Air Pollution Prediction as a Source for Decision Making Framework in Medical Diagnosis

Valerii Lovkin[a], Andrii Oliinyk[a] and Yurii Lukashenko[a]

[a] *National University "Zaporizhzhia Polytechnic", Zhukovsky str., 64, Zaporizhzhia, 69063, Ukraine*

Abstract

The problem of air pollution prediction is presented in the paper. It is considered regarding complex problem of creation of decision making framework in medical diagnosis. Therefore prediction is performed for a day, not for on an hour. The method of air pollution prediction is developed using Long Short Term Memory (LSTM) recurrent neural network. The LSTM-based model is used for prediction of concentration of separate air pollutant during the next day based on its concentration during the previous hours and average traffic data. The experimental investigation of the proposed method is performed by comparing it with ARIMA model, multilayer perceptron, vanilla recurrent neural networks and LSTM. The proposed method should be used in practice inside medical diagnosis tools and separate systems for air pollution analysis, enabling to obtain predicted air pollutant concentration level during the next day.

Keywords

Air pollution, road traffic, medical diagnosis, decision making framework, prediction, machine learning, long short term memory.

1. Introduction

Despite its huge spread for now, urbanization does not stop to increase. This process results in significant rise of concentration of population and human activity per square meter of territory in relatively small space. High human activity leads to large-scale economic changes, as well as to large emissions of heat, gases and waste, which as a result pollute air. Consequence of this process is detected by harmful impact on human health [1].

The described processes are already typical not only for industrial cities, but also for urban centers, where industrial production is not so highly influential. Analyzing the air quality index (AQI) in different cities of the world, it is seen that there are not only industrial centers of the world among the cities with low air quality. The list of top polluted cities also includes cities in Norway (Oslo), Poland (Krakow), Croatia (Zagreb) [2]. In the context of Ukraine, it should be noted that the level of pollution in Kyiv, which is the largest city of the country, currently prevails over industrial centers of the country during some periods of time. All these factors prove that the huge number of people in the world is affected by air pollution, and the problem of determining air quality level is widespread and important.

Air pollution [3] is determined by concentration of particles and gases in the air [4]. Regarding the problem of air pollution, it is important to monitor the current situation, analyze the accumulated data, and to predict the future level of air pollution. Such a prediction is significant in short and long terms. Prediction for medical diagnosis [5] is appropriate for the whole day or longer because of the specific character of medical examination and decisions made on treatment. The paper is aimed at prediction of air pollution during the next day.

IntelITSIS'2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine

EMAIL: vliovkin@gmail.com (V. Lovkin); olejnikaa@gmail.com (A. Oliinyk); lukashenkoyuriii@gmail.com (Y. Lukashenko)
ORCID: 0000-0002-6890-2807 (V. Lovkin); 0000-0002-6740-6078 (A. Oliinyk); 0000-0002-2478-4597 (Y. Lukashenko)

At the same time the whole obtained, calculated and predicted dataset enables complex solution of the problems of city management and medical diagnosis, because a man and a city, as well as biosphere in general, are the main objects of the impact of air pollution in the result.

This paper is devoted to the consideration of air pollution in terms of creating a decision-making framework for medical diagnosis, where prediction of the level of air pollution is actual for determination of the individual impact of air pollution on the patient.

2. Air pollution prediction problem statement

Medical diagnosis mainly consists in determining the patient's diagnosis. The obtained diagnosis becomes the basis for decisions to be made by doctor concerning further treatment of the patient. To determine the diagnosis it is necessary to form a heterogeneous set of data which characterizes the observed situation. On the one part this data is interlinked with subjective information about the patient which is determined during survey and various types of examination and on the other part is interlinked with the environment where patient lives.

The level of air pollution is one of the main indicators describing such an environment. Depending on the environmental conditions, it is possible to plan the specification of medical examination, which results are used in decision-making in diagnosis, and determine the specification of the implementation of decisions made based on diagnosis results. The whole set of decisions [6] made during a medical diagnosis forms a decision-making framework consisting of the following stages:

- making decisions concerning specification of the planned medical examination;
- making decisions concerning choice of diagnosis methods for the patient;
- making decisions concerning determination of the patient's condition;
- making decisions concerning the further treatment of the patient.

This group of decisions requires on the one hand the accumulation of historical data on air pollution, i.e. indicators of air pollution by certain substances collected at the relevant stations, and on the other hand prediction of the level of air pollution at these stations for the future.

The problem of air pollution prediction should be stated as determination of functional dependence between air pollutant concentration level during the next period of time and its concentration level during the previous periods of time together with additional parameter ε:

$$p_a^c(t, t+n-1) = f(p_a(t-1), p_a(t-2), \dots, p_a(t-n), e(t-n, t-1)), \quad (1)$$

where $p_a^c(t, t+n-1)$ is an average level of concentration of air pollutant a which is calculated for the time period which lasts from t till $t+n-1$, $p_a(t-1), p_a(t-2), \dots, p_a(t-n)$ present concentration of air pollutant a at the discrete corresponding moments $t-1, t-2, \dots, t-n$, f is a functional dependence which has to be found in the study, $e(t-n, t-1)$ is an additional parameter, which presents additional factors, which influence p_a^c, should be calculated for the time period which lasts from $t-n$ till $t-1$, and doesn't depend on the corresponding air pollutant.

Each air pollutant a is an element of the set of air pollutants A, including gases and particulates. This set should be created, including all air pollutants which are factors of decision making in medical diagnosis.

The shape of functional dependence f should be investigated based on machine learning methods. The main attention should be paid to recurrent neural networks and LSTM recurrent neural networks, because of the nature of sequence $p_a(t-1), p_a(t-2), \dots, p_a(t-n)$ in the problem (1).

At the same time these discrete air pollutants form Air Quality Index (AQI), which is represented by categorical value used for monitoring and decision making. AQI is obtained on the basis of levels of the following pollutants:

- ground-level ozone (O_3);
- particle pollution ($PM_{2.5}$ and PM_{10});
- carbon monoxide (CO);
- sulfur dioxide (SO_2);
- nitrogen dioxide (NO_2) [7].

It means that accurate prediction of concentration of these pollutants is critical for accurate prediction of AQI.

3. Related works

A number of studies concerning air pollution prediction using machine learning methods has been conducted.

Concentration of $PM_{2.5}$ is estimated in the study [8] using regression models. The proposed models should be applied for countries where there is no possibility to use costly sensors to monitor air pollution and to create dataset which is necessary for prediction by sequence-based methods. Prediction is performed using real-time traffic monitoring based on Google Maps. Separate models were built for different periods of day. Trace gas concentrations are observed only as additional data for the environment where it could be accumulated. Regression models don't enable to take into account complex relations between data sequences and different types of factors which influence air pollution.

Prediction of concentration of particulate matters using regression models was broadened out by prediction of concentration of PM_{10} in the study [9]. Regression models were used to predict concentration level during the next day.

Method for pattern analysis using dynamic time warping was proposed in the study [10]. This method needs data on $PM_{2.5}$ concentration from multiple stations and prediction is performed based on similarity between stations. k-nearest neighbour method, which calculates dynamic time warping as distance between stations by using its geographical coordinates, is used.

In the paper [11] support vector regression model was used to predict concentration of separate air pollutants and to predict general pollution level based on the AQI. The following air pollutants were studied: carbon monoxide, sulfur dioxide, nitrogen dioxide, ground-level ozone, particulate matter 2.5. Prediction was realized on an hourly basis. Appropriate results were obtained for O_3, CO and SO_2, that's why this approach couldn't be recommended for universal usage.

The study [12] is dedicated to the relationship between air pollution and urban transport networks. Artificial neural network model based on multilayer perceptron and the ARIMAX model are compared using experimental investigation. Prediction is performed for an hour. It is proposed to use ensemble model based on both models to process specific situations. Such an ensemble actually models influence of transport network on nitrogen dioxide concentration in the city air, so it doesn't take into account other factors which influence on the air quality as well as other air pollutants. Besides such a model does not consider sequences which exist in the history of air pollutant concentration.

Deep learning model based on LSTM neural networks was investigated in the paper [13] where it is presented in the context of Internet of Things concept [14, 15]. The proposed model is aimed at AQI prediction, so the obtained results are categorical. During experimental investigation separate LSTM models were created for ozone and nitrogen dioxide gases. The obtained results indicated that sequence-based approach for air quality prediction is perspective and could be used in practice.

LSTM-based model is used in the paper [16] to predict $PM_{2.5}$ concentration in the air of South Korea. The prediction was performed for long-term periods. Different time horizons, including 8, 12, 16, 20, 24 hours, were investigated. It confirmed possibility to predict air pollution for intervals longer than 1 hour.

In the paper [17] LSTM neural networks and deep autoencoders were used for PM concentration prediction. $PM_{2.5}$ and PM_{10} were investigated using datasets of Seoul. Prediction was performed for 10 days after period which was studied. During the experimental investigation LSTM models demonstrated better results, therefore there is no practical need in the usage of deep autoencoders for air pollution prediction problem.

The study [18] is aimed at road traffic prediction based on air pollution. CO, NO, NO_2, NO_x and O_3 are the observed air pollutants. Prediction was realized based on LSTM neural network architecture. But at the same time air pollution is not a reason of road traffic but its consequence. So it should be possible to improve air pollution prediction using road traffic data because road traffic is one of the main reasons of polluted air in big cities.

The following study of the features which could be used for LSTM model to perform a prediction in a day is needed. The problem of feature selection was considered in the studies of authors [19] and should be applied to air pollution.

4. Proposed method of LSTM-based air pollution prediction using traffic data

Air pollution prediction is performed separately for each air pollutant, therefore prediction models should be created for each air pollutant from the set *A*.

As a result of learning of air pollution nature and literature review the hypothesis on the dependence between air pollution concentration and traffic data was moved. For further analysis of this hypothesis separate investigation of correlation between city traffic data and concentration of air pollutants was performed. The investigated correlation is positive. Its visualization is presented in the Figure 1.

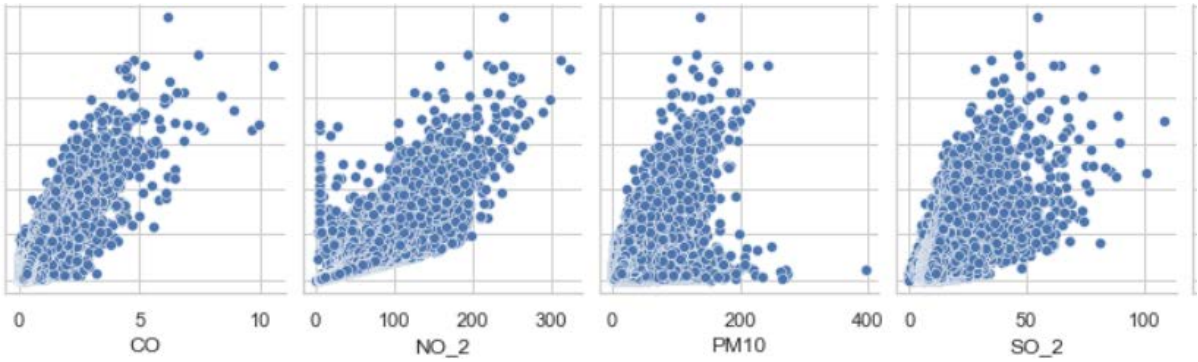

Figure 1: Graph of correlation between road traffic and concentration of air pollutants (CO, NO_2, PM_{10}, SO_2)

The problem (1) is solved using LSTM model [20, 21, 22, 23]. This decision was made because input data are characterized by sequential nature, so this problem is a time series prediction problem. The proposed structure of the model is presented in the Figure 2.

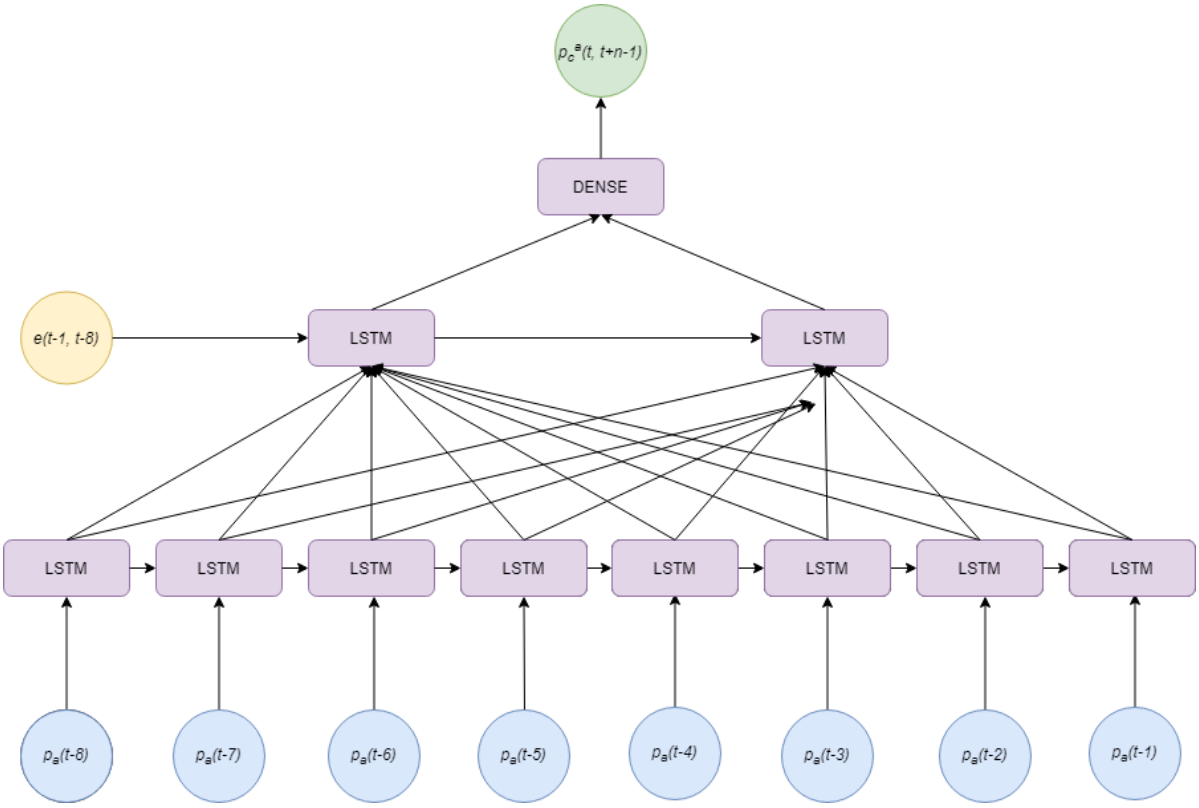

Figure 2: Structure of the model which is used for air pollutant prediction in the proposed method

The proposed model consists of two LSTM layers: the first layer of neurons interacts with input data, the second layer is a hidden layer. The first layer is proposed to build from 8 neurons. Each input neuron gets arithmetic mean value of concentration of air pollutant within 3 hours. Then obtained values are processed by the hidden layer which consists of 2 LSTM-neurons. Amount of traffic impacts on one of the neurons of the hidden layer, and hence impacts on the final prediction.

Dataset which is needed for the model training in the proposed method should be prepared in the following way. Each sample represents values of parameters during a day and air pollutant concentration for the following day registered in a separate station. Parameters include one value of average day traffic in the region of station and 8 values of an air pollutant concentration during a day. 8 values were used instead of 24, because fluctuations of air pollutant within 3 hours are insignificant. Air pollutant concentration during the following day is presented by median value calculated using 24 values of air pollutant concentration. The obtained dataset is normalized before model training.

Concentration data during the current day should be registered in one station and should be used for prediction by the trained models which present each air pollutant in practice. Predictions are performed by the models for the following day. The obtained values of concentration of each air pollutant should be used for decisions on specificity of patient treatment.

5. Results

The main dataset [24], which was used for experimental investigation, represents air pollution concentration levels registered during 18 years (from 2001 to 2018) in different stations in Madrid on hourly basis. Approximately 150 thousands of measurements were performed for each air pollutant in a station. Air pollutants presented in the dataset include SO_2, CO, NO, NO_2, $PM_{2.5}$, PM_{10}, NO_x, O_3, TOL, BEN, EBE, MXY, PXY, OXY, TCH, CH_4, NMHC [24]. Not all stations presented in this dataset had measurements for full list of air pollutants which are included in AQI during full period from 2001 to 2018: some measurements are missed because of the absence of some equipment, its repair or unavailability. The example of missed (white color) and obtained (black) values is presented in the Figure 3.

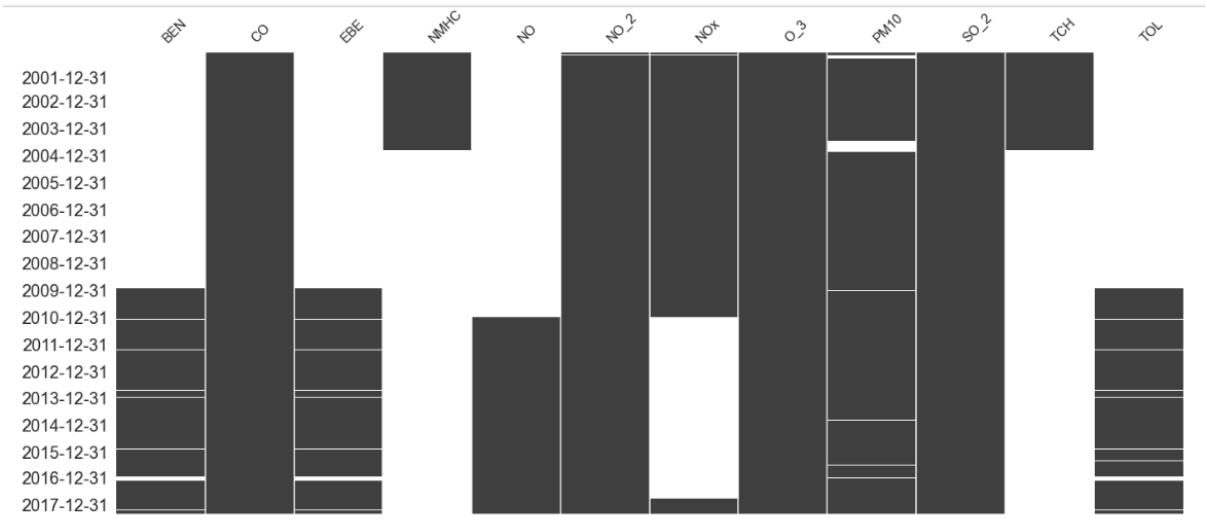

Figure 3: Missed values of concentration of air pollutants in the dataset under investigation

For experimental investigation only stations with equipment for registration of all air pollutants from AQI were chosen. Missed values were replaced by averaging.

Traffic data for the chosen stations were obtained from Madrid's City Council Open Data website [25].

The whole dataset was divided into learning sample (80 %) and test sample (20 %).

During experimental investigation the following models and methods were used for the problem solving: ARIMA model, artificial neural network based on multilayer perceptron [22], vanilla

recurrent neural network [21], LSTM model [20], the proposed method which uses the same LSTM model and traffic data as additional parameter. Software was developed for the investigation using Python programming language. Keras library was used for neural network models realization.

For estimation of the obtained results metrics of root mean square error (RMSE) and mean absolute error (MAE) were used:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - A_i)^2}, \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\lfloor E_i - A_i \rfloor, \quad (3)$$

where n is a number of samples in test dataset, A_i is actual value (i-th sample from test dataset), E_i is a predicted value with number i.

The results of the conducted experimental investigation using these metrics were accumulated in the Table 1.

Table 1

Comparison of RMSE and MAE values calculated for the existing models and the proposed method for air pollutant O_3

Prediction model/method	RMSE	MAE
ARIMA model	15.33	10.95
Multilayer perceptron	19.10	12.87
Recurrent neural network	14.95	10.63
LSTM	13.87	9.13
The proposed method	12.71	7.22

The obtained results demonstrate that LSTM model is characterized by better values of RMSE and MAE than ARIMA model which is a classic solution for time series prediction problem, traditional multilayer perceptron and vanilla neural network. ARIMA model allowed to obtain better results than multilayer perceptron. At the same time additional usage of traffic data for model input allowed to perform prediction with RMSE which is 9.13 % smaller and MAE which is 20.92 % smaller than LSTM model without additional parameter.

Another metric was proposed to estimate accuracy of prediction of all air pollutant concentration from AQI. This estimation was performed calculating percent of samples from dataset for which prediction error was not larger than the limit (for example, 0.5 mcg/m^3 for O_3). The obtained results are presented in the Table 2.

Table 2

Comparison of prediction accuracy for the test sample

Prediction model/method	Accuracy, %
ARIMA model	86.35
Multilayer perceptron	83.96
Recurrent neural network	86.75
LSTM	88.62
The proposed method	90.98

The proposed method is characterized by the best results between the considered models and methods. The accuracy of the proposed method is 4.63 % better than accuracy of ARIMA model and 2.36 % better than accuracy of LSTM model.

Prediction made for Arturo Soria station by LSTM model, created without traffic data and with traffic data based on the procedure of the proposed method, is presented in the Figure 4. To accent the

differences between character of values of the previous day and the next day for which prediction is performed, predictions are visualized with a day interval.

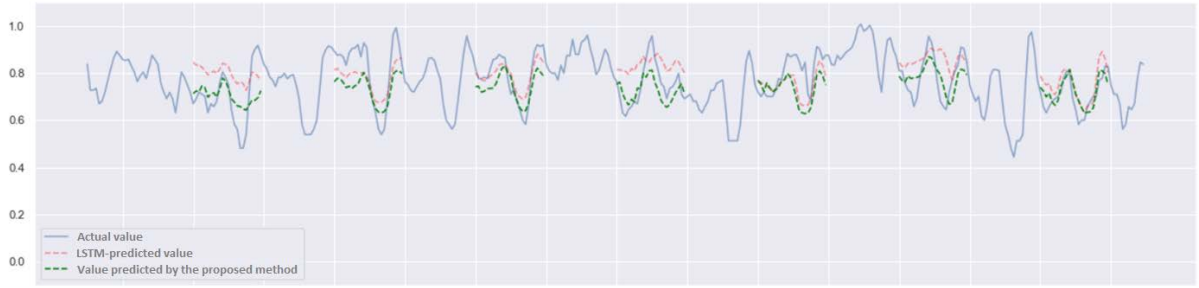

Figure 4: Predicted values of O_3 concentration in Arturo Soria station

6. Conclusion

Air pollution prediction problem is considered from the point of view of decision making in medical diagnosis. Main features of such decisions within the decision making framework are presented.

Mathematical formalization of the air pollution prediction problem is made. Method of the problem solution is presented. Prediction model in the method is organized using LSTM neural network and consists of 2 LSTM layers. Road traffic data is used for additional presentation of environment as a factor which impacts on air pollution. Data preparation procedure is described in the paper.

Experimental investigation of the proposed method is performed using dataset collected in Madrid during 18 years. ARIMA model, multilayer perceptron, vanilla recurrent neural network and LSTM are used as alternatives. The model, which was trained according to the proposed method, allowed to obtain better results, including smaller values of RMSE, MAE and better accuracy level.

The proposed method should be used in practice inside medical diagnosis tools and separate systems for air pollution analysis, enabling to predict air pollutant concentration level during the next day.

7. Acknowledgments

The work was performed as part of the research work "Development of methods and tools for analysis and prediction of dynamic behavior of nonlinear objects" (state registration number 0121U107499) of Software Tools Department of National University “Zaporizhzhia Polytechnic”.

We are particularly grateful for the assistance with data sample which was given by Diego Vicente, Junior Data Scientist at Decide Soluciones in Madrid, Spain.

8. References

[1] R. F. Phalen, R. N. Phalen, Introduction to Air Pollution Science: A Public Health Perspective, Jones & Bartlett Learning, Burlington, MA, 2011.

[2] Air quality and pollution city ranking, 2021, URL: https://www.iqair.com/world-air-quality-ranking.

[3] D. A. Vallero, Fundamentals of Air Pollution, 5th ed., Academic Press, Waltham, MA, 2014.

[4] Air Pollution: MedlinePlus, 2021, URL: https://medlineplus.gov/airpollution.html.

[5] A. Oliinyk, S. Subbotin, V. Lovkin, S. Leoshchenko, T. Zaiko, Development of the indicator set of the features informativeness estimation for recognition and diagnostic model synthesis, in: Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering : 14th International Conference TCSET'2018, Lviv-Slavske, Ukraine, 2018, pp. 903-908. doi: 10.1109/TCSET.2018.8336342.

[6] T. Kolpakova, A. Oliinyk, V. Lovkin, Improved method of group decision making in expert systems, in: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, 2017, pp. 939–943. doi: 10.1109/UKRCON.2017.8100388.
[7] Air Quality Index (AQI) Basics, 2021, URL: https://www.airnow.gov/aqi/aqi-basics/.
[8] Y. Rybarczyk, R. Zalakeviciute, Regression Models to Predict Air Pollution from Affordable Data Collections, in: H. Farhadi (Ed.), Machine Learning - Advanced Techniques and Emerging Applications, InTech, London, 2018, pp. 15-48. doi: 10.5772/intechopen.71848.
[9] M. T. Lei, J. Monjardino, L. Mendes, D. Gonçalves, F. Ferreira, Macao air quality forecast using statistical methods, Air Quality, Atmosphere & Health 3 (2019) 249-258. doi: 10.2495/EI-V2-N3-249-258.
[10] P.-W. Soh, K.-H. Chen, J.-W. Huang, H.-J. Chu, Spatial-temporal pattern analysis and prediction of air quality in Taiwan, in: 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), Pattaya, Thailand, 2017, pp. 1-6. doi: 10.1109/UMEDIA.2017.8074094.
[11] M. Castelli, F. Martins Clemente, A. Popovič, S. Silva, L. Vanneschi, A Machine Learning Approach to Predict Air Quality in California, Complexity 2020 (2020) 1-23. doi: 10.1155/2020/8049504.
[12] M. Catalano, F. Galatioto, M. Bell, A. Namdeo, A. Bergantino, Improving the prediction of air pollution peak episodes generated by urban transport networks, Environmental Science & Policy 60 (2016) 69-83. doi: 10.1016/j.envsci.2016.03.008.
[13] I. Kok, M. Simsek, S. Ozdemir, A deep learning model for air quality prediction in smart cities, in: 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 1983-1990. doi: 10.1109/BigData.2017.8258144.
[14] J. A. Alsayaydeh, V. Shkarupylo, M. S. Hamid, S. Skrupsky, A. Oliinyk, Stratified Model of the Internet of Things Infrastructure, Journal of Engineering and Applied Sciences, 13 (2018) 8634-8638. doi: 10.3923/jeasci.2018.8634.8638.
[15] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo, C. Pellipus, Homes appliances control using bluetooth, ARPN Journal of Engineering and Applied Sciences, 14 (2019) 3344-3357.
[16] T.-C. Bui, V.-D. Le, S. K. Cha, A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM, 2018, URL: https://arxiv.org/abs/1804.07891.
[17] T. Xayasouk, H. Lee, G. Lee, Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models, Sustainability 12 (2020) 2570-2577. doi: 10.3390/su12062570.
[18] F. Awan, R. Minerva, N. Crespi, Improving Road Traffic Forecasting Using Air Pollution and Atmospheric Data: Experiments Based on LSTM Recurrent Neural Networks, Sensors 20 (2020) 3749-3769. doi: 10.3390/s20133749.
[19] A. Oliinyk, S. Subbotin, V. Lovkin, S. Leoshchenko, T. Zaiko, Feature selection based on parallel stochastic computing, in: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018 - Proceedings, Lviv, Ukraine, 2018, pp. 347-351. doi: 10.1109/STC-CSIT.2018.8526729.
[20] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, Cambridge, Massachusetts, 2017.
[21] J. D. Kelleher, Deep Learning, The MIT Press, Cambridge, Massachusetts, 2019.
[22] C. C. Aggarwal, Neural Networks and Deep Learning: A Textbook, Springer, Yorktown, NY, 2018.
[23] S. Leoshchenko, A. Oliinyk, S. Subbotin, T. Zaiko, Using Modern Architectures of Recurrent Neural Networks for Technical Diagnosis of Complex Systems, in: Proceedings of the 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 2018, pp. 411-416. doi: 10.1109/INFOCOMMST.2018.8632015.
[24] Air Quality in Madrid (2001-2018), 2018, URL: https://www.kaggle.com/decide-soluciones/air-quality-madrid.
[25] En portada – Portal de datos abiertos del Ayuntamiento de Madrid, 2021, URL: https://datos.madrid.es/portal/site/egob.