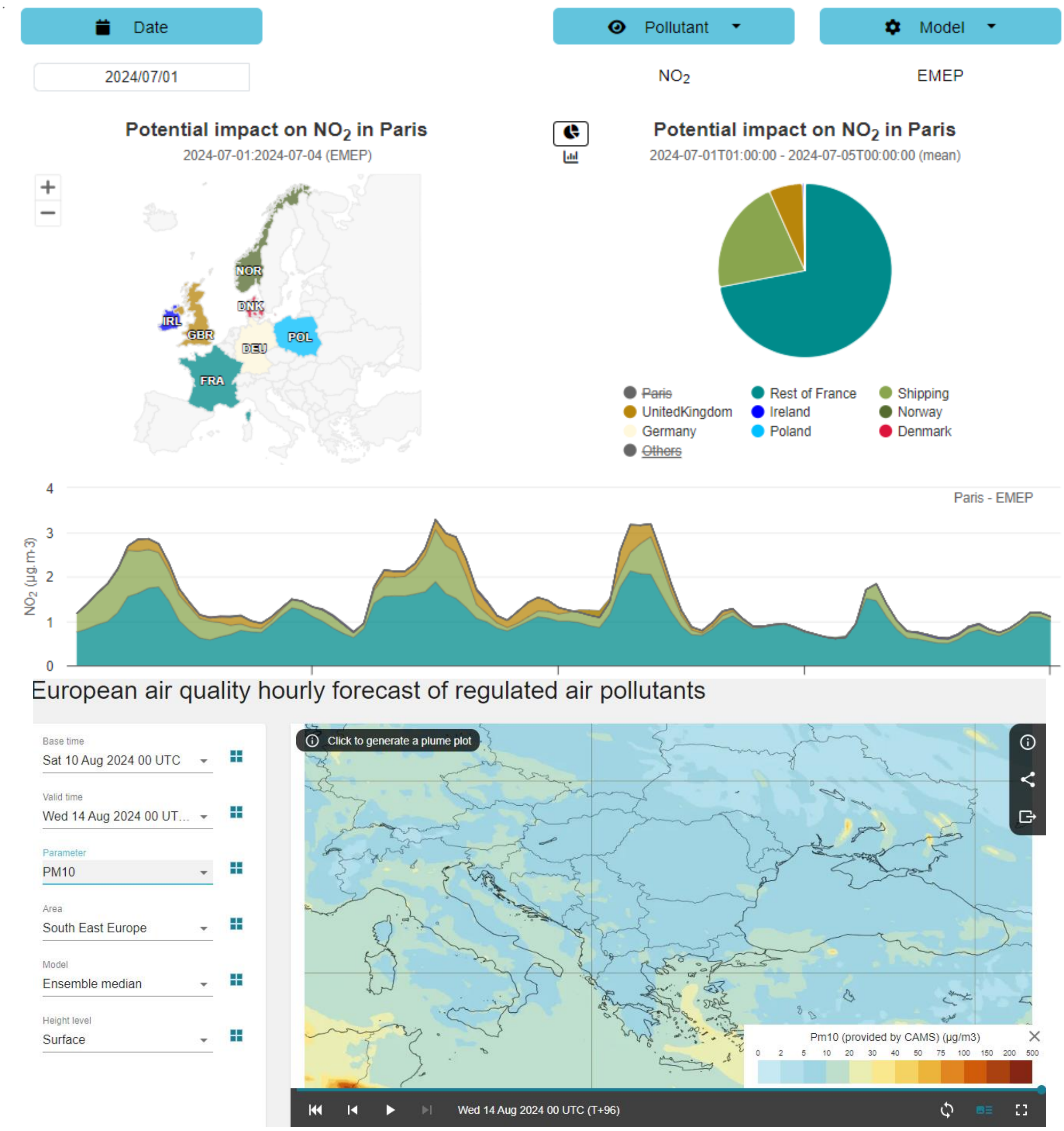
Area: South East Europe

Model: Ensemble median

Height level: Surface

Click to generate a plume plot
Pm10 (provided by CAMS) (µg/m3)
0 2 5 10 20 30 40 50 75 100 150 200 500
Wed 14 Aug 2024 00 UTC (T+96)

Fig.1. continued



Date: 2024/07/01

Pollutant: NO_2

Model: EMEP

Potential impact on NO_2 in Paris

2024-07-01:2024-07-04 (EMEP)

Potential impact on NO_2 in Paris

2024-07-01T01:00:00 - 2024-07-05T00:00:00 (mean)

European air quality hourly forecast of regulated air pollutants

Base time: Sat 10 Aug 2024 00 UTC

Valid time: Wed 14 Aug 2024 00 UT...

Parameter: PM10

Area: South East Europe

Model: Ensemble median

Height level: Surface

Fig.2. The impact of emissions reduction on NO_2 in Paris on July 1, 2024 calculated by the CAMS algorithms (https://policy.atmosphere.copernicus.eu/daily_source_attribution/country_impact.php?date=2024-07-01&pollutant=NO2)

CONCLUSIONS

The problem of air and water pollution is a particularly important one for people's health but also for that of the planet. Starting from this observation, we made a review of the main directions of development of algorithms for predicting the level of air or water pollution. We note that most of these algorithms are based on artificial techniques such as machine learning and neural networks.

Algorithms for predicting the level of air or water pollution have been verified and compared through various case studies. Therefore, there are implementations of them, some being available online for free.

The Copernicus Atmospheric Monitoring Service provides the public and the authorities with large amounts of air quality data, as well as interpretation and estimation products.

A further direction of research could be to compare the algorithms by highlighting the advantages and disadvantages of each one and also to compare the different available applications that these algorithms implement.

Acknowledgement This research was financially supported by Dunărea de Jos University of Galați, under the project RF 2488/31.05.2024.

References

1. Bharat Deshmukh S., Prakash Shirsat K., Prashant Dhotre S., Ravsaheb Jejurkar P., A survey on machine learning-based prediction of air quality index, IJARIIE-ISSN(O)-2395-4396, 7(6) (2021) 1203–1207.
2. Dobrea M., Bădicu A., Barbu M., Șubea O., Bălănescu M., Suciu G., Bîrdici A., Orza O., Dobre C., Machine Learning algorithms for air pollutants forecasting, IEEE 26th International Symposium for Design and Technology in Electronic Packaging (SIITME) (2020).
3. Drewil G.I., Al-Bahadili R.J., Air pollution prediction using LSTM deep learning and metaheuristics algorithms, Sensors 24 (2022) 100546.
4. Gacquer D., Delcroix V., Delmotte F., Piechowiak S., Comparative study of supervised classification algorithms for the detection of atmospheric pollution, Engineering Applications of Artificial Intelligence (2011) 1070–1083.
5. Garlík B., Přívětivý J., Artificial Intelligence Algorithms for Prediction and Diagnosis of Air Pollution Affecting Human Health, Journal of Physics: Conference Series, 2701, (2024) 012072.
6. Iskandaryan D., Ramos F., Trilles S., Air quality prediction in smart cities using machine learning technologies based on sensor data: a review, Applied Sciences, 10(7) (2020) 2401.
7. Koranga K., Pan P., Kumar T., Pant D., Kumar Bhatt A., Pan R.P., Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand, Materials Today: Proceedings 57(4) (2022) 1706-1712.
8. Kotu V., Deshpande B., Data Science (Second Edition). Concepts and Practice, Chapter 12 - Time Series Forecasting (2019) 395-445.
9. Méndez M., Merayo M.G., Núñez M., Machine learning algorithms to forecast air quality: a survey, Artificial Intelligence Review 56 (2023) 10031–10066.
10. Middya A.I., Roy S., Pollutant specific optimal deep learning and statistical model building for air quality forecasting, Environmental Pollution 301 (2022) 118972.
11. Nemade S., Mankar C., A Survey on Different Machine Learning Techniques for Air Quality Forecasting for Urban Air Pollution, International Journal for Research in Applied Science & Engineering Technology 7(IV) (2019) 2185–2194.
12. Oğuz1 K., Pekin M.A., Prediction of Air Pollution with Machine Learning Algorithms, Turkish Journal of Science & Technology 19(1) (2024) 1-12.
13. P. H. Soares, J. P. Monteiro, F. J.Gaioto, L. Ogiboski, C. M.s Gonçalves Andrade, Use of Association Algorithms in Air Quality Monitoring, Atmosphere 14 (2023) 648.
14. Sharma M., Jain S., Mittal S., Sheikh T.H., Forecasting and prediction of air pollutants concentrates using machine learning techniques: the case of India, IOP Conference Series: Materials Science and Engineering, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India, https:// doi. org/10. 1088/ 1757- 899x/ 1022/1/012123 (2020).
15. Wang Y., Liu H., Comparative Study of Algorithms Used in Water Pollution Prevention and Control, Advances in Artificial Intelligence, Big Data and Algorithms, Grigoras G. and Lorenz P. (Eds.), doi:10.3233/FAIA230931 (2023).
16. Yonar A., Yonar H., Modeling air pollution by integrating ANFIS and metaheuristic algorithms, Modeling Earth Systems and Environment 9 (2023) 1621–1631.

17. Zhang Q., Wu S., Wang X., Sun B., Liu H., A $PM_{2.5}$ concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations, Journal of Cleaner Production 275 (2020) 122722.
18. Zhou Y., De S., Ewa G., Perera C., Moessner K., Data-Driven Air Quality Characterization for Urban Environments: A Case Study, IEEE Acess 6 (2018) 77996–78006. doi:10.1109/ACCESS.2018.2884647 (2018).