DOI: https://doi.org/10.32782/ecovis/2026-3-7
UDC 004.85:338.2:504.3:656.1

Vasyl Hryhorkiv
Doctor of Physical and Mathematical Sciences, Professor,
Head of the Department of Economic and Mathematical Modelling,
Yuriy Fedkovych Chernivtsi National University
ORCID: https://orcid.org/0000-0003-4866-946X

Olena Yaroshenko
Candidate of Economic Sciences, Docent,
Associate Professor at the Department of Economic and Mathematical Modelling,
Yuriy Fedkovych Chernivtsi National University
ORCID: https://orcid.org/0000-0002-5976-6559

Bohdan Semenovych
Postgraduate Student at the Department of Economic Theory,
Management and Administration,
Yuriy Fedkovych Chernivtsi National University
ORCID: https://orcid.org/0009-0000-8779-5966

Mariia Hryhorkiv
Doctor of Economic Sciences, Professor,
Professor at the Department of Economic and Mathematical Modelling,
Yuriy Fedkovych Chernivtsi National University
ORCID: https://orcid.org/0000-0003-3327-991X

MODELLING AIR POLLUTION AS AN INDICATOR OF THE IMPACT OF THE REGION'S ECOLOGICAL-ECONOMIC SYSTEM ON SOCIAL WELFARE

Air pollution poses a critical threat to public welfare due to its impact on public health and labour productivity. This paper refines and empirically validates a methodology for short-term forecasting of PM2.5 and PM10 concentrations in the air by analysing transport and environmental data using the ensemble method of extremely randomised trees. The authors previously tested the basic methodology on data from New York City. The scientific novelty lies in the extension of the methodology through the introduction of a three-component structure of time lags, the development of economic activity variables based on the interaction between transport demand and the daily rhythms of urban activity, and the confirmation of the generalisability of the approach using data from two geographically distant cities. The results form the basis for the development of cost-effective environmental monitoring systems and evidence-based decision-making in the fields of environmental policy, urban mobility management and the assessment of environmental and economic benefits in the context of public welfare.

Keywords: machine learning, ensemble methods, ExtraTrees, particulate matter, air pollution forecasting, traffic flow intensity, environmental monitoring, environmental and economic benefits, social welfare.

Григорків Василь Степанович
доктор фізико-математичних наук, професор,
завідувач кафедри економіко-математичного моделювання,
Чернівецький національний університет імені Юрія Федьковича

Ярошенко Олена Іванівна
кандидат економічних наук, доцент,
доцент кафедри економіко-математичного моделювання,
Чернівецький національний університет імені Юрія Федьковича

Семенович Богдан Степанович
аспірант,
Чернівецький національний університет імені Юрія Федьковича

Григорків Марія Василівна
доктор економічних наук, професор,
професор кафедри економіко-математичного моделювання,
Чернівецький національний університет імені Юрія Федьковича



© Vasyl Hryhorkiv, Olena Yaroshenko, Bohdan Semenovych, Mariia Hryhorkiv, 2026
Стаття поширюється на умовах ліцензії відкритого доступу (CC BY 4.0)

МОДЕЛЮВАННЯ РІВНЯ ЗАБРУДНЕННЯ ПОВІТРЯ ЯК ІНДИКАТОРА ВПЛИВУ ЕКОЛОГО-ЕКОНОМІЧНОЇ СИСТЕМИ РЕГІОНУ НА СУСПІЛЬНИЙ ДОБРОБУТ

Забруднення повітря у великих містах становить серйозну загрозу для соціального добробуту через свій вплив на здоров'я населення та продуктивність праці, тоді як створення щільної мережі контрольних вимірювальних станцій залишається фінансово недоступним для більшості муніципальних бюджетів. Мета дослідження полягає в розширенні та емпіричній перевірці методології короткострокового прогнозування концентрацій частинок у атмосфері (PM2,5 та PM10) на основі аналізу даних про транспорт та навколишнє середовище з використанням методів машинного навчання, зокрема ансамблевого методу «Екстремально рандомізованих дерев» (Extremely Randomised Trees). Ця методологія була попередньо розроблена авторами на основі даних про місто Нью-Йорк. Внесок авторів полягає в уточненні структури змінних із часовим лагом, формуванні показників економічної активності та перевірці цього підходу в іншому географічному контексті. Наукова новизна даної роботи полягає у запровадженні трикомпонентної структури часових лагів, що охоплює інтервали в одну, дві та три години; у розробці показників економічної активності на основі взаємодії інтенсивності дорожнього руху з добовим профілем міської активності; а також у підтвердженні можливості застосування цього підходу до двох географічно віддалених міст. Емпіричне дослідження було проведено на основі даних про Київ, що містили 126 000 погодинних спостережень з десяти пунктів моніторингу за 2023–2024 роки. На незалежній тестовій вибірці модель дала пояснення 87,1 % дисперсії дрібної фракції та 85,0 % дисперсії грубої фракції, при цьому середні абсолютні похибки становили відповідно 1,55 та 2,73 мікрограма на кубічний метр. Порівняння з результатами, отриманими для Нью-Йорка, показує, що моделюючий ланцюжок зберігає високу точність, незважаючи на відмінності в кліматичних умовах, транспортній інфраструктурі та структурі економічної діяльності цих двох міст. Практичне значення отриманих результатів полягає в можливості збереження менш щільної мережі вимірювальних станцій, своєчасного попередження вразливих груп населення та попередньої оцінки транспортної та екологічної політики шляхом моделювання сценаріїв, що в сукупності створює основу для оцінки еколого-економічної корисності регіону в контексті соціального добробуту.

Ключові слова: машинне навчання, ансамблеві методи, ExtraTrees, тверді частинки, прогнозування забруднення повітря, інтенсивність транспортного потоку, екологічний моніторинг, еколого-економічна корисність, суспільний добробут.

Problem statement. The quality of ambient air is considered to be one of the most pressing environmental challenges of the twenty-first century, with consequences for public health and quality of life on a global scale. As stated in the Health Effects Institute's (2024) analytical report, air pollution is considered to be the second most significant risk factor for mortality, contributing to more than 8 million deaths per year [5]. The global scale of the problem is confirmed by data from the World Health Organisation, which indicates that almost 99% of the world's population breathes air that exceeds the WHO air quality guideline limits, posing considerable risks to public health and economic development [13]. The issue is particularly pronounced in urban areas, where high population density and extensive industrial and economic activity result in a significant deterioration in ambient air quality.

The transport sector plays a decisive role within the structure of urban pollution. It produces a wide range of pollutants, including fine particulate matter measuring less than 2.5 and 10 micrometres in diameter (PM2.5 and PM10, respectively), that directly contribute to respiratory and cardiovascular diseases [5]. According to the European Environment Agency, the transport sector accounted for 30.9% of PM10 and PM2.5 emissions across the European Union's 27 member states in 2023 [3]. Fig. 1 illustrates the dynamics of their atmospheric release by vehicle category.

Similar challenges with air pollution from the transport sector are faced by Ukrainian cities. Studies of the ecological situation in Kyiv have shown that road transport is one of the main sources of atmospheric pollution in the capital, alongside energy generation and industrial enterprises [9]. Due to limited municipal budgets, establishing a dense network of conventional monitoring stations is not financially viable, making cost-effective alternative air quality monitoring and forecasting systems essential.

Analysis of recent research and publications. Contemporary studies in air quality forecasting may be classified by the technological approaches they employ.

Traditional approaches to the monitoring of atmospheric pollution rely on networks of stationary measuring stations equipped with high-precision instruments. Kumar et al. [8] posit that, while such systems offer high measurement accuracy, they are characterised by limited spatial coverage and high operating costs, which motivates the use of low-cost sensing technologies. An additional alternative is represented by the satellite monitoring methods developed by van Donkelaar et al. [12]. It is evident that such methods provide global coverage; however, they are characterised by low temporal resolution and limited accuracy in urban areas due to cloud cover.

While traditional monitoring methods were being developed, advances in computer vision opened up new possibilities for analysing traffic flows. Redmon et al. [10] developed the YOLO (You Only Look Once) architecture, which enables real-time object detection. Subsequent refinements of this architecture, including

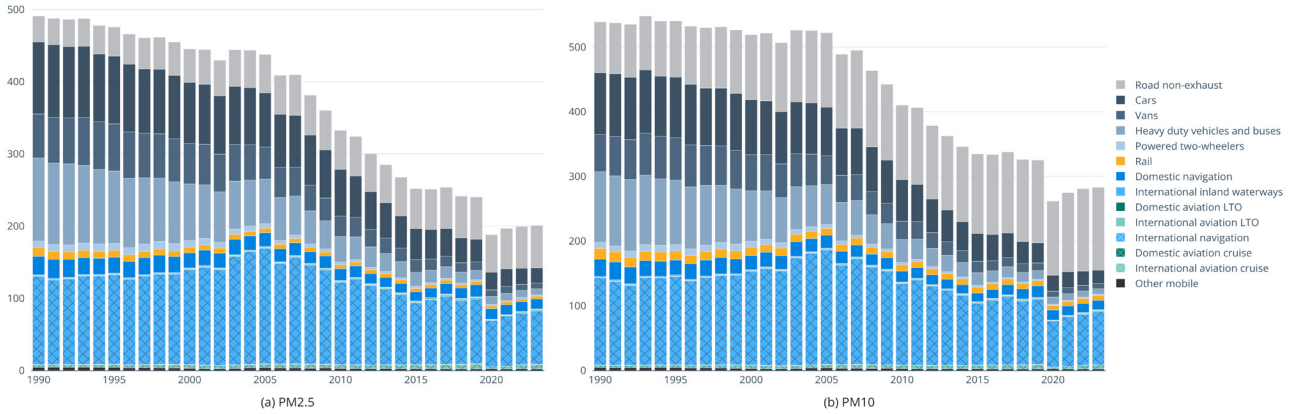


Figure 1. Emissions of PM2.5 (a) and PM10 (b) from the transport sector in the EU-27 by transport mode, 1990-2023, Gg per year

Source: adapted by the authors from [3]

YOLO11, have achieved vehicle detection accuracy of over 95%, while maintaining high processing speeds. However, most studies in this area concentrate on the technical aspects of detection without integrating them with environmental modelling.

In the field of algorithmic support for forecasting, ensemble machine learning methods have made a significant contribution. Breiman [2] proposed the Random Forests method, which has become fundamental to many regression analysis applications. Geurts et al. [4] then built on this by developing the Extremely Randomised Trees (ExtraTrees) algorithm. This differs from Random Forests in the way that splitting thresholds are selected, and it exhibits reduced prediction variance.

Zheng et al. [15] further advanced the integration of machine learning methods with environmental monitoring by proposing the U-Air system, which uses big data to infer urban air quality. Meanwhile, Karagulian et al. [6] conducted a systematic review of the contributions of local sources to ambient particulate matter concentrations in cities worldwide, confirming the importance of the transport sector in this regard. In a recent study, Kostadinov et al. [7] applied deep learning methods to air pollution forecasting, focusing on the impact of urban traffic on PM10 concentrations. Notably, the study found that urban car traffic was not the dominant contributor in the examined city. Previous work by the same research group [1] evaluated recurrent architectures for PM10 forecasting at 6-, 12-, and 24-hour time intervals, and compared their performance with that of an ARIMA model. Both studies relied on meteorological and historical concentration data rather than direct measurements of traffic volumes.

A previous study by the present research team [14] demonstrated the feasibility of using automated traffic flow analysis to forecast PM2.5 and PM10 concentrations by integrating computer vision technologies with ensemble methods. The system uses the YOLO11s model to recognise and categorise six types of vehicle – private cars, taxis, commercial vehicles, medium trucks, heavy trucks and buses – in complex urban traffic conditions. The ExtraTrees algorithm uses multivariate regression models to provide high-accuracy forecasts of PM2.5 and PM10 concentrations. Empirical validation of more than 50,000 observations in New York City confirmed the stability and suitability of the methodology for automated environmental monitoring systems.

Nevertheless, two key issues remain unresolved: the geographical transferability of the methodology to regions with different climatic, infrastructural and economic characteristics; and the optimal configuration of lagged variables for different types of pollutants, taking into account their physicochemical properties and how they behave in the atmosphere. These gaps define the aim and objectives of the present study.

The purpose of the article. This study aims to extend and empirically validate a short-term forecasting methodology for atmospheric concentrations of PM2.5 and PM10 particulate matter. This methodology is based on machine learning and the analysis of transport and environmental data [14]. It involves refining the structure of temporally lagged variables and developing features of economic activity based on the interaction between traffic intensity and the city's diurnal patterns. The approach's geographical generalisability is confirmed through comparative validation using data from two cities: Kyiv and New York.

Presentation of the main research material. The research methodology comprises six sequential stages: data preparation and preprocessing, construction of an extended system of lagged variables, engineering of economic activity features, training of a multi-output ExtraTrees ensemble model, assessment of forecasting accuracy, and comparative validation of the results.

Data sources and sample structure. This empirical study integrates environmental and transport data from Kyiv, covering the period from 2023 to 2024. Hourly measurements of atmospheric pollutant concentrations

were obtained from the open SaveEcoBot platform [11], which aggregates data from a network of automated air quality monitoring stations equipped with high-precision sensors that continuously measure fine particulate matter concentrations and meteorological parameters. To maximise the relevance of the environmental data to traffic intensity, ten monitoring stations were selected that are located in the immediate vicinity of the city's main traffic routes and junctions. For each location, mass concentrations of PM2.5 (particles with a diameter of less than 2.5 μm) and PM10 (less than 10 μm) expressed in micrograms per cubic metre (μg/m³), together with air temperature (°C), relative humidity (%), and a time stamp accurate to the hour, were recorded at hourly resolution. Table 1 shows the structure of the integrated dataset. This includes hourly observations of environmental and transport parameters at monitoring location 14 in Kyiv.

Information on traffic intensity was obtained from Kyiv City Council's traffic monitoring system. As complete historical data is not publicly available, a traffic intensity model was developed based on detailed data from a representative period, with extrapolation over a two-year period. This model accounts for diurnal patterns, including morning and evening peak hours, as well as weekly and annual seasonality, the latter reflecting vacation periods and public holidays.

The final dataset comprises 126,000 hourly observations, which is sufficient to reliably train and validate machine learning models, as well as detect dependencies between predictors and target variables. Data preprocessing was performed on this basis to ensure the correctness of the modelling procedure.

Data preprocessing. To limit the influence of outliers, extreme values were filtered using the thresholds $PM2.5 \le 50$ μg/m³ and $PM10 \le 100$ μg/m³, which correspond to the upper bounds of the concentration range observed in urbanised areas in the absence of extreme events such as forest fires or dust storms. To construct lagged variables correctly, the dataset must be sorted by location identifier and timestamp. This ensures that measurements are ordered chronologically for each location, prevents future values from being used as predictors and preserves the model's causal structure.

To account for the autocorrelation structure of the concentrations, a three-component lag system with lags of 1, 2 and 3 hours was applied to both pollutants. Grouping by location identifier when constructing the lagged variables ensures that lags are computed separately for each observation point. This prevents data from different locations being mixed together, which would violate the temporal structure. Taking into account the time delay between vehicle emissions and measurement by air quality sensors, arising from dispersion and atmospheric advection, the traffic intensity variable was introduced with a three-hour lag.

In order to capture the city's diurnal patterns and the differences in vehicle operating regimes, a system of temporal characteristics and economic activity indicators was developed. These features include the time of day and day of the week, which model diurnal and weekly traffic flow cycles. Two binary indicators were introduced: the first distinguishes between working days and weekends, and the second identifies business hours – that is, the period of peak economic activity characterised by intensive movement of commercial and private transport (08:00–19:00). Additionally, an interaction term was constructed between traffic intensity and the business hours indicator, enabling the model to capture different regimes of pollutant emission depending on the time of day. The physical rationale for this approach is that, during periods of high economic activity, vehicles more frequently operate in idling mode in congested areas with frequent stops. This is accompanied by higher emission factors than during periods of low activity when vehicles can move freely.

In the final stage, observations with missing values were removed to ensure the training data was complete. A forecasting model with the specifications outlined below was then constructed.

Model specification. The final forecasting model uses 15 numerical features, which cover five categories of predictors. The transport component is represented by current traffic intensity, its lagged value shifted by three hours, and the interaction between traffic intensity and the temporal context of economic activity. Meteorological conditions are accounted for by air temperature and relative humidity. The temporal structure is modelled using the hour of the day and the day of the week, as well as binary indicators for weekends and periods of peak business activity. The autocorrelation dynamics of pollutants are captured using six lagged PM2.5 and PM10 concentration variables at 1-, 2-, and 3-hour

Table 1

Structure of air quality and traffic intensity monitoring data in Kyiv

Location ID	Traffic intensity, veh/h	PM2.5, μg/m³	PM10, μg/m³	Temperature, °C	Humidity, %	Date and time
14	319	10.77	19.94	7.70	93.50	2023-01-01 00:00
14	150	13.79	24.56	9.35	93.50	2023-01-01 01:00
14	97	12.08	22.47	8.23	93.40	2023-01-01 02:00
14	108	10.04	19.30	8.90	93.40	2023-01-01 03:00
14	148	11.31	22.17	8.40	93.20	2023-01-01 04:00

Source: compiled by the authors based on [11]

horizons. To account for spatial heterogeneity among the ten observation points, reflecting differences in local transport infrastructure, topography, and microclimatic conditions, a categorical feature denoting the monitoring location identifier is also included.

The following preprocessing tools from the scikit-learn library were used for the differentiated processing of numerical and categorical features:

– ColumnTransformer, which ensures the differentiated processing of numerical and categorical features by applying specific transformations to each type of variable.

– StandardScaler standardises numerical variables to zero mean and unit variance, which is important for the stability of machine learning algorithms and for ensuring a uniform scale across all features.

– OneHotEncoder, which converts the categorical location identifier variable into binary indicators and ensures proper handling of potential new categories in the test data.

The core model is an ExtraTreesRegressor comprising an ensemble of 600 decision trees. The algorithm uses strong randomisation: at each node, it randomly selects a subset of features and generates split thresholds, then adopts the best candidate split. Each tree is fitted to the entire training sample, without bootstrap resampling. This randomisation increases the diversity of the base estimators and reduces the variance of the ensemble forecast.

To forecast the two target variables (PM2.5 and PM10) simultaneously, the scikit-learn meta-estimator MultiOutputRegressor was used to adapt the model to multiple output variables. This approach involves training separate ExtraTrees models for each pollutant, while maintaining a single input feature space. This ensures that the parameters for PM2.5 and PM10 can be optimised independently. Advantages of this approach include a unified data preprocessing pipeline for both target variables, simplified system architecture and the ability to generate forecasts for both pollutants from a single model call.

The entire analysis process, from preliminary data preparation to generating forecasts, is integrated into a single automated system. This system's technical architecture ensures consistent and coherent information processing throughout the modelling process.

Structure of the monitoring system and system requirements. The developed monitoring system consists of two main components: a video analytics module and an air quality forecasting module. The video analytics module is based on the YOLO11s architecture and performs real-time detection and classification of vehicles (see Figures 2 and 3).

Figures 2 and 3 demonstrate the technical feasibility of automatically collecting detailed traffic data, eliminating the need for manual counting or sensors embedded in the road surface. The system can distinguish between different vehicle categories, which is important for assessing the contribution of road traffic to air pollution since emission factors can differ substantially between vehicle types.

This module has the following minimum hardware requirements: a four-core processor; 8 GB of random-access memory; and 6 GB of disk space to store the model, dependencies and logs. The system can recognise objects at a rate of 5 frames per second when processing an input video stream with a resolution of up to 720 p.

The air quality forecasting module uses a trained ExtraTrees model to produce hourly PM2.5 and PM10 concentration forecasts. This module has considerably lower hardware requirements, needing only a two-core processor, 4 GB of random-access memory and 1 GB of disk space for the stored model and software dependencies. Forecasting takes

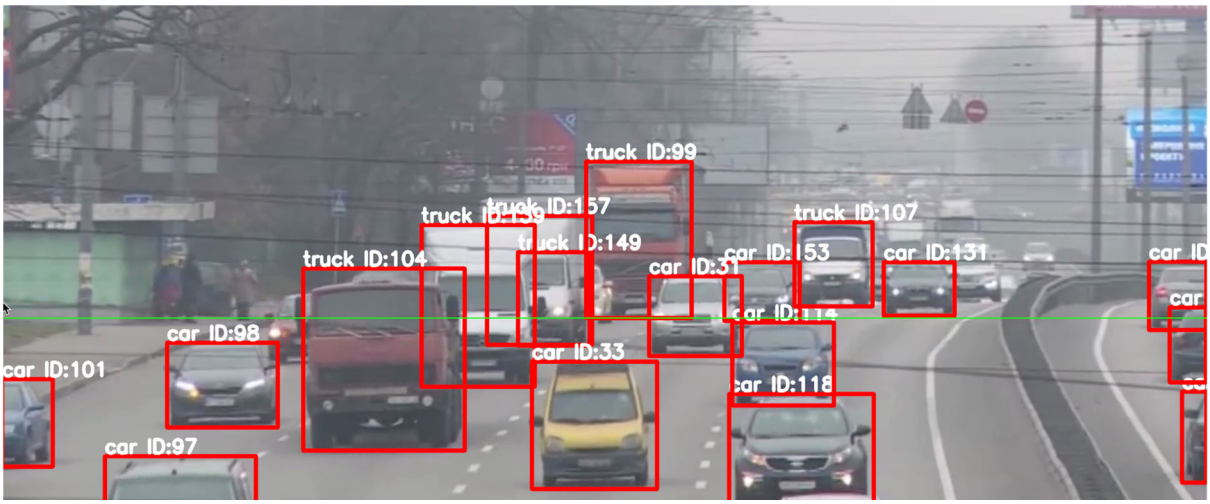


Figure 2. Example of the automatic classification of vehicles by computer vision system under urban traffic conditions for the collection of traffic intensity data

Source: compiled by the authors

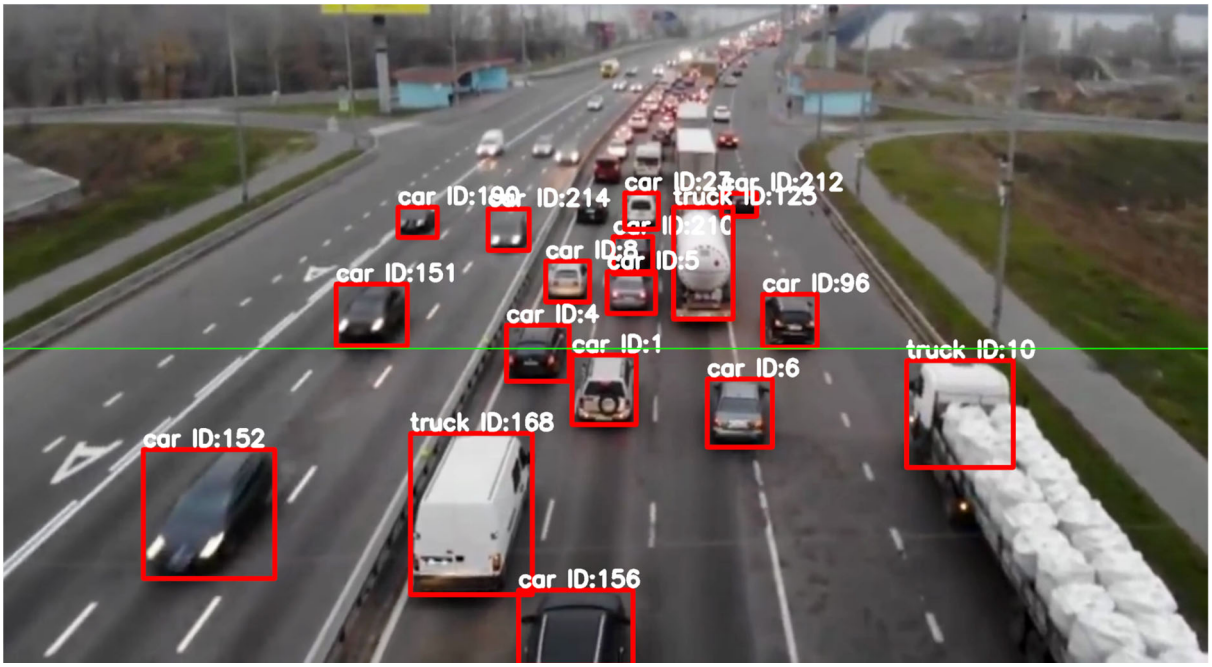


Figure 3. Detection and identification of different categories of vehicles in a mixed traffic flow in the evening for the monitoring of traffic intensity

Source: compiled by the authors

20–30 milliseconds per prediction, enabling real-time processing of data from multiple monitoring locations. The forecasting module does not require a graphics processing unit, meaning the system can be deployed on standard server hardware without the need for additional investment in specialised computing resources.

The pipeline applies the same processing procedures to all data and prevents information from the hold-out sample from being used in model construction. This ensures an objective assessment of forecasting accuracy. This accuracy was evaluated using an independent test sample.

Model validation. To validate the model, the dataset was divided into training and test samples, comprising 100,800 and 25,200 observations, respectively. The 80:20 ratio is standard practice in machine learning. This provides a balance between having sufficient data for training and being able to make a reliable statistical assessment of the quality of forecasts on independent data.

Two principal metrics were used to evaluate forecasting accuracy. The mean absolute error (MAE) is computed as the mean of the absolute deviations of the forecasts $\hat{y}_i$ from the actual values y_i. It is measured in the same units as the target variable (μg/m³):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \qquad (1)$$

where n denotes the number of observations in the test sample.

The coefficient of determination R² is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \qquad (2)$$

where $\bar{y}$ is the arithmetic mean of the actual values. The coefficient represents the proportion of variance in the target variable explained by the model. A value of 1 indicates a perfect forecast, while a value of 0 indicates a forecast based solely on the mean.

Empirical forecasting results. The training of the model on the training sample and the validation on the independent test sample demonstrated high forecasting accuracy for both pollutants. For PM2.5, a coefficient of determination of $R^2 = 0.871$ was attained with a mean absolute error of $MAE = 1.55$ μg/m³. For PM10, $R^2 = 0.850$ was obtained with $MAE = 2.73$ μg/m³. These results indicate that the model explains 87.1% of the variability in PM2.5 concentrations and 85.0% of the variability in PM10, which suggests the adequacy of the selected feature set.

A graphical analysis of the relationship between actual and predicted values is presented in Figures 4 and 5. The scatter plots demonstrate a substantial concentration of data points in close proximity to the diagonal of the perfect forecast, visually corroborating the high correlation between predicted and observed concentrations. The model reproduces PM2.5 somewhat more closely $(R^2 = 0.871)$ than PM10 $(R^2 = 0.850)$, which is consistent with the more heterogeneous sources of the coarse fraction. The residual dispersion visible

for both pollutants is indicative of factors that are not represented by the feature set, including secondary formation in the atmosphere via chemical reactions, long-range transport from regional sources, and complex interactions with meteorological parameters.

As illustrated in Figure 5, a comparable trend is evident for PM10, with the distribution of data points expanding in the upper segment of the concentration range. This phenomenon may be associated with rare episodes of elevated pollution caused by atypical meteorological conditions or by temporary emission sources that are not represented in the model. The visual analysis indicates a high degree of agreement between predicted and observed values across most of the observed range, with a tendency towards underestimation at high concentrations.

To demonstrate the model's capacity to replicate the temporal evolution of pollutant concentrations, Figures 6 and 7 present a comparison of the actual and predicted PM2.5 and PM10 values at location 14.

Figures 6 and 7 show a high degree of agreement between the actual and predicted values, with the model successfully reproducing both short-term fluctuations and the general trend. The absence of systematic deviations confirms that the model's forecasting accuracy is stable over time.

Comparative analysis of the results. Comparing the results of the present study on Kyiv data with those of a previous study on New York data reveals important regularities and confirms that the methodology can be generalised. For PM2.5, forecasting accuracy improves from $R^2 = 0.776$ in New York to $R^2 = 0.871$ in Kyiv, a gain of 9.5 percentage points. This improvement can be attributed to several factors that reflect the methodological refinements of the present study.

The extended lag structure, which spans three temporal horizons (1, 2 and 3 hours), enables a more complex autocorrelation structure of PM2.5 concentrations to be modelled. In contrast to coarse particles, fine particles have a longer atmospheric lifetime, meaning that information on concentrations in the distant past is relevant for forecasting the current state. Including lags at 2 and 3 hours enables the model to account for the accumulation and gradual dispersion of pollutants, which occur at distinct characteristic timescales.

The engineering of economic activity features, particularly the development of the business-hours indicator and the interaction variable between traffic intensity and temporal context, was important for modelling the impact of transport. The operating regime of vehicles has a significant impact on emission intensity; idling in congestion during business hours, for example, generates higher emission factors than free-flow movement during periods of low activity. The interaction variable enables the model to

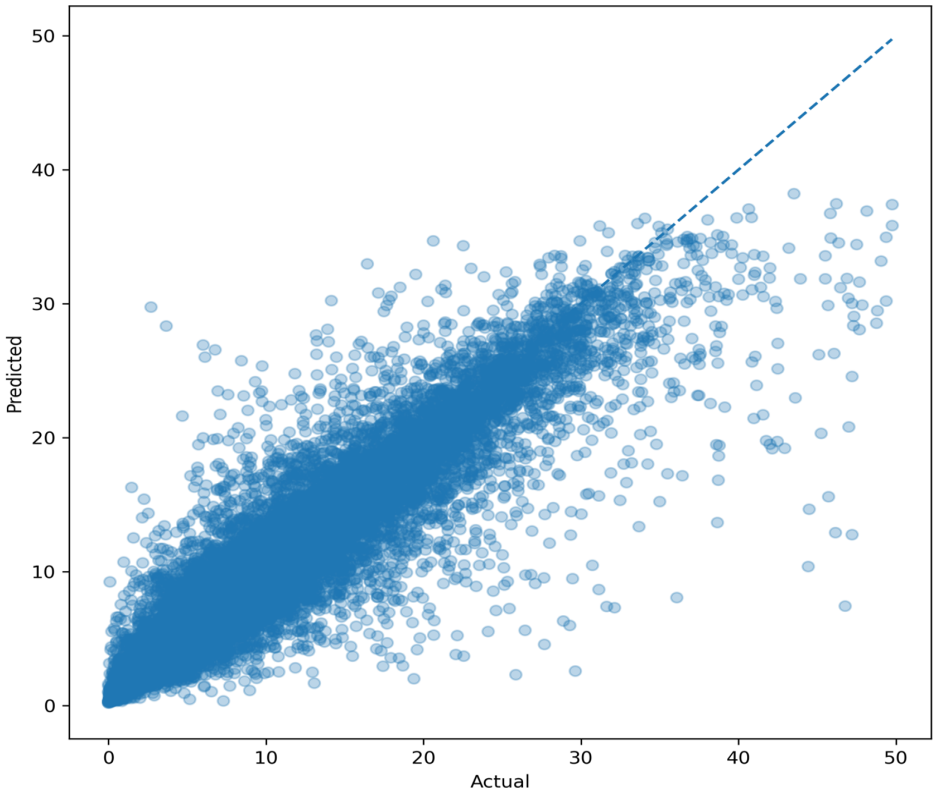


Figure 4. Comparison of actual and predicted PM2.5 concentration values using the three-component lag structure $R^2 = 0.871$

Source: compiled by the authors

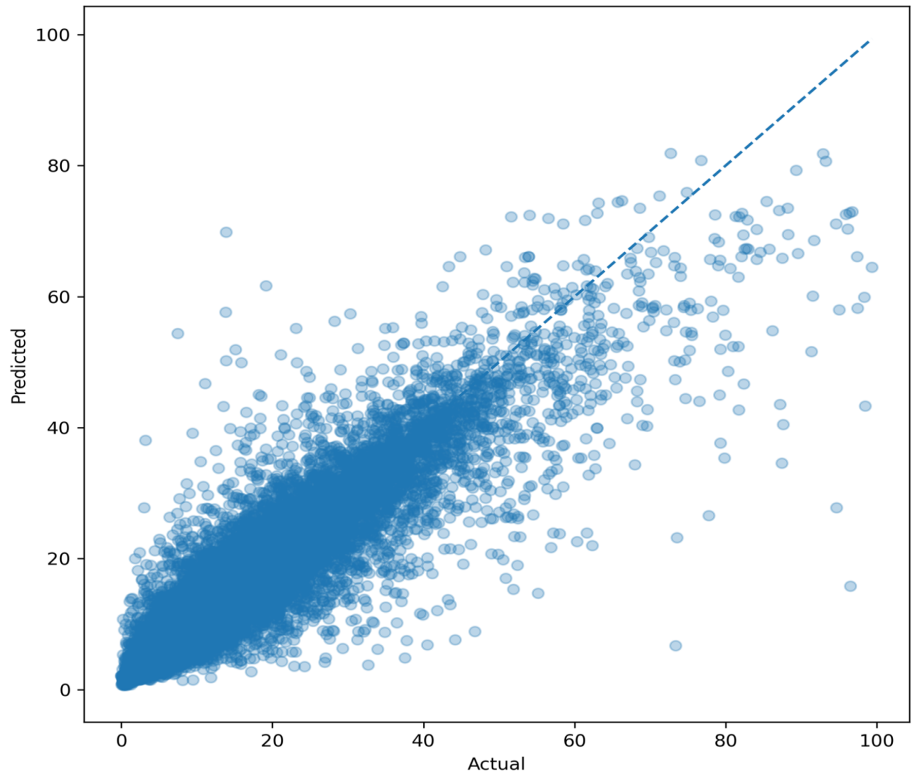


Figure 5. Comparison of actual and predicted PM10 concentration values using the three-component lag structure $R^2 = 0.850$

Source: compiled by the authors

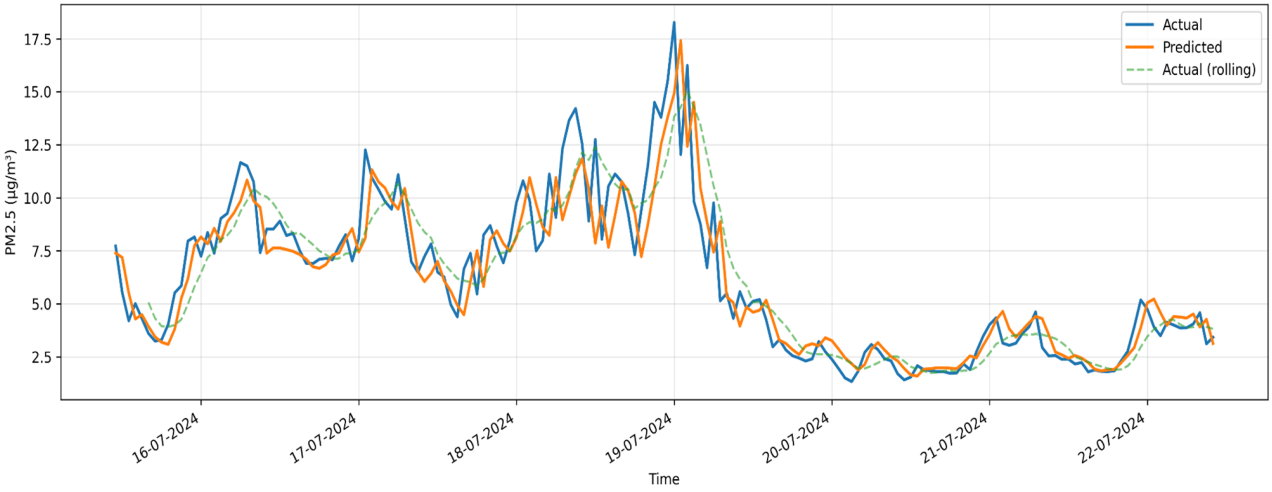


Figure 6. Temporal dynamics of actual (blue line) and predicted (orange line) PM2.5 concentration values for monitoring location 14 over 150 hours

Source: compiled by the authors

automatically detect and account for this effect, eliminating the need for the explicit modelling of complex nonlinear dependencies.

The larger sample size of 126,000 observations, compared to the previous study, ensures better representation across a range of meteorological conditions and traffic intensity levels. This is particularly important for training ensemble methods, since a greater quantity of data enables more diverse base estimators to be constructed and the variance of the final forecast to be reduced. Geographical specificity may also be a factor, as Kyiv's transport infrastructure is less complex than that of New York, which could simplify modelling the impact of transport on air pollution.

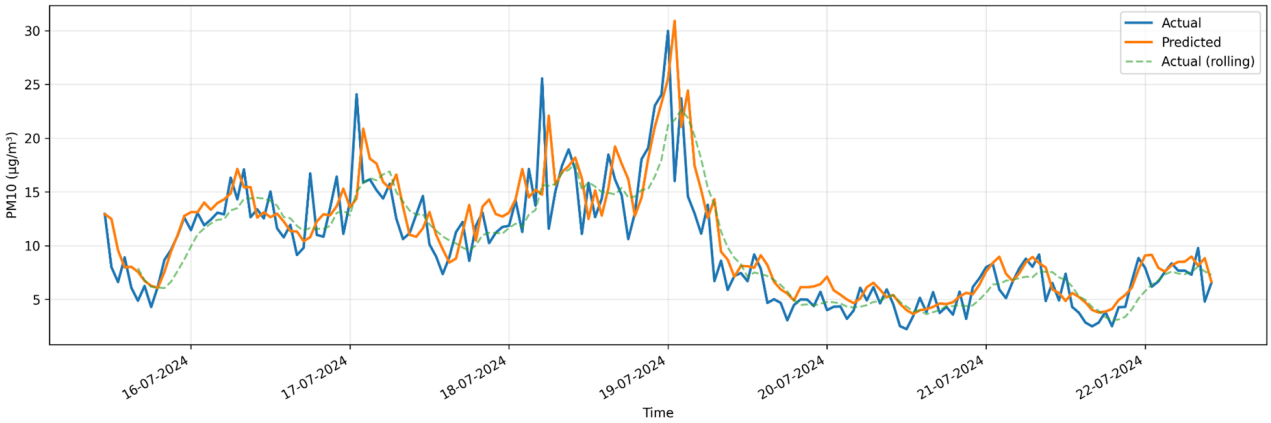


Figure 7. Temporal dynamics of actual (blue line) and predicted (orange line) PM10 concentration values for monitoring location 14 over 150 hours

Source: compiled by the authors

For PM10, the opposite trend is observed: accuracy declines from $R^2 = 0.972$ in New York to $R^2 = 0.850$ in Kyiv, a 12.2 percentage-point reduction. This can be explained by the fundamental differences in the sources and dynamics of the coarser fraction. PM10 includes larger particles that are generated not only by exhaust emissions, but also by non-exhaust emissions from tyre, brake and road surface abrasion, construction work and the resuspension of road dust. The spatial and temporal variability of these sources is difficult to predict solely from aggregated traffic intensity. This limitation is becoming increasingly relevant over time. While total particulate emissions from transport in the EU-27 fell by 53% between 1990 and 2023, the non-exhaust fraction increased by 63.1% during this period [3].

The paucity of meteorological data, particularly the absence of information on wind speed and direction, has a critical effect on PM10 forecasting. While coarser PM10 particles settle onto surfaces more rapidly under gravity than finer PM2.5 particles, wind flows can substantially affect their spatial distribution. Indeed, intense wind is capable of rapidly dispersing local concentrations or, conversely, of transporting particles from industrial zones or construction sites, thereby creating episodes of elevated pollution in the absence of immediate emission sources at a given location. The influence of wind conditions is even more critical for finer PM2.5 particles, which remain airborne for longer and are transported over greater distances. However, the inclusion of lagged concentration variables partially compensates for the model's lack of direct meteorological data.

Kyiv has a more continental climate with more pronounced seasonal variations than New York, which is influenced by the Atlantic Ocean. This results in greater variability in PM10 concentrations due to seasonal effects, such as increased emissions from residential heating in winter and dust resuspension during summer droughts. The model's inability to account for these seasonal sources constrains its PM10 forecasting capability.

Despite these differences, both studies demonstrate consistently high forecasting accuracy, with coefficients of determination exceeding 70% for both pollutants in both geographical contexts. This confirms the generalisability of the present approach across continents, combining the ExtraTrees ensemble method, lagged variables to model autocorrelation and feature engineering to model the influence of economic activity. The methodology can be adapted to different urbanised territories without the need for a fundamental redesign of the model architecture. This has important practical implications for scaling the system up in other cities.

The results obtained allow the following conclusions to be drawn from the study.

Conclusions. The study validated and expanded upon a methodology for forecasting ambient air pollution in Kyiv using ensemble machine learning methods. High forecasting accuracy was attained, with coefficients of determination of $R^2 = 0.871$ for PM2.5 with a mean absolute error of 1.55 µg/m³ and $R^2 = 0.850$ for PM10 with an error of 2.73 µg/m³ on a sample of 126,000 hourly observations, which demonstrates the practical applicability of the approach for automated environmental monitoring systems.

The comparative analysis of results from two geographically distinct cities confirms the cross-continental generalisability of the developed methodology. Despite substantial differences in climatic conditions, transport infrastructure, the structure of economic activity, and the regulatory environment between New York and Kyiv, the basic approach employing ExtraTrees, lagged variables, and feature engineering exhibits consistently high forecasting accuracy. This has important practical implications for the adaptation of the methodology to diverse urbanised contexts without the need for a fundamental redesign of the model architecture.

The results of the study are of practical significance for the development of environmental monitoring systems, the determination of the overall state of ecological-economic interaction, and the formulation of air quality management policy in urbanised territories. The principal benefit of the methodology is its capacity to precisely predict pollutant concentrations using readily available traffic intensity data, thus facilitating significant avenues for implementation. First and foremost, it enables cost-effective monitoring systems to be created without the need for a dense network of expensive air-quality sensors. In addition, short-term forecasts can provide vulnerable population groups with timely warnings about periods of elevated pollution. Finally, the model can be used to assess the potential impact of various transport, environmental and ecological-economic policies in advance by simulating scenarios with changing traffic intensity.

Further research should include: integrating detailed meteorological data on wind speed and direction to improve the accuracy of PM10 forecasting; differentiating vehicles by engine type and category for a more precise assessment of emissions; using spatio-temporal models to model dependencies between locations; including data on the chemical composition of particles to identify specific pollution sources; validating the methodology using data from other cities in different climatic zones and with different levels of economic development. This will confirm the generalisability of the approach on a global scale and its relevance for determining the environmental components of a region's ecological-economic system and its impact on social welfare.

Конфлікт інтересів: співавтори є членами редакційної колегії журналу. Для забезпечення незалежності редакційного процесу рукопис розглядався редактором, який не мав конфлікту інтересів зі співавторами. Співавтори не брали участі у відборі рецензентів, процесі рецензування та прийнятті рішення щодо публікації статті. Рукопис пройшов стандартну процедуру подвійного сліпого рецензування.

REFERENCES

1. Arsov M., Zdravevski E., Lameski P., Corizzo R., Koteli N., Gramatikov S., Mitreski K., Trajkovik V. (2021) Multi-horizon air pollution forecasting with deep neural networks. *Sensors*, vol. 21, no. 4, 1235. DOI: https://doi.org/10.3390/s21041235
2. Breiman L. (2001) Random forests. *Machine Learning*, vol. 45, no. 1, pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324
3. European Environment Agency (2026) Sustainability of Europe's mobility systems 2025. Air pollution section. Available at: https://www.eea.europa.eu/en/analysis/publications/sustainability-of-europes-mobility-systems-2025/air-pollution (accessed September 17, 2026)
4. Geurts P., Ernst D., Wehenkel L. (2006) Extremely randomized trees. *Machine Learning*, vol. 63, no. 1, pp. 3–42. DOI: https://doi.org/10.1007/s10994-006-6226-1
5. Health Effects Institute, Institute for Health Metrics and Evaluation, UNICEF (2024) State of Global Air 2024. Special Report. Boston, MA: Health Effects Institute. Available at: https://www.healthdata.org/sites/default/files/2024-06/soga-2024-report.pdf (accessed September 10, 2026)
6. Karagulian F., Belis C. A., Dora C. F. C., Prüss-Ustün A. M., Bonjour S., Adair-Rohani H., Amann M. (2015) Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmospheric Environment*, vol. 120, pp. 475–483. DOI: https://doi.org/10.1016/j.atmosenv.2015.08.087
7. Kostadinov M., Zdravevski E., Lameski P., Coelho P. J., Stojkoska B., Herzog M. A., Trajkovik V. (2024) Forecasting air pollution with deep learning with a focus on impact of urban traffic on PM10 and noise pollution. PLOS ONE, vol. 19, no. 12, e0313356. DOI: https://doi.org/10.1371/journal.pone.0313356
8. Kumar P., Morawska L., Martani C., Biskos G., Neophytou M., Di Sabatino S., Bell M., Norford L., Britter R. (2015) The rise of low-cost sensing for managing air pollution in cities. *Environment International*, vol. 75, pp. 199–205. DOI: https://doi.org/10.1016/j.envint.2014.11.019
9. Popov O. O., Iatsyshyn A. V., Kovach V. O., Artemchuk V. O., Kameneva I. P., Taraduda D. V., Sobyna V. O., Sokolov D. L., Dement M. O., Yatsyshyn T. M. (2020) Risk Assessment for the Population of Kyiv, Ukraine as a Result of Atmospheric Air Pollution. *Journal of Health and Pollution*, vol. 10, no. 25, 200303. DOI: https://doi.org/10.5696/2156-9614-10.25.200303
10. Redmon J., Divvala S., Girshick R., Farhadi A. (2016) You only look once: unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779–788. DOI: https://doi.org/10.1109/CVPR.2016.91
11. SaveEcoBot (2024) API dostupu do danykh pro yakist povitria v Ukraini [API for accessing air quality data in Ukraine]. Available at: https://www.saveecobot.com/static/api (in Ukrainian)
12. van Donkelaar A., Martin R. V., Brauer M., Hsu N. C., Kahn R. A., Levy R. C., Lyapustin A., Sayer A. M., Winker D. M. (2016) Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, vol. 50, no. 7, pp. 3762–3772. DOI: https://doi.org/10.1021/acs.est.5b05833
13. World Health Organization (2022) Billions of people still breathe unhealthy air: new WHO data. Available at: https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data (accessed September 10, 2026)
14. Yaroshenko O., Semenovych B., Hryhorkiv V., Hryhorkiv M. (2025) Zastosuvannia mashynnoho navchannia ta kompiuternoho zoru dlia prohnozuvannia pokaznykiv zabrudnennia povitria yak faktoru dobrobutu suspilstva [Application of machine learning and computer vision for forecasting air pollution indicators as a factor of social welfare]. *Naukovyi visnyk Chernivetskoho natsionalnoho universytetu imeni Yuriia Fedkovycha. Seriia Ekonomika – Scientific Journal of Yuriy Fedkovych Chernivtsi National University. Economics*, no. 1, pp. 70–77. DOI: https://doi.org/10.32782/ecovis/2025-1-11 (in Ukrainian)

15. Zheng Y., Liu F., Hsieh H. P. (2013) U-Air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1436–1444. DOI: https://doi.org/10.1145/2487575.2488188

СПИСОК ЛІТЕРАТУРИ

1. Arsov M., Zdravevski E., Lameski P. et al. Multi-horizon air pollution forecasting with deep neural networks. *Sensors*. 2021. Vol. 21, № 4. 1235. DOI: https://doi.org/10.3390/s21041235
2. Breiman L. Random forests. *Machine Learning*. 2001. Vol. 45, № 1. P. 5–32. DOI: https://doi.org/10.1023/A:1010933404324
3. European Environment Agency. Sustainability of Europe's mobility systems 2025 : web report. Розділ «Air pollution». 2026. URL: https://www.eea.europa.eu/en/analysis/publications/sustainability-of-europes-mobility-systems-2025/air-pollution (дата звернення: 17.09.2026).
4. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees. *Machine Learning.* 2006. Vol. 63, № 1. P. 3–42. DOI: https://doi.org/10.1007/s10994-006-6226-1
5. Health Effects Institute, Institute for Health Metrics and Evaluation, UNICEF. State of Global Air 2024. Special Report. Boston, MA : Health Effects Institute, 2024. URL: https://www.healthdata.org/sites/default/files/2024-06/soga-2024-report.pdf (дата звернення: 10.09.2026).
6. Karagulian F., Belis C. A., Dora C. F. C. et al. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmospheric Environment*. 2015. Vol. 120. P. 475–483. DOI: https://doi.org/10.1016/j.atmosenv.2015.08.087
7. Kostadinov M., Zdravevski E., Lameski P. et al. Forecasting air pollution with deep learning with a focus on impact of urban traffic on PM10 and noise pollution. *PLOS ONE*. 2024. Vol. 19, № 12. e0313356. DOI: https://doi.org/10.1371/journal.pone.0313356
8. Kumar P., Morawska L., Martani C. et al. The rise of low-cost sensing for managing air pollution in cities. *Environment International*. 2015. Vol. 75. P. 199–205. DOI: https://doi.org/10.1016/j.envint.2014.11.019
9. Popov O. O., Iatsyshyn A. V., Kovach V. O. et al. Risk Assessment for the Population of Kyiv, Ukraine as a Result of Atmospheric Air Pollution. *Journal of Health and Pollution.* 2020. Vol. 10, № 25. 200303. DOI: https://doi.org/10.5696/2156-9614-10.25.200303
10. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016. P. 779–788. DOI: https://doi.org/10.1109/CVPR.2016.91
11. SaveEcoBot. API для доступу до даних про якість повітря в Україні. 2024. URL: https://www.saveecobot.com/static/api
12. van Donkelaar A., Martin R. V., Brauer M. et al. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology.* 2016. Vol. 50, № 7. P. 3762–3772. DOI: https://doi.org/10.1021/acs.est.5b05833
13. World Health Organization. Billions of people still breathe unhealthy air: new WHO data. 2022. URL: https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data (дата звернення: 10.09.2026).
14. Ярошенко О., Семенович Б., Григорків В., Григорків М. Застосування машинного навчання та комп'ютерного зору для прогнозування показників забруднення повітря як фактору добробуту суспільства. *Науковий вісник Чернівецького національного університету імені Юрія Федьковича. Серія Економіка*. 2025. № 1. С. 70–77. DOI: https://doi.org/10.32782/ecovis/2025-1-11
15. Zheng Y., Liu F., Hsieh H. P. U-Air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013. P. 1436–1444. DOI: https://doi.org/10.1145/2487575.2488188

Дата надходження статті: 19.08.2026
Дата прийняття статті до публікації: 07.10.2026
Дата публікації статті: 27.10.2026