Application of machine learning methods to the prediction of NO_2 concentration in the air environment

Iryna Didych[1,*,†] , Andrii Mykytyshyn[1,*,†] , Andrii Stanko[1,†], Mykola Mytnyk [1,†]

[1] *Ternopil Ivan Puluj National Technical University, Ruska 56, 46001 Ternopil, Ukraine*

Abstract

Air quality significantly impacts public health, with nitrogen dioxide (NO_2) being a key pollutant linked to respiratory and cardiovascular diseases. In this study, we developed a machine learning model to accurately predict hourly NO_2 concentrations in Ternopil, Ukraine, using readily available meteorological and temporal data. The model was trained on a large dataset and tested using data from the Ecocity monitoring station, known for recording NO_2 levels exceeding legal limits. By employing neural networks, the model demonstrated high accuracy in predicting NO_2 concentrations, with the error of 3.9% and 1.4%, respectively, in the test samples. Our findings underscore the potential of machine learning techniques to enhance air quality monitoring and forecasting, particularly in urban areas with limited resources. This approach offers a valuable tool for real-time pollution management and public health protection.

Keywords

Air quality, nitrogen dioxide, prediction, machine learning

1. Introduction

Air quality is a complex, multifactorial set of chemical, physical, and biological characteristics of air, and at the same time a very relevant topic because of its connection to human health. Numerous studies have demonstrated the link between cardiovascular and lung diseases and long-term exposure to pollutants, in particular nitrogen dioxide (NO_2) and particulate matter (PM2.5 and PM10). According to the European Environment Agency [1], in 2018, about 55,000 premature deaths in the EU could be attributed to exposure to NO_2. The results of several clinical and epidemiological studies show that there is at least moderate evidence that adverse health effects occur even with short-term exposure to pollutants, such as exposure below established limits [2].

Increasing concentrations of pollutants in the atmosphere have changed its properties, making it a harmful environment for humans and other living organisms [3]. Pollutants include

ITTAP'2024: 4th International Workshop on Information Technologies: Theoretical and Applied Problems, October 23-25, 2024, Ternopil, Ukraine, Opole, Poland

[1,*] Corresponding author.

[†] These authors contributed equally.

iryna.didych@tntu.edu.ua (I. Didych); mikitishin@gmail.com (A. Mykytyshyn); stanko.andrjj@gmail.com (A. Stanko) ; mytnyk@networkacad.net (M. Mytnyk).

0000-0003-2846-6040 (I. Didych); 0000-0002-2999-3232 (A. Mykytyshyn); 0000-0002-5526-2599 (A. Stanko); 0000-0003-3743-6310 (M. Mytnyk)

a variety of gases, droplets, and particles that degrade air quality; therefore, their exposure to humans is believed to lead to serious health problems, especially in urban areas where pollution levels are high [4]. Air pollutants are chemical, physical (e.g., particulate matter), or biological agents that alter the natural characteristics of the atmosphere. Particulate matter, cited as an example of air pollutants, is the main factor that negatively affects human health due to its high toxicity. Air pollution results in the presence of certain gases in the atmosphere in concentrations that exceed the standard and can be seriously harmful to human health. Examples of such gases are nitrogen oxides, sulfur oxides, carbon monoxide, photochemical oxidants (e.g. ozone), lead, as well as various heavy metals and volatile organic compounds that are released into the atmosphere as a result of industrialization and transport emissions, thereby degrading air quality. Air quality is defined as the state of the atmosphere in the environment that can be affected by pollution from these sources.

Air quality is considered good when it meets a certain level of purity and transparency, and there are no gaseous pollutants such as smoke, dust, smog, and other impurities in the atmosphere. The Air Quality Index (AQI) is a numerical indicator used by government agencies to inform the public about the current state of the air or projected levels of air pollution. When the AQI rises due to an increase in air pollutants (e.g., during peak traffic hours or when there are downwind wildfires), a growing proportion of the population can quickly experience serious negative health effects [3]. The US Environmental Protection Agency has classified these air pollutants into six main categories [5].

The issue of constant monitoring of air quality in real time is also relevant, as the authors says in paper [6] and [7]. Several studies have been conducted to monitor air quality in the environment. Air quality assessments were carried out in the city of Ternopil, Ukraine, at the Ivan Puluj National Technical University of Ternopil.

The purpose of this study is to assess and analyse the impact of atmospheric pollution on air quality, as well as to establish its dependence on atmospheric factors.

Many studies have established a link between cardiovascular and pulmonary diseases and long-term exposure to pollutants, including nitrogen dioxide (NO_2) and particulate matter (PM2.5 and PM10). According to the European Environment Agency [1], in 2018, approximately 55,000 premature deaths in the EU were attributed to NO_2 exposure. Data from several clinical and epidemiological studies show at least moderate evidence that adverse health effects can occur even with short-term exposure to pollutants, including exposures below established limit values [2].

Natural NO_2 emissions have very low background concentrations. Emissions associated with anthropogenic activities are the most significant factor affecting human health. The main source of NO_2 emissions is human-caused combustion processes, such as heat, electricity, or internal combustion engines. In particular, motor vehicles are the main sources of nitrogen oxides [8]. However, it is important to investigate the dependence of NO_2 concentration on changes in temperature and humidity in the environment.

Such an analysis requires models that can accurately reflect air quality in order to identify periods of increased pollution and respond quickly to short-term fluctuations, especially in urban areas. Prediction of pollutants by deterministic methods is accompanied by significant uncertainties due to the complexity of the physical and chemical processes that determine the formation and transport of pollutants in the urban atmosphere [9-11].

In this regard, sophisticated machine learning methods are becoming increasingly common in air quality modelling, outperforming traditional statistical approaches. In the review by Cabañeros et al. [12], which analysed the number of studies using artificial neural networks (ANNs) to model pollutants since 2001, found 139 papers. Of these, 51 studies applied this method to predict nitrogen oxides, while others focused on modelling various pollutants such as particulate matter (PM), carbon dioxide (CO_2), and ozone (O_3).

ANNs are capable of detecting complex, nonlinear relationships between meteorological variables and pollutant concentrations, and generalizing information from training datasets to form functional relationships between variables, even if the nature of these relationships is unknown. Unlike regression analysis, ANNs work effectively in the presence of significant noise in the data [13].

The first successful applications of ANNs for modelling NO_2 concentrations in urban areas were presented in the works of Gardner and Dorling [14], Kolemainen et al. [15]. The authors in [16], which demonstrated the advantages of the proposed approaches over regression models. Since then, many modern studies have also obtained significant results in the use of neural networks for modelling nitrogen oxides, simulating both national emissions over long periods [17] and local emissions on an hourly basis [11]. Some studies have included several air pollutants in the models, such as Jiang et al. [18], where a combination of a neural network and a heuristic algorithm was used to develop an early warning system for five different pollutants.

A key aspect in developing machine learning-based air quality models is the selection of appropriate input parameters. The concentration of NO_2 in urban air is influenced by many variables that reflect meteorological conditions and pollution sources. Several studies have identified meteorological variables such as temperature, humidity, and wind speed as important predictors, as well as concentrations of other pollutants [19].

An alternative approach is to use previously measured concentrations of the target pollutant as predictors, relying on temporal autocorrelation between successive values of the same variable. This is especially effective for forecasting several hours in advance [20]. In some studies, this method was combined with long short-term memory (LSTM) recurrent networks to successfully predict NO_2 up to eight hours in advance [21]. Dai et al. [22] combined LSTMs with convolutional neural networks (CNNs) to create a model suitable for predicting six different pollutants. While these models often perform well, they are more computationally intensive than simple feedforward networks.

Other studies have used traffic data obtained by vehicle counts or other models [23]. Since traffic is one of the main contributors to elevated NO_2 concentrations, traffic statistics have a high predictive value.

2. Materials and Methods

2.1. Study Area

In our study, we set out to develop a model that would accurately estimate hourly NO_2 concentrations in Ternopil, Ukraine, using only available standard meteorological and temporal data as input variables.

As in many cities, air quality in Ternopil is monitored by several separate measurement stations. In previous years, several of them have recorded high levels of NO_2 that exceeded the

regulatory limits. For this study, the Ecocity station in the central part of Ternopil, known as a pollution hotspot, was chosen, where NO_2 concentrations often exceed the legal thresholds. This location is particularly important due to the high density of residential development in the immediate vicinity of the station.

Ternopil is located in western Ukraine, near the Seret River, on the Ternopil Plateau of the Podillia Upland of the Eastern European Plain. The city is located in the temperate climate zone of the broadleaf forest zone. Ternopil has a moderately continental climate with warm and humid summers and mild winters [24].

The street where the air quality monitoring station used in this study is located is in the central part of the city. It is a heavily trafficked main transportation artery connecting the highway with the centre of Ternopil. At the same time, there is a very high density of residential buildings along this street. For several months, both the concentrations of pollutants measured at this station and the number of days with peak pollution levels exceeded the established permissible limits.

2.2. AirFresh air quality monitoring station

The correct choice of input parameters, is crucial for predicting pollutant concentrations. Ivan Puluj National Technical University in Ternopil (Ukraine), in cooperation with the program "Clean Air for Ukraine" of the NGO Arnika (Prague, Czech Republic), NGO Free Arduino (Ivano-Frankivsk, Ukraine) and the public monitoring network EcoCity, installed an AirFresh air quality measurement station at the university to expand the network and conduct research. The AirFresh air quality monitoring station is a device that enables real-time monitoring and recording of ambient air conditions, namely temperature, humidity, and dust concentrations of PM2.5 and PM10. AirFresh measures the concentrations of dust microparticles (PM2.5 and PM10), carbon monoxide (CO), ammonia (NH_3), ground-level ozone (O_3), and nitrogen dioxide (NO_2) as shown in Figure 1. Each station may be outfitted with radiation background sensors or additional sensors for 16 pollutants tracked by EcoCity.[25-27].

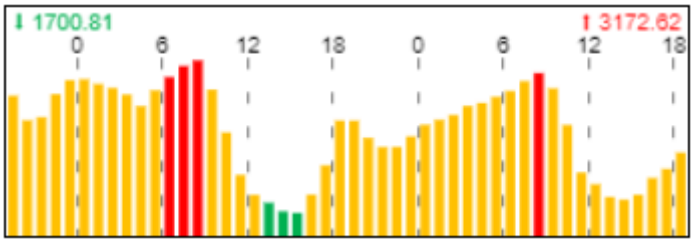


Figure 1: Daily measurements of nitrogen dioxide (NO_2) levels by the AirFresh station in Ternopil

Meteorological parameters, such as the height of the measurement site, have a significant impact on the concentration of NO_2. At the monitoring station, data on pollutant concentrations, as well as temperature and relative humidity, are measured at a height of 3 meters.

Experimental data on NO_2 concentrations, temperature and humidity were collected at the station during the period from August 6 to 13, 2024.

To ensure the ability to recognize temporal patterns in NO_2 fluctuations, the model was given the ability to detect typical hourly and daily variations of this pollutant. To this end, the analysis included time variables that reflect different frequency components in the observed data.

3. Results and discussion

The development of a machine learning model consists of several stages, each of which plays an important role in creating an efficient and accurate model. The main stages of developing a machine learning model are: data collection; data preprocessing and analysis; selection of a model and machine learning algorithm; splitting the data into training and test samples; model training; and evaluation and validation. Tracking NO_2 emissions, which is the most active pollutant gas, and predicting its concentration are important steps towards pollution control. Therefore, nitrogen dioxide (NO_2) was predicted using experimental data obtained from the Meteorological Station AirFresh. During learning, the dataset was divided into two unequal parts training and test samples.

The study was divided into two stages. At the first stage, the training set contained experimental dependencies of NO_2 concentration on temperature, humidity, and measurement time for six days, and a sample of a one-day dataset unknown to the system was chosen to test the quality of forecasting. And at the second stage, the test sample was randomly selected by the computer from all the experimental data for different days (Figure 2).

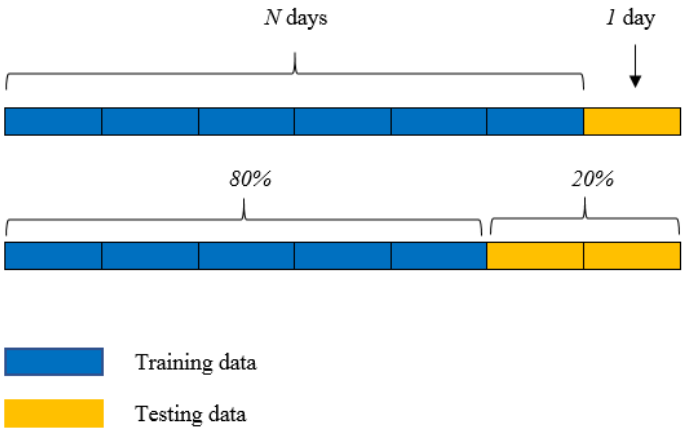
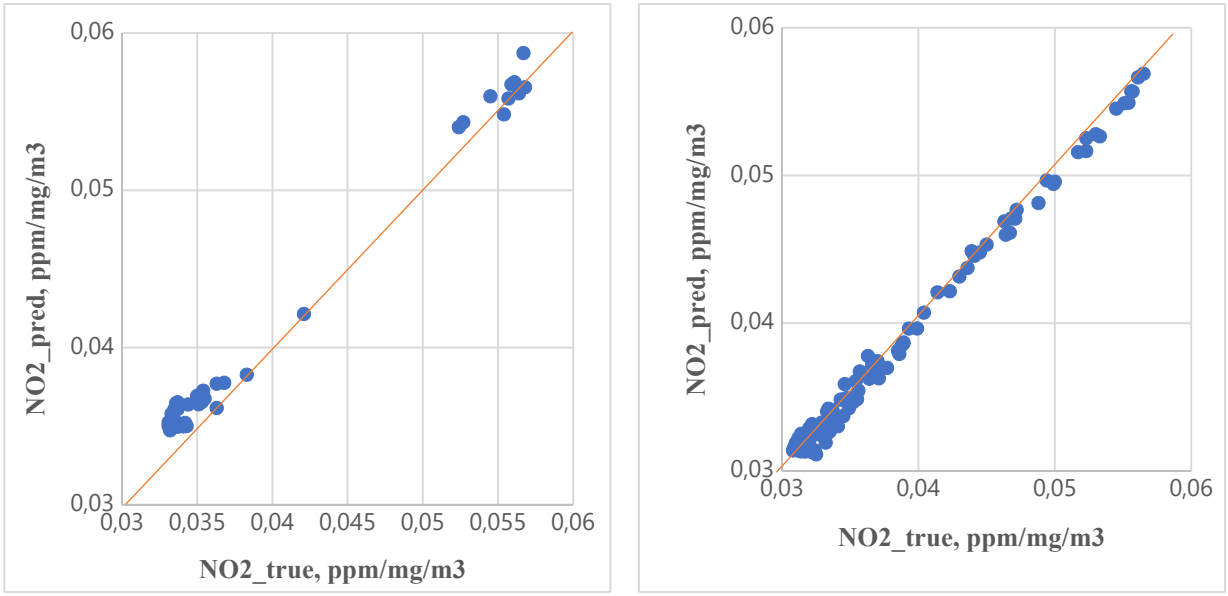


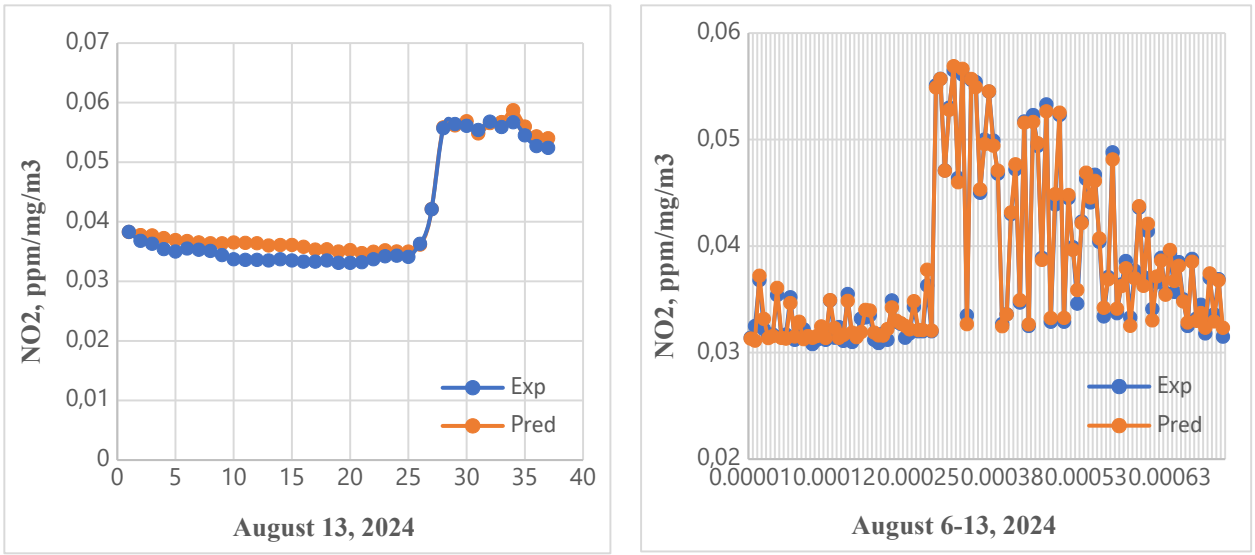
Figure 2: Comparison of the two predicting architectures for training and testing

The dataset contained 541 elements. In particular, the training set contained 504 elements characterising temperature, humidity and measurement time over N days (six days in our study). The NO_2 concentration was predicted by neural networks and selected as the output parameter. Based on the experimental results of the NO_2 concentration for one day, a test set of 37 elements was formed to evaluate the quality of prediction. It was found that the built models can make predictions based on data that were not used in the training sample. Therefore, such results are informative for studying their quality.



a) b)

Figure 3: The predicted ($NO2_{pred}$) and experimental ($NO2_{true}$) concentrations during August 2024, in particular, a) August 13 and b) August 6-13 in test sample by method of neural networks



a) b)

Figure 4: The predicted and experimental dependences of NO_2 concentrations during August 2024, in particular, a) August 13 and b) August 6-13 in test sample.

The neural networks were used to build (Figure 3 and 4) the dependence of experimental concentrations ($NO2_{true}$) on the predicted ones ($NO2_{pred}$), as well as NO_2-August 13, 2024 for one day. The NN method gives an error of 3.9%.

It is important that in Figure 3, the points are located quite close to the bisector of the first coordinate angle, which indicates the consistency of the predicted and experimental data.

The NO_2 concentration was predicted by temperature, humidity and time of measurement over seven days. The sample contained 541 elements, of which 80% were randomly selected for the training set and 20% were left to evaluate the quality of the prediction. The parameters of the neural network are shown in Table 1.

Table 1
The parameters of neural network

Dependencies	Name of network	Algorithm of learning	Error function	Function of hidden activation	Function of output activation
NO_2 -August 13, 2024	MLP 3-23-1	BFGS	SOS	Tangential	Exponential
NO_2 -August 6-13, 2024	MLP 3-29-1	BFGS	SOS	Tangential	Exponential

The prediction error was calculated using the Mean Absolute Percent Error (MAPE) formula:

$$MAPE = 100\,\% \cdot \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{true} - y_{prediction}\right|}{\left|y_{true}\right|}, \quad (1)$$

It was found that the predicting results are in good agreement with the experimental ones. The error of the NN method is 1.4%.

Conclusions

In this study, a machine learning model was developed to predict NO_2 concentrations based on meteorological and temporal data in Ternopil, Ukraine.

The concentration of nitrogen dioxide (NO_2) was predicted using experimental data obtained from the Weather Station during 6-13 August 2024. It was found that regardless of the type of study (self-selected test sample or randomly selected by a computer), the forecasting results are in good agreement with the experimental data. The error of the NM method is 3.9% and 1.4%, respectively, in the test samples.

The proposed model allows for real-time forecasting of pollutant emissions, which is an important tool for monitoring air quality in urban areas with limited resources.

The use of such models can be an important step toward creating early warning systems for elevated levels of pollution and rapid response to short-term fluctuations in the concentration of harmful substances. This could have a positive impact on public health, especially in areas with

heavy traffic and high building density. Future research could focus on integrating additional factors, such as traffic and other pollutants, to enable even more accurate predictions of air quality changes in modern urban ecosystems.

References

[1] González Ortiz, A., Guerreiro, C., & Soares, J. (2020). Air Quality in Europe: 2020 Report. In European Environment Agency. EU Publications: Luxembourg.

[2] Latza, U., Gerdes, S., & Baur, X. (2009). Effects of nitrogen dioxide on human health: Systematic review of experimental and epidemiological studies conducted between 2002 and 2006. In International Journal of HEH, 212, 271–287.

[3] Okudo, C.C., Ekere, N.R., & Okoye, C.O.B. (2022). Evaluation of Particulate Matter (PM2.5 and PM10) Concentrations in the Dry and Wet Seasons As Indices of Air Quality in Enugu Urban, Enugu State, Nigeria. In Journal of CSN, 47(5), 998-1015.

[4] Impacts of air pollution and acid rain on wildlife. In Air Pollution. http://www.air-quality.org.uk

[5] U.S. Environmental Protection Agency (USEPA).. Particulate Matter (PM) Basics. In EPA. http://www.epa.gov.gov/pm-pollution/particulate-matter-pm-basics.

[6] Stanko, A., Wieczorek, W., Mykytyshyn, A., Holotenko, O., & Lechachenko, T. (2024). Real-time air quality management: Integrating IoT and Fog computing for effective urban monitoring. CITI'2024: 2nd International Workshop on Computer Information Technologies in Industry 4.0, June 12–14, 2024, Ternopil, Ukraine.

[7] Duda, O., Mykytyshyn, A., Mytnyk, M., & Stanko, A. (2020). The network platform cyber-physical systems application for smart buildings air pollution indicators monitoring," Časopis Manažérska Informatika, Univerzita Komenského v Bratislave, Slovakia, vol. 1, no. 1, 2023, ISSN 2729-8310.

[8] Environmental Protection Agency. (2023). Nitrogen dioxide (NO2) pollution: Basic information about NO2. In EPA. www.epa.gov

[9] Baklanov, A., Molina, L.T., & Gauss, M. (2016). Megacities, air quality and climate. In Atmospheric Environment, 126, 235–249.

[10] Canepa, E., & Builtjes, P.J.H. (2017). Thoughts on Earth System Modeling: From global to regional scale. In Earth-Science Reviews, 171, 456–462.

[11] Arhami, M., Kamali, N., & Rajabi, M.M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. In Environmental Science and Pollution Research, 20, 4777–4789.

[12] Cabaneros, S.M., Calautit, J.K., & Hughes, B.R. (2019). A review of artificial neural network models for ambient air pollution prediction. In Env. Modelling & Software, 119, 285–304.

[13] Wu, Y., & Zhang, Y. (2020). Artificial neural network approaches for modeling air pollutants concentrations: A case study in Jinan, China. In Atm. Env., 224, 117333.

[14] Gardner, M., & Dorling, S. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. In Atm. Env., 33, 709–719.

[15] Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. In Atm. Env., 35, 815–825.

[16] Perez, P., & Trier, A. (2001). Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile. In Atmospheric Environment, 35, 1783–1789.

[17] Stamenković, L.J., Antanasijević, D.Z., Ristić, M., Perić-Grujić, A.A., & Pocajt, V.V. (2017). Prediction of nitrogen oxides emissions at the national level based on optimized artificial neural network model. In Air Quality, Atmosphere & Health, 10, 15–23.
[18] Jiang, P., Li, C., Li, R., & Yang, H. (2019). An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation. In Knowledge-Based Systems, 164, 174–192.
[19] Ding, W., Zhang, J., & Leung, Y. (2016). Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. In Environmental Science and Pollution Research, 23, 19481–19494.
[20] Liu, H., Wu, H., Lv, X., Ren, Z., Liu, M., Li, Y., & Shi, H. (2019). An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China. In Sustainable Cities and Society, 47, 101471.
[21] González-Enrique, J., Ruiz-Aguilar, J.J., Moscoso-López, J.A., Urda, D., Deka, L., & Turias, I.J. (2021). Artificial neural networks, sequence-to-sequence LSTMs, and exogenous variables as analytical tools for NO2 (air pollution) forecasting: A case study in the bay of algeciras (Spain). In Sensors, 21, 1770.
[22] Dai, H., Huang, G., Wang, J., Zeng, H., & Zhou, F. (2021). Prediction of Air Pollutant Concentration Based on One-Dimensional Multi-Scale CNN-LSTM Considering Spatial-Temporal Characteristics: A Case Study of Xi'an, China. In Atmosphere, 12, 1626.
[23] Yeganeh, B., Hewson, M.G., Clifford, S., Tavassoli, A., Knibbs, L.D., & Morawska, L. (2018). Estimating the spatiotemporal variation of NO2 concentration using an adaptive neuro-fuzzy inference system. In Env. Modelling & Software, 100, 222–235.
[24] Makhortykh, M., & Shevchuk, V. (2013). Ternopil Region: Geographical Features and Climate Overview. In Ukrainian Geographical Journal, 4, 55-67.
[25] EcoCity. EcoCity https://eco-city.org.ua
[26] NGO "Arnika". Arnika https://arnika.org
[27] Clean air for Ukraine: Clean Air for Ukraine https://cleanair.org.ua