УДК 504.5:004.942 **doi: 10.20998/2079-0821.2025.02.13**

Є. О. ЛЕБЬОДКІН, Є. М. ВАРЛАМОВ, О. М. СКАКАЛЬСЬКИЙ, О. А. ПАЛАГУТА, Н. С. ЦАПКО

ПОРІВНЯННЯ МЕТОДІВ ПРОГНОЗУВАННЯ КОНЦЕНТРАЦІЙ PM_{10} В КРИВОМУ РОЗІ В ЗИМОВИЙ ПЕРІОД

У статті порівнюються два підходи для прогнозування концентрації дрібнодисперсних частинок PM_{10} - класичне статистичне моделювання (множинна лінійна регресі, МЛР) та сучасний алгоритм машинного навчання Random Forest (RF). Об'єктом дослідження обрано одне з найбільших промислових міст України - Кривий Ріг, яке відоме своєю складною екологічною ситуацією. Розглянуто зимовий період 2024-2025 рр., протягом якого виконано безперервний моніторинг PM_{10} та метеорологічних показників за допомогою автоматизованої міні-станції Cairnet із сертифікованими датчиками. Проведено попередню обробку даних (очищення від аномалій, заповнення пропусків, нормалізація) та формування ознак, зокрема введено категоріальні змінні для частини доби, типу дня (вихідний/робочий) та інтенсивності викидів. Обидві моделі показали схожі добові та тижневі цикли концентрації PM_{10}: пікові значення спостерігаються у вечірній і нічний час, найнижчі - вдень, що узгоджується з очікуваннями та літературними даними. Водночас точність прогнозу суттєво різниться: Random Forest забезпечив значно вищу детермінацію ($R^2 \approx 0{,}72$ проти $R^2 \approx 0{,}27$ у регресії) і вдвічі меншу середню абсолютну похибку. Наведено порівняння важливості факторів для обох моделей: Random Forest виділив атмосферний тиск, температуру та вологість як ключові чинники, тоді як лінійна регресія приписує найбільшу вагу впровадженим штучним змінним (індикаторам часу доби та інтенсивності викидів). Проаналізовано причини цих розбіжностей з огляду на нелінійні взаємодії та мультиколінеарність. Зроблено висновок, що для високоточного оперативного прогнозування рівня PM_{10} доцільно застосовувати Random Forest, тоді як проста лінійна модель може використовуватися для швидких попередніх оцінок та інтерпретації впливу окремих факторів.

Ключові слова: прогнозування; дрібнодисперсні частки; PM_{10}; лінійна регресія; Random Forest; важливість змінних; якість повітря.

O. LEBODKIN, E. M. VARLAMOV, O. M. SKAKALSKYI, O. A. PALAHUTA, N. S. TSAPKO

COMPARISON OF METHODS FOR FORECASTING PM_{10} CONCENTRATIONS IN KRYVYI RIH IN THE WINTER PERIOD

The paper compares two approaches to forecasting PM_{10} particulate matter concentrations - a classical statistical model (multiple linear regression) and a modern machine learning algorithm (Random Forest). The study object is Kryvyi Rih, one of the largest industrial cities in Ukraine known for its challenging environmental situation. The winter period of 2024-2025 was considered, during which continuous monitoring of PM_{10} and meteorological parameters was carried out using an automated Cairnet mini-station with certified sensors. Data preprocessing was performed (outlier noise removal, gap filling, normalization) and feature engineering applied, including categorical variables for time of day, day type (weekend/weekday) and emissions intensity level. Both models revealed similar daily and weekly cycles in PM_{10} concentration: peak values occurred in the evening and night, lowest - during daytime, consistent with expectations and literature. However, the forecast accuracy differed significantly: Random Forest achieved much higher determination ($R^2 \approx 0.72$ vs $R^2 \approx 0.27$ for regression) and halved mean absolute error. A comparison of factor importance for both models is presented: Random Forest identified atmospheric pressure, temperature and humidity as key drivers, whereas the linear regression assigns greatest weight to introduced artificial variables (time-of-day and emissions intensity indicators). The reasons for these discrepancies are analyzed in view of nonlinear interactions and multicollinearity. It is concluded that for high-precision real-time PM_{10} forecasting, Random Forest is advisable, while a simple linear model can be used for quick preliminary assessments and interpretation of individual factor effects.

Keywords: forecasting; particulate matter; PM_{10}; linear regression; Random Forest; feature importance; air quality.

Вступ. Пилові частинки фракції PM_{10} (аерозоль розміром до 10 мкм) є одним з найнебезпечніших забруднювачів атмосферного повітря у містах. Високі концентрації PM_{10} становлять загрозу для здоров'я населення, спричиняючи захворювання дихальної та серцево-судинної систем і підвищуючи смертність [1]. За даними Європейського агентства з довкілля, дрібнодисперсний пил залишається найбільшим екологічним ризиком для здоров'я у Європі [2]. В Україні промислові центри, зокрема місто Кривий Ріг, систематично фіксують перевищення граничних концентрацій пилу в приземному шарі атмосфери. Це обумовлює актуальність задачі прогнозування рівнів PM_{10} для своєчасного інформування населення та впровадження заходів із зниження викидів. Значний внесок у запиленість атмосфери міст вносять місцеві промислові та транспортні джерела [3]. У Кривому Розі - одному з найбільших металургійних центрів Європи - на забруднення повітря пилом впливають гірничо-збагачувальні комбінати, металургійні підприємства та інтенсивний автотранспорт. У зимовий період ситуація ускладнюється метеорологічними умовами: часті температурні інверсії і слабкі вітри сприяють накопиченню домішок у приземному шарі. В цих умовах традиційні підходи моніторингу (стаціонарні пости контролю) потребують підсилення засобами математичного моделювання, щоб робити короткострокові прогнози концентрацій забруднювачів. Нові регуляторні вимоги також стимулюють розвиток систем прогнозування. Зокрема, нова Директива ЄС 2024/2881 встановлює суттєво жорсткіші нормативи якості повітря (граничні річні концентрації $PM_{2.5}$ та PM_{10} знижено до 10 та 20 μг/м³ відповідно) і зобов'язує застосовувати моделювання для інформування про перевищення [4]. В Євросоюзі такі прогностичні моделі вже інтегруються у практику оцінки якості повітря. Для України актуальним є запровадження сучасних методів, зокрема машинного навчання, до задач екологічного моніторингу промислових регіонів.

Попередні дослідження Kamińska [5] у Вроцлаві (Польща) показало високу точність моделі Random Forest при прогнозуванні забруднення повітря з урахуванням дорожнього руху та метеорологічних параметрів. Rubal і Kumar [6] розробили еволюційно-адаптований підхід, який комбінує алгоритм диференційної еволюції та Random Forest для прогнозування концентрацій забруднювачів повітря. Stoimenova та співавт. [7] використали метод регресійних дерев для прогнозування рівнів PM_{10} у міських умовах. Chen та співавт. [8] проаналізували просторово-часові закономірності концентрації PM_{10} у Китаї із застосуванням підходу Random Forest до супутникових даних. Plocoste і Laventure [9] прогнозували концентрацію PM_{10} у країнах Карибського басейну за допомогою моделей машинного навчання. Zárate та Rodríguez [10] застосували Random Forest у моделі прогнозування рівнів PM_{10} у Мехіко. Abuouelezz та співавт. [11] провели порівняльний аналіз моделей машинного навчання для короткотермінового прогнозування $PM_{2.5}$ і PM_{10} в умовах OAE. Adamenko і Arkhypova [12] досліджували закономірності змін рівнів $PM_{2.5}$ і PM_{10} у атмосферному повітрі Прикарпаття. Чугай і Терземан [13] продемонстрували можливість прогнозування забруднення повітря NO_2 в Одесі з використанням моделей машинного навчання. Gupta та співавт. [14] здійснили порівняльний аналіз моделей машинного навчання для прогнозування індексу якості повітря.

Таким чином. алгоритми машинного навчання (дерева рішень, ансамблі, нейронні мережі) здатні врахувати нелінійні багатофакторні залежності і підвищити точність прогнозів забруднення повітря порівняно з традиційними статистичними моделями [15].

Мета роботи. Мета роботи полягає в підвищенні точності короткострокового прогнозування концентрацій PM_{10} у великому промисловому місті під час зимового періоду на основі даних моніторингу та сучасних методів аналізу даних. Для реалізації поставленої мети у дослідженні порівнюються два підходи до побудови прогнозних моделей: традиційний статистичний метод множинної лінійної регресії та сучасний метод машинного навчання Random Forest. Задачі дослідження передбачають збір і підготовку вихідних даних про концентрації PM_{10} та метеорологічні параметри, розробку та оптимізацію моделей обох типів, оцінювання їхньої точності та здатності відтворювати відомі закономірності, а також аналіз чинників, що найбільше впливають на рівень забруднення повітря взимку. За результатами порівняння визначений найбільш перспективний підхід до прогнозування та сформовані рекомендації щодо підвищення якості короткострокових прогнозів для подальшого впровадження у систему моніторингу атмосферного повітря.

Методика дослідження. Об'єктом дослідження обрано атмосферне повітря промислового міста Кривий Ріг у зимовий період (листопад 2024 р. - березень 2025 р.). Вимірювання концентрацій PM_{10} та пов'язаних метеорологічних параметрів здійснювалися за допомогою автономної міні‑станції ENVEA Cairnet, оснащеної лазерним сенсором Cairsens і датчиками газових домішок (H_2S, NH_3, NO_2, O_3, CO, SO_2). Станція встановлена приблизно на висоті 3 м над ґрунтом у Центрально‑Міському районі міста, забезпечує безперервний збір даних із кроком 15 хвилин і передачу їх онлайн та відповідає вимогам міжнародних стандартів якості даних EN 15267, MCERTS і EPA. Вона вимірює масову концентрацію часток PM_{10}, $PM_{2.5}$ та PM_1, газові домішки й метеорологічні величини (температуру, відносну вологість, атмосферний тиск, швидкість та напрям вітру); діапазон вимірювання для PM_{10} становить 0-1000 μг/м³, межа виявлення <5 μг/м³, дискретність 0,01 μг/м³.

Набір даних охоплює період з 1 листопада 2024 р. до 31 березня 2025 р. та містить понад 9000 середніх (15‑хвилинних) спостережень концентрацій PM_{10}. Паралельно реєструвалися температура, відносна вологість, атмосферний тиск, швидкість і напрям вітру, а також часові атрибути - година доби, день тижня, дата. В прцесі обробки даних впроваджено додаткові пояснювальні ознаки: категоріальний показник «час доби» з чотирма інтервалами (ніч - 00:00-06:00, ранок - 06:00-12:00, день - 12:00-18:00, вечір - 18:00-24:00), бінарну змінну «тип дня» (0 - робочий, 1 - вихідний) та числовий індекс «інтенсивність викидів», що описує умовний рівень антропогенного навантаження (ніч = 1,00, ранок = 1,75, день = 1,55, вечір = 2,15).

Перед побудовою моделей виконано очистку й підготовку даних. Спочатку видалено грубі аномальні сплески у рядах концентрацій та метеопараметрів (менше 0,5 % записів) за допомогою меж квартильного інтервалу; пропущені значення заповнено лінійною інтерполяцією або перенесенням попереднього показника. Для забезпечення рівнозначного впливу змінних усі кількісні показники було нормалізовано до нульового середнього і одиничної стандартної девіації.

Для короткострокового прогнозування концентрацій PM_{10} застосовано два підходи: базовий статистичний метод множинної лінійної регресії (МЛР) та ансамблевий метод машинного навчання Random Forest (RF).

Лінійна регресія (1), передбачає, що залежність між концентрацією PM_{10} та набором пояснювальних змінних є лінійною:

$$PM_{10,i}=\beta_0+\beta_1\cdot T_i+\beta_2\cdot Hi+\beta_3\cdot P_i+\beta_4\cdot W_i+\beta_5\cdot D_{ніч,i}+\beta_6\cdot D_{ранок,i}+\beta_7\cdot D_{день,i}+\beta_8\cdot D_{вечір,i}+\beta_9\cdot D_{вихідний,i}+\beta_{10}\cdot I_i+\varepsilon_i \quad (1)$$

Де PM10,i - концентрація PM_{10}, $\beta 0$ - $\beta 10$ - вагові коефіцієнти метеорологічних і часових змінних, ε і – випадкова похибка.

Таке припущення дозволяє інтерпретувати кожен коефіцієнт: наприклад, у нашій моделі зростання температури на 1 °C супроводжувалося зменшенням прогнозованої PM_{10} приблизно на 6,7 μг/м³

Random Forest (2), це ансамблевий алгоритм, який поєднує багато простих моделей та усереднює їхні результати. На відміну від лінійної регресії, він здатен відображати складні й нелінійні залежності між параметрами. У нашому випадку було використано 200 таких моделей

$$f(x) = (1/B)\ \Sigma\_\{b=1\}\^\{B\}\ f\_b(x) \qquad (2)$$

де: B - кількість дерев у моделі, f_b(x) - прогноз b-го дерева, f(x) - узагальнений прогноз Random Forest.

Обидві моделі навчалися на одній навчальній вибірці (1 листопада 2024 р. - 15 березня 2025 р.), що охоплює приблизно 85 % даних; тестування проводилося на вибірці за останні 16 днів березня (16-31.03.2025 р.), що дозволяє оцінити здатність моделей до узагальнення. Параметри підбиралися шляхом перехресної перевірки на навчальному наборі: для МЛР використано стандартну реалізацію з бібліотеки statsmodels у Python 3; для RF застосовано Random Forest Regressor із 200 деревами, максимальною глибиною 20 та критерієм розбиття MSE (Mean Squared Error - середня квадратична похибка). Якість прогнозів оцінювали за коефіцієнтом детермінації R^2, середньою абсолютною похибкою (MAE) та кореневою середньоквадратичною похибкою (RMSE), (наскільки в середньому наші прогнози відрізняються від реальних значень).

Результати та обговорення У цьому розділі наведено результати моделювання та аналіз їхньої достовірності. Обидва підходи підтвердили наявність вираженого добового циклу змін концентрації PM_{10} у Кривому Розі: вечірні та нічні години характеризуються підвищеним рівнем забруднення, а вдень фіксується спад до мінімуму. У середньому у нічний час (00:00-06:00) концентрація на ≈5 μг/м³ (≈12 %) вища за добовий рівень, тоді як вдень (12:00-18:00) вона на ≈10 μг/м³ (≈25 %) нижча за середнє значення. Ці закономірності зумовлюються як накопиченням пилу під час приземних інверсій та відсутності сонячного прогріву, так і зростанням антропогенного навантаження у вечірні години. На рис.1 рамкові діаграми показують медіани, квартилі та аномальні значення.

Вплив умовної інтенсивності викидів виявив тісний зв'язок із рівнем PM_{10}. Обидві моделі показують підвищення медіан концентрації зі збільшенням коефіцієнта інтенсивності від 1,00 (ніч) до 2,15 (вечір), що відображає внесок промислових і транспортних джерел. Водночас Random Forest демонструє більш згладжену прогресію: відсутні різкі стрибки, помітно зменшений вплив поодиноких пікових точок, особливо в категорії «вечір».

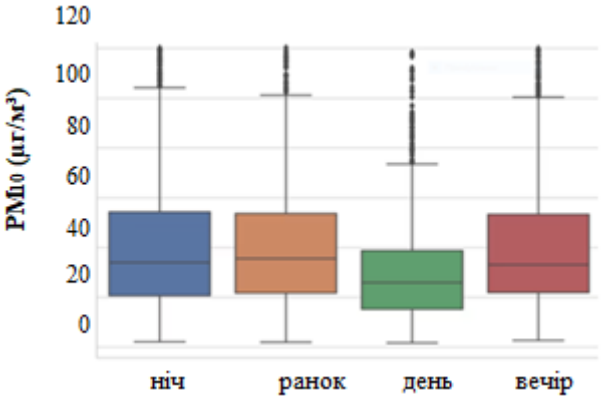

Рисунок 1 - Розподіл концентрацій PM_{10} для різних періодів доби

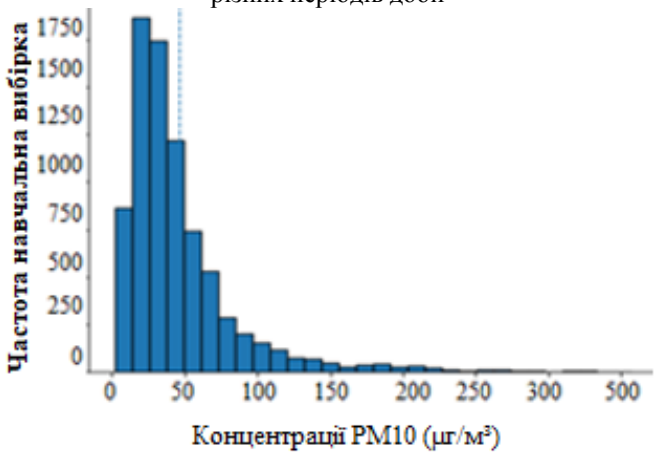

Рисунок 2 - Гістограми розподілу виміряних концентрацій PM_{10}, навчальна вибірка (грудень-середина березня)

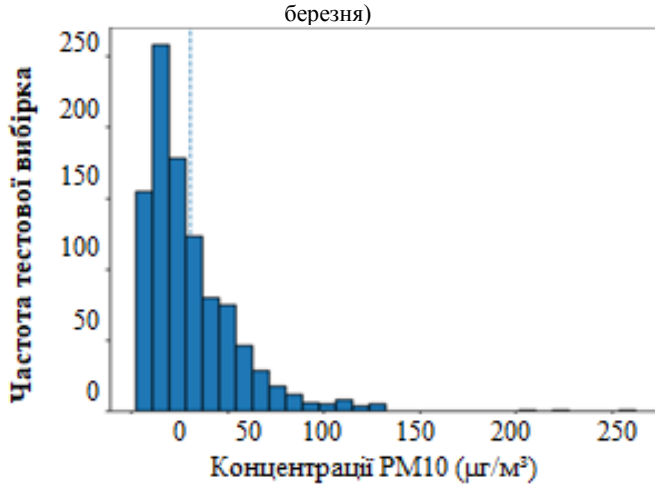

Рисунок 3 - Гістограми розподілу виміряних концентрацій PM_{10}, тестова вибірка (кінець березня)

Множинна лінійна регресія, навпаки, переоцінює розкид для максимального рівня і породжує значні відхилення.

Статистичний розподіл виміряних концентрацій наведено на гістограмах (рис. 2, рис.3). Понад 90 % значень зосереджено в діапазоні до 60 μг/м³, тоді як у «хвості» трапляються окремі піки понад 200 μг/м³. У навчальній вибірці таких аномалій більше, ніж у тестовій (кінець березня), вона характеризується сприятливішими умовами та нижчим середнім рівнем (~31 μг/м³ проти ~41 μг/м³ у повному наборі). Загалом більшу частину часу ситуація відносно стабільна, але іноді виникають короткочасні сплески через поєднання несприятливих метеорологічних умов і емісій.

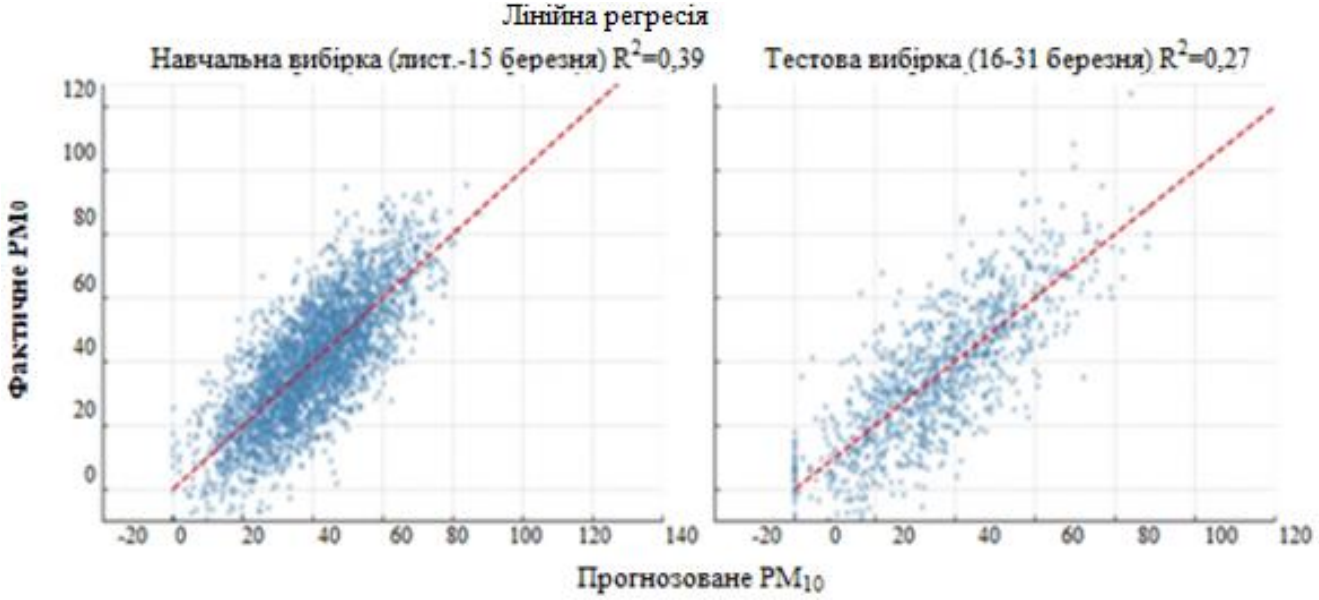

Рисунок 4 - Оцінка точності лінійної регресії

Порівняння точності моделей засвідчило перевагу Random Forest (рис. 4, рис5). На навчальній вибірці коефіцієнт детермінації R^2 становив ~0,39 для МЛР та Random Forest, тобто зменшилася більш ніж удвічі. ~0,96 для RF; на тестовій - відповідно ~0,27 і ~0,72. Середня абсолютна похибка склала близько 18 μг/м³ для регресійної моделі і лише ~8,5 μг/м³ для Random Forest, тобто зменшилася більш ніж удвічі.

Графіки фактичних проти прогнозованих значень (рис. 4, рис.5), демонструють сильне розсіювання у випадку лінійної регресії і щільне групування точок вздовж діагоналі y=x для RF (рис. 4, 5)

Аналіз залишків (помилок) показує, що у регресії вони зміщені й дисперсія зростає зі збільшенням прогнозованого рівня (гетероскедастичність): модель систематично недооцінює піки PM_{10} та переоцінює низькі концентрації. У Random Forest залишки розподілені симетрично навколо нуля, без помітної залежності від значення прогнозу, що вказує на відсутність систематичної похибки.

Рисунок 6 – показує залежність помилки прогнозу від рівня прогнозованої концентрації: для лінійної регегресії (залишки зміщені і зростають при збільшенні PM_{10}); для моделі - Random Forest (залишки рівномірні, систематична складова відсутня).

Таблиця 1 узагальнює оцінки впливу незалежних факторів у побудованих моделях. Для МЛР важливість фактора визначається модулем його коефіцієнта, тоді як у Random Forest - відносним внеском у зменшення помилки

У Random Forest найвпливовішими виявилися атмосферний тиск (~27 %), відносна вологість (~25 %) і температура повітря (~25 %), які сумарно забезпечують понад 77 % загального впливу на результат. Ці результати відповідають фізичним закономірностям: антициклони з низькою температурою і високою вологістю сприяють накопиченню пилу. Швидкість вітру (~8 %) має помірний вплив, решта змінних отримали невелику вагу (1-4 %). Натомість лінійна регресія надає високі коефіцієнти штучно введеним змінним: інтенсивності викидів, категоріям часу доби та типу дня; метерологічним умовам відведено меншу роль, оскільки кореляція між змінними розподіляє вагу між ними. Це свідчить, що лінійна модель не здатна адекватно врахувати нелінійні взаємозв'язки та приховані кореляції.

Переваги та недоліки розглянутих методів узагальнено в табл. 2. Множинна лінійна регресія є простою та легко інтерпретованою: вона швидко обчислюється, дозволяє чітко бачити внесок кожної змінної й придатна для попередньої оцінки. Проте лінійний підхід не враховує нелінійність і взаємодію факторів, чутливий до мультиколінеарності та викидів, що призводить до низької точності. Random Forest, навпаки, забезпечує високу точність, враховує складні

залежності і стійкий до аномалій; недоліками є більша ресурсомісткість та менша прозорість інтерпретації (модель - «чорний ящик»).

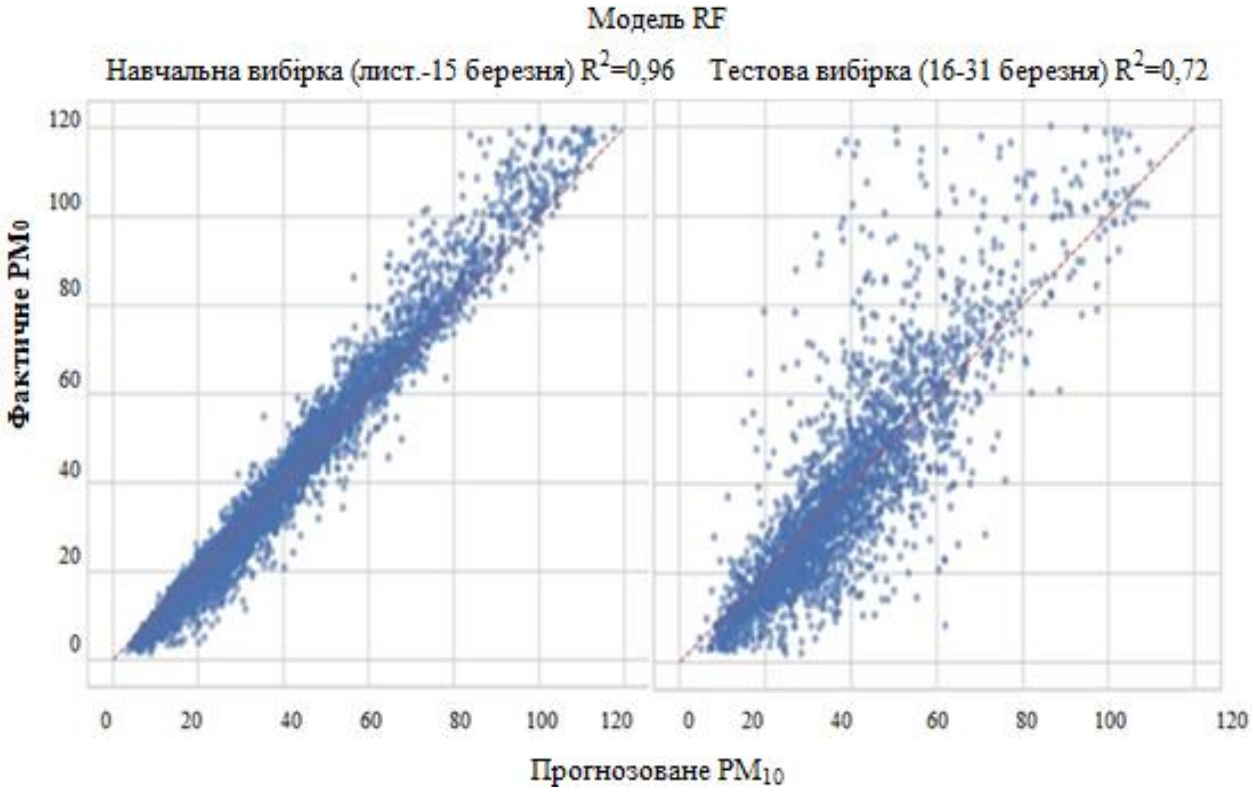

Рисунок 5 - Оцінка точності моделі Random Forest

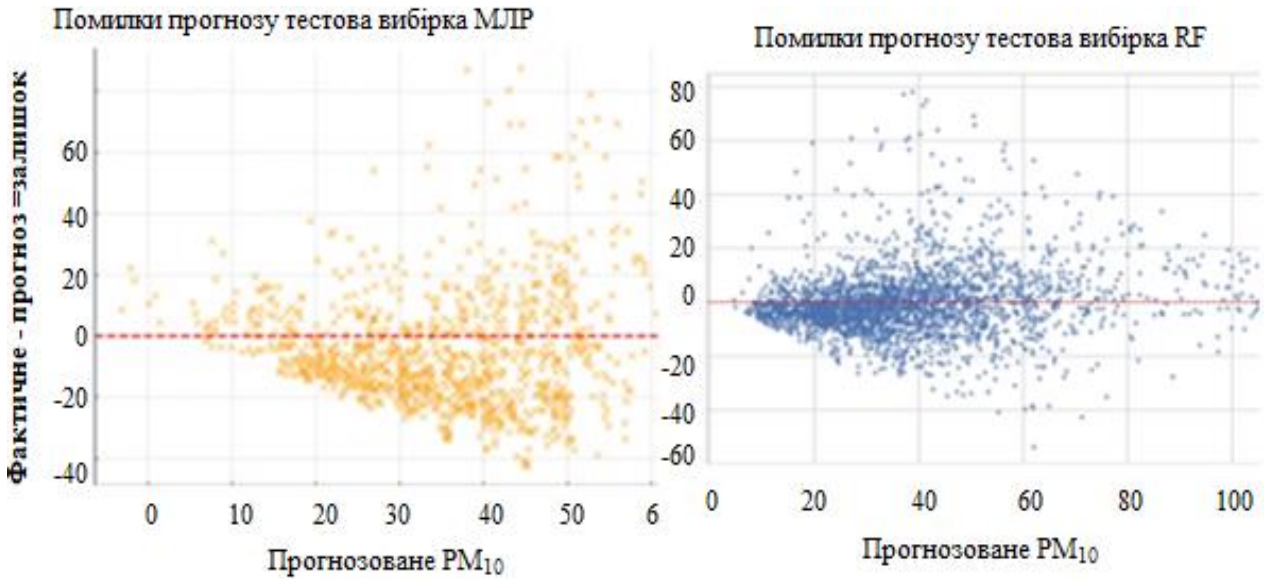

Рисунок 6 - Залежність помилки прогнозу від рівня прогнозованої концентрації

Таблиця 1 - Порівняння оцінок впливу факторів у моделях

Незалежна змінна	Модель лінійної регресії (μг/м³ на 1 од.)	Random Forest: відносна важливість (%)
Температура T (°C)	-6.70 - зі збільшенням температури на 1 °C PM_{10} зменшується приблизно на 6.7 мкг/м³.	≈ 25 % - одна з трьох найважливіших ознак; охолодження повітря суттєво підвищує рівень пилу.
Відносна вологість RH (%)	+0.20 - кожний відсоток вологості підвищує PM_{10} на 0.2 мкг/м³; ефект слабкий.	≈ 25 % - має великий вплив; висока вологість сприяє накопиченню аерозолю.
Атмосферний тиск P (мм рт. ст.)	-0.82 - підвищення тиску на 1 мм рт. ст. зменшує PM_{10} приблизно на 0.82 мкг/м³.	≈ 27 % - найважливіша ознака; високий тиск асоціюється з антициклонічними умовами та накопиченням частинок, тому модель активно враховує його зміну.
Напрям вітру SW (градуси)	+0.01 - ефект практично нульовий; напрямок вітру у градусах мало впливає на PM_{10}.	≈ 3-4 % - відіграє другорядну роль, але деякі сектори напрямків можуть мати локальні джерела пилу.
Швидкість вітру DW (м/с)	-15.89 - збільшення швидкості на 1 м/с зменшує PM_{10} приблизно на 16 мкг/м³, що відображає ефективне очищення атмосфери вітром.	≈ 8 % - четверта за впливом ознака; сильний вітер зменшує концентрації, хоча й може переносити пил з інших районів.
Тип дня (вихідний/робочий)	-6.10 - у вихідні PM_{10} нижчий на ~6.1 мкг/м³ порівняно з робочими днями.	≈ 3-4 % – має помітний, але невеликий вплив; зниження транспортної активності у вихідні зменшує забруднення.
Інтенсивність викидів	+20.37 - збільшення умовної шкали емісій на 1 пункт підвищує PM_{10} на 20.4 мкг/м³ (використовувались коефіцієнти 1.00, 1.55, 1.75, 2.15 для ночі, дня, ранку та вечора відповідно).	≈ 1-2 % - має невелику відносну вагу, оскільки частково корелює з метеопараметрами.
D_ранок (06:00–12:00)	-4.31 - вранці PM_{10} у середньому на 4.3 мкг/м³ нижче, ніж уночі.	≈ 1-2 % - добові категорії загалом мають низьку вагу.
D_день (12:00–18:00)	-10.67 - вдень PM_{10} у середньому на 10.7 мкг/м³ нижче, ніж уночі.	≈ 1-2 % - невеликий внесок.
D_вечір (18:00–24:00)	+5.08 - ввечері PM_{10} на 5.1 мкг/м³ вище від нічного рівня.	≈ 1-2 % - внесок невеликий.

Таблиця 2 - Переваги та недоліки прогнозних моделей

Змінна	Чому різна важливість
Атмосферний тиск (P)	У Random Forest тиск входить до трійки найважливіших параметрів (>25 % сумарної важливості), бо невеликі коливання тиску у поєднанні з температурою та вологістю істотно змінюють режим розсіювання часток. Лінійна регресія дає невеликий коефіцієнт ($\beta_3 \approx -0{,}82$) і, відповідно, низьку відносну важливість, оскільки тиск колінеарний з іншими метеопараметрами і його лінійний ефект частково «поглинається» температурою та швидкістю вітру.
Температура (T)	Random Forest фіксує, що низька температура сприяє накопиченню пилу (≈25 % важливості), оскільки за низьких температур часто спостерігається висока інтенсивність викидів і слабкий вітер. У лінійній регресії температура має помітний негативний коефіцієнт ($\beta_1 \approx -6{,}70$), але за абсолютною величиною вона поступається викидам і швидкості вітру.
Відносна вологість (RH)	Вологість у Random Forest отримує високу важливість (понад 25 %), бо модель враховує нелінійний вплив: підвищена вологість сприяє конденсації та злипанню часток, а в поєднанні з низькою температурою може значно підвищувати PM_{10}. У регресії коефіцієнт $\beta_2 \approx 0{,}20$ є малим; RH слабко корелює з PM_{10} й частково залежить від температури, тому її лінійний вплив невеликий.
Швидкість вітру (DW)	Для лінійної регресії це один із найсильніших факторів: $\beta_5 \approx -15{,}89$, що означає, що кожен метр/секунду зменшує PM_{10} майже на 16 мкг/м³. Random Forest оцінює швидкість вітру на рівні ≈8 % важливості: ефект вітру може перекриватися змінами температури та тиску, і модель розподіляє його внесок між корельованими ознаками.
Напрямок вітру (SW)	У лінійній моделі коефіцієнт $\beta_4 \approx +0{,}01$ практично нульовий, оскільки напрямок кодували одним числом (0–360°) і не врахували циклічність; модель не розрізняє близькі напрями (наприклад, 350° і 10°). Random Forest теж оцінює напрямок як малозначущий (~3–4 %), але трохи вище, бо дерева рішень можуть виявляти окремі «сектори», що асоціюються з підвищеним PM_{10}.
Weekend / тип дня	Вихідні дні мають негативний вплив у регресії ($\beta_6 \approx -6{,}10$) і помірну важливість у Random Forest (~3–4 %). Обидві моделі показують, що у вихідні PM_{10} нижче через менший транспорт і промислову активність, але Random Forest вважає цю змінну менш важливою, бо добові цикли вже частково враховані інтенсивністю викидів та метеопараметрами.
Інтенсивність викидів	У лінійній регресії це найвагоміший фактор ($\beta_7 \approx +20{,}37$), оскільки змінна відображає перехід від нічних до вечірніх пік, і модель не може захопити нелінійність добового циклу і взаємодію з погодою. У Random Forest інтенсивність викидів має лише 1–2 % важливості: добові цикли вже враховуються температурою, вітром і тиском, тому окрема ознака добавляє мало інформації.
Категорії часу доби (ранок, день, вечір)	У регресії ці даммі-змінні мають помітні коефіцієнти: $\beta_8 \approx -4{,}31$, $\beta_9 \approx -10{,}67$, $\beta_{10} \approx +5{,}08$. Вони потрібні, щоб описати добовий цикл у лінійній формі. Random Forest відносить їх до менш важливих (1–2 %), оскільки добові коливання моделюються через інші показники: температура, швидкість вітру і тиск змінюються протягом доби і дають змозі деревам рішень відокремити нічні та денні умови без спеціальних даммі-змінних.

Висновки. Короткострокове прогнозування концентрацій суспендованих часток PM_{10} у великому промисловому центрі (м. Кривий Ріг) виконано із застосуванням традиційної множинної лінійної регресії та алгоритму Random Forest. Отримані результати дозволяють сформулювати такі висновки:

1. Виявлено виражені добові та тижневі цикли зміни рівнів PM_{10}: накопичення пилу у вечірньо-нічний

період і денне зниження, а також зменшення середніх концентрацій у вихідні порівняно з буднями. Обидві моделі однаково відтворюють ці циклічні коливання, що відображено на відповідних рисунках.

2. Використання Random Forest забезпечує суттєве підвищення точності прогнозів. На відкладеній тестовій вибірці цей метод показав коефіцієнт детермінації близько 0,72 проти 0,27 для лінійної регресії, а середня абсолютна похибка зменшилася більш ніж удвічі. Random Forest не має систематичного зміщення, достовірно описує низькі та високі концентрації та краще відтворює статистичний розподіл вимірювань.

3. Проведений аналіз визначивосновними детермінантами рівня PM_{10} атмосферний тиск, температуру і відносну вологість, сумарний внесок яких у прогноз Random Forest перевищує 75 %. Антициклональні умови з високим тиском та морозною погодою сприяють накопиченню пилу, а циклональні процеси і вітер зменшують концентрацію. Швидкість і напрям вітру, тип дня та рівень викидів мають менший, але помітний вплив, що відображено у таблиці впливу факторів.

4. Модель множинної лінійної регресії доцільно застосовувати для швидких оцінок та інтерпретації впливу окремих змінних, але через припущення лінійності вона поступається Random Forest за точністю та стабільністю.

5. Рекомендовано впроваджувати системи прогнозування, що поєднують дані міні-станцій моніторингу з сучасними алгоритмами машинного навчання, зокрема Random Forest. Такі системи дозволять в режимі реального часу попереджати про епізоди високого забруднення та сприятимуть прийняттю управлінських рішень щодо зниження викидів. Подальші дослідження варто спрямувати на використання складніших моделей (градієнтний бустинг, нейронні мережі) та врахування додаткових чинників для підвищення точності і пояснювальної здатності прогнозів.

Список літератури

1. World Health Organization (WHO). WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: WHO; 2021. https://www.who.int/publications/i/item/9789240034228

2. Директива (ЄС) 2024/2881 Європейського Парламенту та Ради від 23 жовтня 2024 року про якість атмосферного повітря та чистіше повітря для Європи (перероблена редакція). Official Journal of the EU, L 309, 20.11.2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024L2881&qid=1758282016559

3. European Environment Agency (EEA). Air Quality in Europe - 2022 Report. EEA Report No.13/2022. Copenhagen, 2022. https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2022

4. Arnika NGO. Poisonous air: Satellites of the European Space Agency testify to the most polluted areas of Ukraine. News release, 23.11.2020. [Електронний ресурс]. https://arnika.org/en/news/poisonous-air-satellites-of-the-european-space-agency-testify-to-the-most-polluted-areas-of-ukraine

5. Kamińska J.A. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław. Journal of Environmental Management. 2018;217:164-174. https://doi.org/10.1016/j.jenvman.2018.03.094

6. Rubal S., Kumar D. Evolving differential evolution method with random forest for prediction of air pollution. Procedia Computer Science. 2018;132:824-833. https://doi.org/10.1016/j.procs.2018.05.094

7. Stoimenova M., et al. Regression trees modeling and forecasting of PM_{10} air pollution in urban areas. AIP Conference Proceedings. 2017;1895:030005. http://dx.doi.org/10.1063/1.5007364

8. Chen G., et al. Spatiotemporal patterns of PM_{10} concentrations over China during 2005-2016: a satellite-based estimation using the random forests approach. Environmental Pollution. 2018;242(A): 605-613. https://doi.org/10.1016/j.envpol.2018.07.012

9. Plocoste T., Laventure S. Forecasting PM_{10} concentrations in the Caribbean area using machine learning models. Atmosphere. 2023;14(1):134. https://doi.org/10.3390/atmos14010134

10. Zárate A.R., Rodríguez A.A. Application of Random Forest in a predictive model of PM_{10} particles in Mexico City. Nature Environment and Pollution Technology. 2024;23(2):711-724. https://neptjournal.com/upload-images/(9)D-1554.pdf

11. Abuouelezz W., Ali N., Aung Z., Altunaiji A., Shah S.B., Gliddon D. Exploring $PM_{2.5}$ and PM_{10} ML forecasting models: a comparative study in the UAE. Scientific Reports. 2025;15:9797. https://doi.org/10.1038/s41598-025-94013-1

12. Адаменко С. Я., Архипова Л. М. Дослідження закономірностей змін $PM_{2.5}$ та PM_{10} в атмосферному повітрі Прикарпаття. Екологічна безпека та природокористування. 2024;3(51): 47-58. https://ekmair.ukma.edu.ua/items/3fdcfc77-483d-4890-bd28-f11ff41c0a19

13. Чугай А.В., Терземан В.В. Прогнозування забруднення атмосферного повітря міста Одеса діоксидом азоту. Одеський держ. екологічний ун-т, 2020. Препринт, 12 с. http://dx.doi.org/10.32782/pcsd-2022-1-12

14. Gupta N.S., et al. Prediction of air quality index using machine learning techniques: a comparative analysis. J. Environ. Public Health. 2023;2023:4916267. http://dx.doi.org/10.1155/2023/4916267

15. Breiman L. Random Forests. Machine Learning. 2001;45(1): 5-32. https://link.springer.com/article/10.1023/A:1010933404324

References (transliterated)

1. World Health Organization (WHO). WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. Geneva: WHO; 2021. https://www.who.int/publications/i/item/9789240034228
2. Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe (recast). Official Journal of the EU, L 309, 20.11.2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024L2881&qid=1758282016559
3. European Environment Agency (EEA). Air Quality in Europe - 2022 Report. EEA Report No.13/2022. Copenhagen, 2022. https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2022
4. Arnika NGO. Poisonous air: Satellites of the European Space Agency testify to the most polluted areas of Ukraine. News release, 23.11.2020. [Online]. https://arnika.org/en/news/poisonous-air-satellites-of-the-european-space-agency-testify-to-the-most-polluted-areas-of-ukraine
5. Kamińska J.A. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wroclaw. Journal of Environmental Management. 2018;217:164-174. https://doi.org/10.1016/j.jenvman.2018.03.094
6. Rubal S., Kumar D. Evolving differential evolution method with random forest for prediction of air pollution. Procedia Computer Science. 2018;132:824-833. https://doi.org/10.1016/j.procs.2018.05.094
7. Stoimenova M., et al. Regression trees modeling and forecasting of PM10 air pollution in urban areas. AIP Conference Proceedings. 2017;1895:030005. http://dx.doi.org/10.1063/1.5007364
8. Chen G., et al. Spatiotemporal patterns of PM10 concentrations over China during 2005-2016: a satellite-based estimation using the random forests approach. Environmental Pollution. 2018;242(A):605-613. https://doi.org/10.1016/j.envpol.2018.07.012
9. Plocoste T., Laventure S. Forecasting PM10 concentrations in the Caribbean area using machine learning models. Atmosphere. 2023;14(1):134. https://doi.org/10.3390/atmos14010134
10. Zarate A.R., Rodriguez A.A. Application of Random Forest in a predictive model of PM10 particles in Mexico City. Nature Environment and Pollution Technology. 2024;23(2):711-724. https://neptjournal.com/upload-images/(9)D-1554.pdf
11. Abuouelezz W., Ali N., Aung Z., Altunaiji A., Shah S.B., Gliddon D. Exploring PM2.5 and PM10 ML forecasting models: a comparative study in the UAE. Scientific Reports. 2025;15:9797. https://doi.org/10.1038/s41598-025-94013-1
12. Adamenko S.Ya., Arkhypova L.M. Study of patterns of PM2.5 and PM10 changes in Prykarpattia air. Ekologichna Bezpeka ta Pryrodokorystuvannya. 2024;3(51):47-58. https://ekmair.ukma.edu.ua/items/3fdcfc77-483d-4890-bd28-f11ff41c0a19
13. Chuhai A.V., Terzeman V.V. Forecasting air pollution of the city of Odesa by nitrogen dioxide. Odessa State Environmental University, 2020. Preprint, 12 p. http://dx.doi.org/10.32782/pcsd-2022-1-12
14. Gupta N.S., et al. Prediction of air quality index using machine learning techniques: a comparative analysis Journal of Environmental Public Health. 2023;2023:4916267 2001;45(1):5-32. http://dx.doi.org/10.1155/2023/4916267.
15. Breiman L. Random Forests. Machine Learning https://link.springer.com/article/10.1023/A:1010933404324

Відомості про авторів / About the Author

Лебьодкін Є. О. (Lebodkin Ievgen) – аспірант, Науково-дослідна установа «Український науково-дослідний інститут екологічних проблем», м. Харків, Україна; ORCID: https://orcid.org/0009-0006-9188-9037; e-mail: lebyodkin@gmail.com.

Варламов Є. М. (Varlamov Yevhenii) – кандидат технічних наук, старший науковий співробітник, Науково-дослідна установа «Український науково-дослідний інститут екологічних проблем», м. Харків, Україна; ORCID: https://orcid.org/0000-0002-3405-1784; e-mail: varlamov.niiep@gmail.com

Скакальський О. М. (Skakalskyi Oleksandr) – начальник управління екології, Криворізька міська рада, м. Кривий Ріг, Україна; ORCID: https://orcid.org/0009-0009-9025-8438, e-mail: cedaplus@gmail.com

Палагута О. А. (Palahuta Oksana) - кандидат технічних наук, Науково-дослідна установа «Український науково-дослідний інститут екологічних проблем», м. Харків, Україна; ORCID: https://orcid.org/0009-0008-4641-9903; E- mail: ksu_ksenichka@ukr.net

Цапко Н. С. (Tsapko Nataliia) – кандидат технічних наук., доцент, Науково-дослідна установа «Український науково-дослідний інститут екологічних проблем», м. Харків, Україна; ORCID: https://orcid.org/0000-0003-2480-3636; e-mail: tsapkonatali@gmail.com