УДК 004.85: 004.942

А. М. Дудар
Є. М. Крижановський

ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ЯКОСТІ АТМОСФЕРНОГО ПОВІТРЯ МІСТА ВІННИЦІ

Вінницький національний технічний університет

Анотація

Дану роботу присвячено розробці інтелектуальної системи прогнозування концентрації дрібнодисперсного пилу PM2.5 атмосферного повітря міста Вінниці з використанням ансамблевих методів машинного навчання. В рамках дослідження сформовано автоматизований ETL-пайплайн для збору та передобробки даних, здійснено розвідувальний аналіз часових рядів, побудовано та навчено ансамбль моделей градієнтного бустингу (LightGBM, XGBoost, CatBoost) із байєсівською оптимізацією гіперпараметрів. Найкраща модель LightGBM досягла коефіцієнта детермінації $R^2 = 0.90$ на горизонті t+1 та $R^2 = 0.45$ на горизонті t+24, що підтверджує ефективність запропонованого підходу.

Ключові слова: машинне навчання, PM2.5, градієнтний бустинг, передобробка даних, часові ряди, прогнозування якості повітря.

Abstract

This paper presents an intelligent system for forecasting PM2.5 fine particulate matter concentration in the atmospheric air of Vinnytsia city using ensemble machine learning methods. The study encompasses the development of an automated ETL pipeline for data collection and preprocessing, exploratory time series analysis, and training of a gradient boosting ensemble (LightGBM, XGBoost, CatBoost) with Bayesian hyperparameter optimization. The best-performing LightGBM model achieved $R^2 = 0.90$ at the t+1 forecast horizon and $R^2 = 0.45$ at t+24, confirming the effectiveness of the proposed approach.

Keywords: machine learning, PM2.5, gradient boosting, data preprocessing, time series, air quality forecasting.

Вступ

Забруднення атмосферного повітря дрібнодисперсним пилом PM2.5 є однією з найгостріших екологічних проблем сучасності, що безпосередньо впливає на здоров'я населення та якість міського середовища [1]. Традиційні статистичні методи прогнозування часових рядів — зокрема авторегресійні моделі SARIMAX — є недостатньо ефективними в умовах складних нелінійних залежностей між метеорологічними факторами та концентрацією забруднювача. Тому актуальним є застосування ансамблевих методів машинного навчання, зокрема градієнтного бустингу, що здатні автоматично виявляти приховані просторово-часові закономірності у великих масивах екологічних даних [2].

Підготовка даних

Вхідними даними системи є часові ряди вимірювань станції моніторингу атмосферного повітря Вінницького національного технічного університету за 2022–2026 роки, що налічують 41 723 погодинних спостереження. Збір даних реалізовано через автоматизований ETL-пайплайн з використанням бібліотек BeautifulSoup та requests для вилучення CSV-ресурсів з відкритого порталу моніторингу Вінницької міської ради. Зібрані дані зберігаються у базі даних MySQL, спроектованій за схемою «Зірка», що забезпечує швидке виконання OLAP-запитів над понад 7 млн записів.

Ключовою особливістю набору даних є значна часова фрагментарність: попри загальний горизонт у 1 237 днів, валідні спостереження зафіксовано лише для 633 діб, що формує рівень денного покриття 51.2%. Теплова карта повноти даних (рис. 1) наочно демонструє хаотичне чергування активних і «мовчазних» сегментів моніторингу протягом 2022–2026 років без чітко вираженої регулярності.

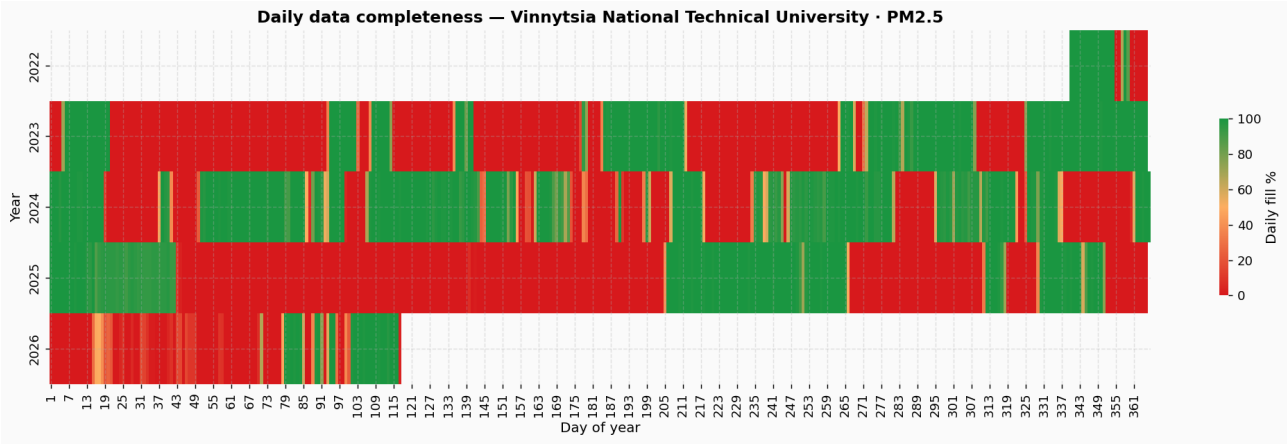

Рис. 1. Теплова карта повноти даних PM2.5 на станції ВНТУ (2022–2026)

Така критична фрагментарність унеможливлює застосування класичних моделей SARIMAX, що вимагають суворої безперервності часового ряду для коректного оцінювання авторегресійних та сезонних компонент. Натомість методи градієнтного бустингу природно стійкі до нерегулярних пропусків, оскільки оперують вектором ознак окремого спостереження, а не цілісною послідовністю. Тому вибір бустингових ансамблів є статистично обґрунтованим рішенням для даного набору даних [2, 3].

Передобробка включає: фізичну верифікацію значень та чотирирівневу корекцію аномалій (лінійна інтерполяція в часовому вікні ±6 годин, метод найближчого сусіда, медіана за 30 діб, глобальна крос-станційна медіана); дискретизацію часової сітки з кроком 1 година; тригонометричне кодування циклічних дескрипторів (година, день тижня, місяць). Простір ознак розширено авторегресійними лагами PM2.5 (1–72 год), ковзними статистиками у вікнах 3–48 год, експоненційно зваженими середніми та просторовими показниками крос-кореляції між станціями. Для запобігання витоку інформації всі динамічні ознаки обчислювались зі зміщенням [3, 4].

Імплементація моделей машинного навчання

Задача реалізована як пряме багатоступеневе прогнозування (Direct Multi-Step Forecasting): для кожного з 24 часових горизонтів навчається окрема незалежна модель, що мінімізує накопичення помилок, властиве рекурсивним підходам. Весь хронологічно впорядкований масив розділено на тренувальну (80%) та тестову (20%) вибірки без перемішування, що гарантує відсутність витоку майбутніх значень у навчання. Якість кожної моделі додатково контролювалась часовою крос-валідацією з п'ятьма фолдами (TimeSeriesSplit), де кожен наступний фолд містить більш пізні спостереження, ніж попередній.

В системі реалізовано конвеєр із трьох базових алгоритмів градієнтного бустингу та одного мета-алгоритму стекінгу. LightGBM застосовує гістограмний метод апроксимації розподілу ознак та стратегію росту дерева по листках (leaf-wise), що забезпечує швидке навчання при великому числі ознак і є основним алгоритмом з індивідуальним автоналаштуванням для кожного горизонту. XGBoost будує дерева по рівнях (level-wise) з вбудованою L1- та L2-регуляризацією, що обмежує перенавчання і забезпечує кращу узагальнюючу здатність на середніх горизонтах. CatBoost використовує симетричні дерева та власний механізм Ordered Target Statistics для обробки категоріальних ознак (час доби, місяць, сегмент) без попереднього one-hot кодування, що зменшує зсув градієнта і підвищує стабільність на середньострокових прогнозах [4, 5].

Фінальний рівень стекінгу поєднує прогнози трьох базових моделей для кожного горизонту у мета-ознаки, на яких навчається Ridge-регресія з L2-регуляризацією ($\alpha = 1.0$). Мета-модель навчається виключно на out-of-fold прогнозах базових алгоритмів, отриманих під час крос-валідації, що запобігає перенавчанню стекінгу. Такий дворівневий ансамбль дозволяє згладити систематичні відхилення окремих алгоритмів та підвищити стабільність прогнозу в перехідних зонах концентрацій.

Центральним елементом імплементації є автоматичний підбір гіперпараметрів засобами фреймворку Optuna. Байєсівська оптимізація на основі алгоритму TPE (Tree-structured Parzen Estimator) будує імовірнісну сурогатну модель простору параметрів і за 50 ітерацій знаходить конфігурацію з мінімальною MSE на крос-валідаційній вибірці. Пошук проводиться індивідуально для кожного з 24 горизонтів у просторі: кількість листків (num_leaves: 20–300), швидкість навчання (learning_rate: 0.01–

0.15), частка ознак (feature_fraction: 0.4–1.0), мінімальна кількість зразків у листку (min_child_samples: 5–100) та коефіцієнти регуляризації (λ_1, λ_2: 1e-8–10.0). Завдяки такій спеціалізації моделі для коротких горизонтів отримують більш глибоку структуру дерев для вловлювання миттєвих змін, тоді як моделі для далеких горизонтів тяжіють до більшої регуляризації для узагальнення добових трендів [5].

Навчені структури дерев, коефіцієнти мета-моделі та параметри нормування зберігаються у спеціалізованих форматах (.joblib для LightGBM та XGBoost, .cbm для CatBoost, .json для метаданих) у локальному репозиторії ./saved_models/. Це забезпечує миттєве розгортання системи в продуктивному середовищі без повторного навчання та можливість інкрементального оновлення окремих горизонтів при надходженні нових даних моніторингу.

Результати дослідження

Комплексна евалюація ансамблю проведена за метриками R^2, RMSE та MAPE на тестовій вибірці для всіх 24 горизонтів прогнозування (рис. 2). LightGBM з байєсівською оптимізацією є беззаперечним лідером: $R^2 = 0.90$ при t+1 та $R^2 = 0.45$ при t+24 — найкращий результат серед усіх моделей у довгостроковій перспективі. CatBoost демонструє $R^2 = 0.86$ при t+1 та найбільш рівномірну деградацію, залишаючись оптимальним для середньострокового прогнозування (6–18 год). XGBoost фіксує $R^2 = 0.73$ при t+1 та відносну стійкість у зоні t+5–t+8 завдяки вбудованій регуляризації [4, 5].

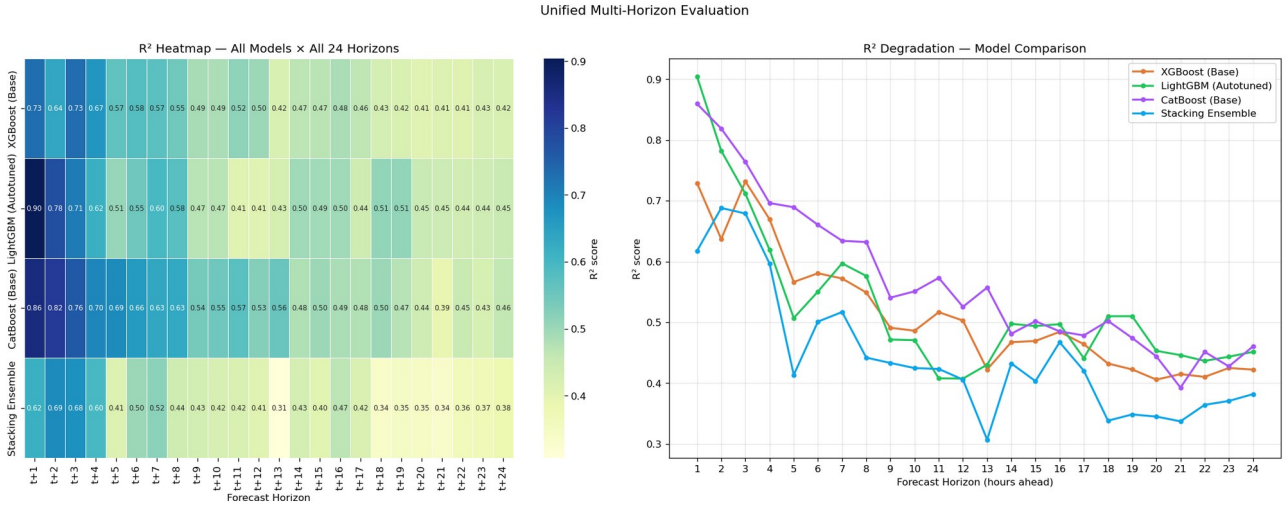

Рис. 2. Зведена мультигоризонтна оцінка: теплова карта R^2 та графік деградації точності моделей

Аналіз важливості ознак виявив тристадійну структуру прогнозування: на коротких горизонтах (t+1–t+4, MAPE 5.3–9.4%) домінує авторегресійний режим із ключовою роллю pm25_lag1; на середніх (t+5–t+12, MAPE 10.4–14.2%) зростає вплив метеопараметрів та ковзних середніх; на довгих (t+13–t+24, MAPE 14.1–16.2%) модель переходить до кліматологічно-просторового режиму з домінуванням сезонних ознак (dayofyear, month_sin) та просторових концентрацій сусідніх станцій. Така структура відображає багатомасштабну природу динаміки PM2.5 і свідчить про коректне засвоєння моделлю фізичних закономірностей атмосферних процесів.

Висновки

У результаті дослідження розроблено та програмно реалізовано систему багатокрокового прогнозування концентрації PM2.5 на основі ансамблю алгоритмів градієнтного бустингу. Застосування SARIMAX є статистично некоректним для даного набору через критичну фрагментарність часових рядів (48.8% пропущених діб), тоді як бустингові методи природно стійкі до нерегулярних пропусків. Запропонований підхід до передобробки та інжинірингу ознак у поєднанні з дворівневою архітектурою стекінгу та байєсівською оптимізацією гіперпараметрів забезпечив $R^2 = 0.90$ при короткостроковому та $R^2 = 0.45$ при добовому прогнозуванні. Практична цінність системи полягає у можливості інтеграції до міських інформаційно-аналітичних платформ для оперативного управління якістю повітря та своєчасного попередження населення про небезпечні рівні забруднення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Шмундяк Д. О., Мокін В. Б., Крижановський Є. М. Системний аналіз стану атмосферного повітря регіону, з урахуванням впливу аномалій : монографія. Вінниця : ВНТУ, 2025. 169 с.
2. Ахмадіанфар І. та ін. Towards intelligent air quality forecasting using integrated machine learning framework with variational mode decomposition and catboost feature selection. Scientific Reports. 2026. DOI: 10.1038/s41598-025-33785-y.
3. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Навчальний посібник. Вінниця : ВНТУ, 2024. С. 89–128.
4. Джеймс Г., Віттен Д., Хасті Т., Тібширані Р. Деревоподібні методи. An Introduction to Statistical Learning. Springer, Cham, 2023. P. 331–342. DOI: 10.1007/978-3-031-38747-0_8.
5. Сібінді Р., Мвангі Р. В., Вайтіту А. Г. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports. 2023. Vol. 5, Iss. 4. DOI: 10.1002/eng2.12599.

Дудар Анатолій Михайлович *– студент групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: gigsollt@gmail.com*

Крижановський Євгеній Миколайович *– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця.*

Dudar Anatolii M. *– student of Faculty of Intelligent Information Technologies and Automation, SA-22b, Vinnytsia National Technical University, Vinnytsia, e-mail: gigsollt@gmail.com*

Kryzhanovsky Yevhenii M. *– Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia.*