

Д. О. Шмундяк¹
І. В. Варчук¹

СТЕКІНГОВА ГІБРИДНА МОДЕЛЬ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

¹Вінницький національний технічний університет

В умовах зростання екологічних ризиків задача прогнозування рівнів забруднення набуває все більшої актуальності. Традиційні статистичні моделі, наприклад, ARIMA демонструють високу точність на фонових рівнях забруднення, проте втрачають якість прогнозу у разі пікових концентрацій, що є критичними з погляду впливу на здоров'я. Моделі машинного навчання, зокрема градієнтний бустинг, навпаки, краще відтворюють нелінійні піки, проте їхня точність суттєво знижується зі збільшенням терміну прогнозу через поступове накопичення похибок на кожному кроці розрахунку. Зазначені обмеження зумовлюють необхідність розробки гібридних підходів, здатних поєднати переваги обох класів моделей.

Розроблено та досліджено гібридну стекінгову архітектуру прогнозування погодинного часового ряду, що поєднує структурну модель Prophet та алгоритм градієнтного бустингу LightGBM. На відміну від традиційного підходу корекції залишків, у запропонованій архітектурі прогноз Prophet використовується як детермінована сезонна ознака поряд з авторегресійними та календарними ознаками, на основі яких LightGBM безпосередньо передбачає цільове значення.

Реалізацію та тестування запропонованої моделі виконано з використанням реальних даних моніторингу якості атмосферного повітря, отриманих від мережі громадського моніторингу атмосферного повітря EcoCity. Статистичну значущість результатів підтверджено критерієм Діболда–Маріано. Порівняння з базовими моделями показало, що гібридна модель статистично значуще перевершує моделі-базиси та демонструє найзбалансованішу точність. Отримані результати свідчать про перевагу гібридної архітектури над моделями-базисами як за загальною точністю, так і за стабільністю на довгих горизонтах прогнозування.

Ключові слова: прогнозування часових рядів, якість атмосферного повітря, PM2.5, гібридна модель, Prophet, LightGBM, градієнтний бустинг, стекінг, машинне навчання, EcoCity.

Вступ

Прогнозування показників стану атмосферного повітря є актуальною науковою та практичною задачею. Тривалий вплив підвищених концентрацій різних показників, в першу чергу PM2.5 (тверді частинки діаметром менше 2,5 мкм), веде до захворювань серцево-судинної та дихальної систем, що зумовлює необхідність своєчасного і точного аналізу та прогнозування таких даних. Задача ускладнюється нелінійним характером часового ряду концентрації PM2.5: разом з регулярною добовою та тижневою сезонністю спостерігаються раптові пікові викиди, зумовлені метеорологічними умовами, промисловою діяльністю тощо.

В науковій літературі для прогнозування якості атмосферного повітря широко застосовуються як класичні статистичні методи (ARIMA та Prophet) [1]—[4], так і різні алгоритми машинного навчання [5], наприклад, XGBoost і LightGBM [6], [7]. Статистичні моделі ефективно описують лінійну авторегресійну структуру ряду, проте не здатні моделювати нелінійні піки концентрації. Водночас алгоритми градієнтного бустингу краще вловлюють нелінійні залежності.

Перспективним напрямом подолання зазначених обмежень є гібридні моделі, що поєднують структурне моделювання часових рядів з алгоритмами машинного навчання. Більшість наявних гібридних підходів реалізують схему корекції залишків: структурна модель описує тренд і сезонність, а алгоритм машинного навчання будується на залишках. Проте така схема має суттєвий недолік — якість компонента машинного навчання повністю визначається якістю залишків, і якщо структурна модель погано описує ряд, залишки містять змішаний сигнал, що створює залежність якості моделі від точності структурної компоненти. Альтернативним підходом є стекінг, де прогноз моделі



використовується як одна з ознак для алгоритму машинного навчання, який безпосередньо передбачає цільове значення. Такий підхід дозволяє самостійно визначати міру довіри до структурного прогнозу в кожній конкретній ситуації.

Окремий клас підходів до прогнозування часових рядів становлять рекурентні нейронні мережі, зокрема LSTM [5], які ефективно моделюють довготривалі залежності у послідовних даних. Проте їх застосування обмежується необхідністю значного обсягу даних для стабільного навчання, низькою інтерпретованістю результатів. З огляду на зазначені обмеження, у цій роботі вибрано підхід на основі інтерпретованих моделей Prophet та LightGBM, що забезпечують прийнятну точність на наявному обсязі даних.

Мета дослідження — підвищення збалансованості точності погодинного прогнозування концентрації PM_{2.5} на всьому діапазоні спостережуваних значень, зокрема пікові концентрації, шляхом використання стекінгової гібридної архітектури на основі Prophet та LightGBM.

Формалізація задачі

Нехай $y(t)$ — погодинна концентрація PM_{2.5} у момент часу t , де $t=1,2,\dots,N$ з кроком 1 година. Задача полягає у побудові моделі, яка за відомими до моменту t даними передбачає значення $\hat{y}(t)$. Ряд концентрації PM_{2.5} складається з кількох складових [1]

$$y(t) = g(t) \cdot s(t) \cdot h(t) + \varepsilon_t, \quad (1)$$

де $g(t)$ — тренд, $s(t)$ — сезонна складова, $h(t)$ — вплив аномалій, ε_t — випадкова складова з нелінійними піками.

У моделі Prophet використано мультиплікативну схему декомпозиції, оскільки вона дозволяє масштабувати сезонну складову пропорційно рівню ряду та є придатнішою для опису часових рядів концентрацій PM_{2.5} з вираженими коливаннями інтенсивності забруднення. Прогноз Prophet $\hat{y}_P(t)$ є детермінованою функцією часу і може бути отриманий для будь-якого майбутнього моменту без накопичення похибки.

У класичному гібридному підході Prophet буде базовий прогноз, а алгоритм машинного навчання навчається коригувати його залишки

$$\hat{y}(t) = \hat{y}_P(t) + \mathcal{M}[y(t) - \hat{y}_P(t)]. \quad (2)$$

Недолік такого підходу в тому, що якість корекції повністю залежить від якості залишків. Якщо Prophet помиляється на певному відрізку ряду, то залишки містять змішаний сигнал і алгоритм навчається на шумі. У запропонованому підході прогноз $\hat{y}_P(t)$ не коригується, а використовується як одна з вхідних ознак і LightGBM [5]—[7] безпосередньо передбачає цільове значення $y(t)$

$$\hat{y}(t) = f_{\text{LGBM}}[x(t)], \quad (3)$$

де $x(t)$ — вектор ознак, що включає авторегресійні лаги, ковзні статистики, прогноз Prophet та календарні ознаки.

Основною відмінністю запропонованого підходу від підходу корекції залишків є те, що модель f_{LGBM} сама визначає, наскільки довіряти прогнозу Prophet у кожний момент часу: у режимі нормальних концентрацій спирається на нього більше, у пікових — менше, компенсуючи це через лагові ознаки (значення ряду у попередні моменти часу відносно поточного кроку прогнозу).

Архітектура запропонованої моделі

Етап 1. Навчання структурної компоненти (Prophet)

Модель Prophet навчається на повній тренувальній вибірці з усіма параметрами сезонності за замовчуванням та додатковою місячною сезонністю. Після навчання моделі формуються in-sample оцінки для тренувального періоду та прогноз для тестового горизонту (рис. 1)

$$\hat{y}_P(t), \quad t = 1, \dots, N_{\text{train}} + N_{\text{test}}, \quad (4)$$

де N_{train} — обсяг тренувальної вибірки, N_{test} — обсяг тестової вибірки.

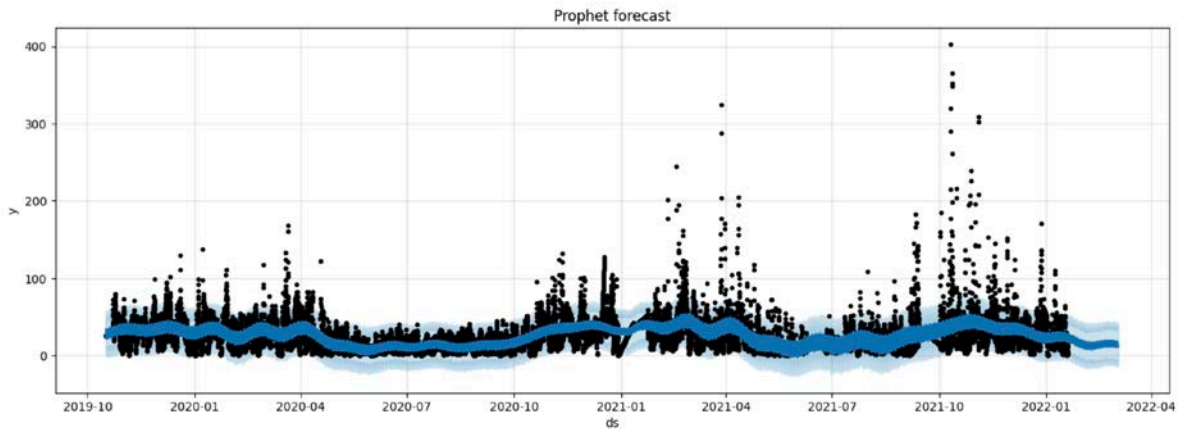


Рис. 1. Приклад прогнозу моделі Prophet

На тренувальному горизонті як ознака використовуються in-sample оцінки моделі Prophet, отримані після її навчання на тренувальній вибірці. Оскільки тестова вибірка не бере участі ні в навчанні Prophet, ні у формуванні ознак для навчання LightGBM, витік інформації з тестового горизонту відсутній. Прогноз Prophet у запропонованій архітектурі використовується не як фінальна модель прогнозування, а як джерело структурної інформації про тренд і сезонність.

Етап 2. Формування матриці ознак

Матриця ознак X формується на основі повного тренувального ряду $y(t)$ та прогнозу Prophet $\hat{y}_p(t)$ і включає п'ять груп ознак:

1. Лагові ознаки $y(t-1)$, $y(t-2)$, $y(t-3)$, $y(t-6)$, $y(t-12)$, $y(t-24)$, $y(t-48)$, $y(t-168)$ відображають авторегресійну структуру ряду на горизонтах від 1 години до 1 тижня.
2. Rolling-ознаки — середнє та стандартне відхилення $\bar{y}_6$, $\bar{y}_{12}$, $\bar{y}_{24}$, $\bar{y}_{168}$, що описують короткострокові та тижневі патерни.
3. Структурний прогноз $\hat{y}_p(t)$, що передає сезонний контекст від Prophet.
4. Годинні ознаки $\sin / \cos(2\pi \cdot \text{hour} / 24)$, 1_{rush} , 1_{night} , що додають добовий цикл підвищеного трафіку та промислової активності.
5. Календарні ознаки $\sin / \cos(2\pi \cdot \text{DoW} / 7)$, 1_{weekend} , $\sin / \cos(2\pi \cdot \text{month} / 12)$, що відображають тижневу та річну сезонність.

Етап 3. Навчання LightGBM

Модель LightGBM навчається на матриці X з цільовою змінною $y(t)$ з early stopping на валідаційній вибірці (останні 10 % тренувального ряду)

$$\hat{y}(t) = f_{\text{LGBM}}[x(t); \theta^*], \quad (5)$$

де θ^* — оптимальні параметри моделі, знайдені мінімізацією MSE на валідаційній вибірці.

Значення гіперпараметрів вибрано емпірично на основі попередніх експериментів з метою балансування між точністю моделі та запобіганням перенавчанню — рання зупинка після 50 раундів без покращення на валідаційній вибірці використовується як основний механізм регуляризації.

Етап 4. Рекурсивний прогноз

На тестовому горизонті прогноз здійснюється рекурсивно: передбачене значення $\hat{y}(t)$ додається до буфера history і використовується для обчислення лагових та rolling-ознак наступного кроку

$$\hat{y}(t) = f_{\text{LGBM}}[x(t)], \text{history} \leftarrow \text{history} \cup \{\hat{y}(t)\}. \quad (6)$$

Ознака $\hat{y}_p(t)$ та всі календарні ознаки є детермінованими і відомими для будь-якого майбутнього моменту часу, що принципово відрізняє їх від авторегресійних ознак і запобігає накопиченню похибки для цих компонент. Блок-схему алгоритму запропонованого методу показано на рис. 2.

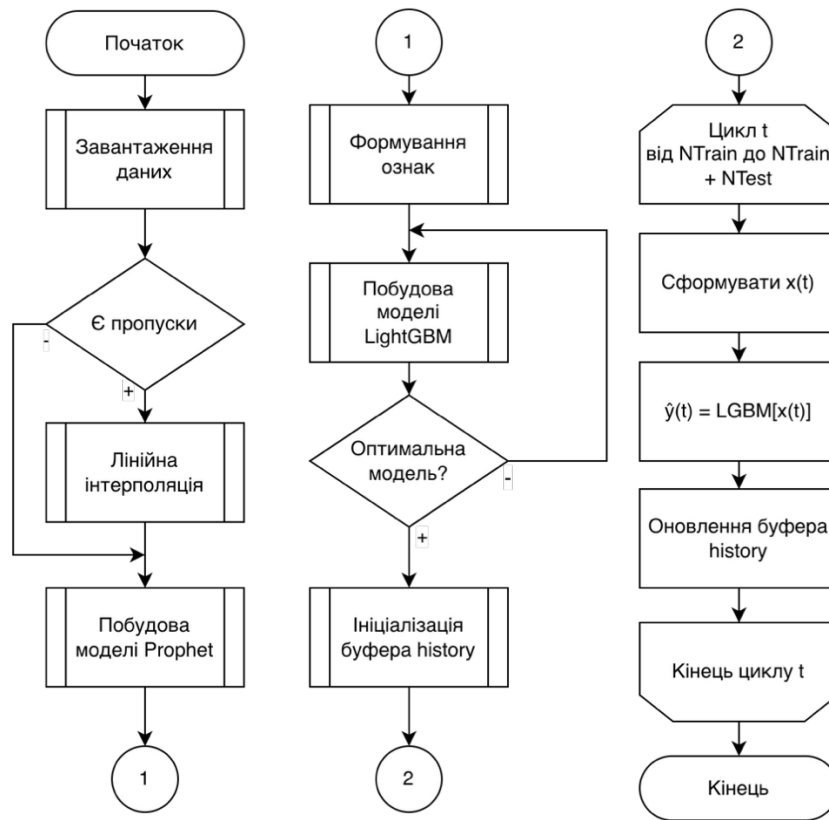


Рис. 2. Блок-схема алгоритму запропонованого методу

Приклад розв'язання задачі

Для випробування ефективності запропонованого методу використано реальні дані моніторингу стану атмосферного повітря, надані мережею громадського моніторингу EcoCity (<https://reborn.eco-city.org.ua/>), що функціонує у межах міжнародної програми «Чисте повітря для України» [8] (рис. 3). Дані для дослідження безпосередньо отримані через сервіс «Кабінет дослідника», доступ до якого існує завдяки угоді між EcoCity і ВНТУ. Вибрано станцію № 650 (с. Турбів) та показник $PM_{2.5}$, оскільки серед доступних станцій вона має найбільший обсяг даних за досліджуваний період — 35064 погодинних спостережень за 2021—2024 рр.

Вхідний набір даних становить 35064 погодинних спостережень. Після попередньої обробки, заповнення пропусків методом лінійної інтерполяції з подальшим застосуванням $bfill$ та $ffill$, кількість пропущених значень становила 0. Тренувальна вибірка охоплює 95 % спостережень (~3310 годин), тестова — 5 % (~1754 години), що відповідає приблизно 73 добам.

Застосуємо описаний метод.

Етап 1. Аналіз вхідного ряду та навчання Prophet.

Під час аналізу тренувального ряду виявлено аномальні значення концентрації $PM_{2.5}$ у період 2021 року, що суттєво перевищує значення наступних років і досягає $140...180 \text{ мкг/м}^3$. Модель

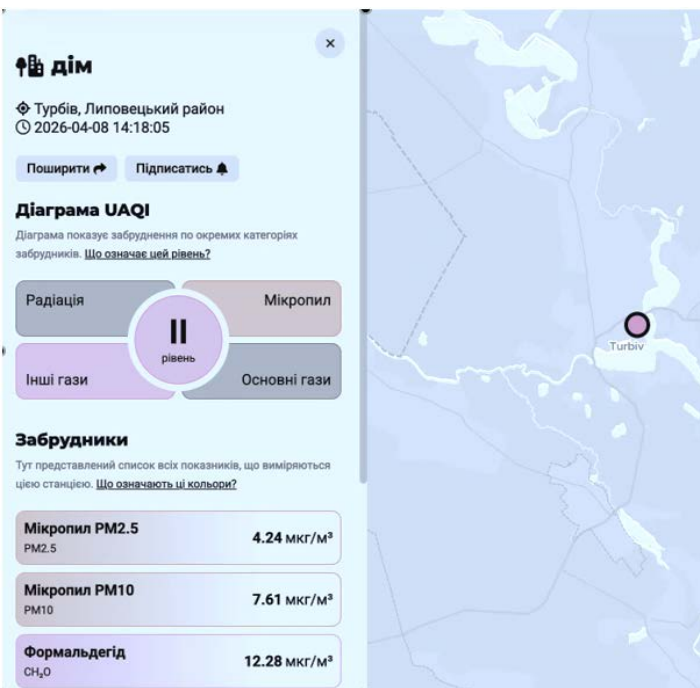


Рис. 3. Станція № 650 мережі EcoCity

Prophet навчається на повній тренувальній вибірці з параметрами сезонності, як це зазначалося раніше. Результат роботи моделі Prophet показано на рис. 4.

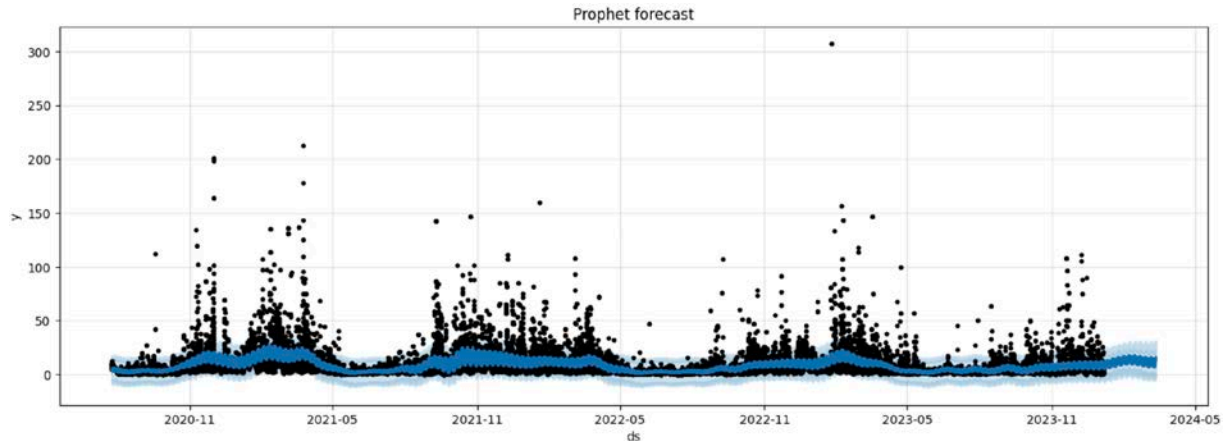


Рис. 4. Прогноз моделі Prophet на тренувальній вибірці

Для кількісної оцінки якості структурного прогнозу обчислено кореляцію між $\hat{y}_p(t)$ та реальними значеннями $y(t)$ на тестовій вибірці. Отримано значення $r = 0,077$, що свідчить про слабкий лінійний зв'язок, це пояснюється систематичним завищенням прогнозу Prophet у зимовий тестовий період через навчену на аномальних даних 2021 року сезонність. Водночас, прогноз Prophet виконує роль сезонного контексту для LightGBM, а не безпосередньо точного прогнозу, що обумовлює його ефективність у складі гібридної моделі навіть за низької абсолютної точності.

Етап 2. Формування матриці ознак та навчання LightGBM.

На основі тренувального ряду $y(t)$ та прогнозу Prophet $\hat{y}_p(t)$ сформовано матрицю ознак. Після видалення рядків з NaN-значеннями, що виникають через лагові та rolling-ознаки, тренувальна матриця має розмірність 30071×27 (рис. 5).

```
Features (27): ['lag_1', 'lag_2', 'lag_3', 'lag_6', 'lag_12', 'lag_24', 'lag_48', 'lag_168', 'roll_mean_6', 'roll_std_6', 'roll_mean_12', 'roll_std_12', 'roll_mean_24', 'roll_std_24', 'roll_mean_168', 'roll_std_168', 'prophet_yhat', 'sin_dow', 'cos_dow', 'is_weekend', 'sin_month', 'cos_month', 'hour', 'sin_hour', 'cos_hour', 'is_rush_hour', 'is_night']
X_train: (27064, 27) | X_val: (3007, 27)
```

Рис. 5. Склад та розмірність матриці ознак моделі LightGBM

Модель LightGBM навчається з early stopping (50 раундів) на валідаційній вибірці (останні 10 %

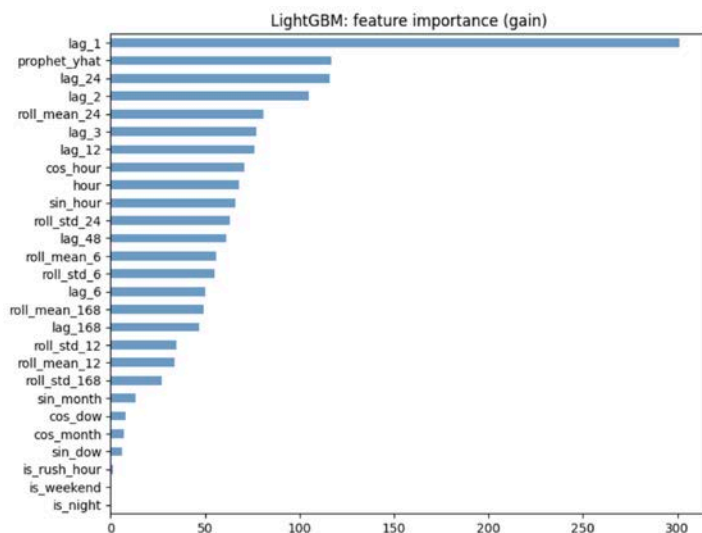


Рис. 6. Діаграма важливості ознак LightGBM

тренувального ряду). Розмір валідаційної вибірки становить 10 % тренувального ряду (близько 3330 годин або 139 діб), що дозволяє охопити декілька повних тижневих та місячних сезонних циклів і забезпечити стабільну оцінку узагальнювальної здатності моделі. Найкраща ітерація становить $N^* = 53$. Очікувано, найвищу важливість демонструють короткострокові лагові ознаки $y(t-1)$, $y(t-2)$, $y(t-3)$ та rolling-середні, що підтверджує сильну авторегресійну структуру ряду PM2.5. Ознака $\hat{y}_p(t)$ посідає помітне місце у рейтингу важливості, що свідчить про внесок структурного прогнозу Prophet у загалом

льну точність гібридної моделі. Графік важливості ознак подано на рис. 6.

Етап 3. Прогнозування та порівняння результатів.

Прогнозування на тестовому горизонті здійснюється рекурсивно. Для порівняння розраховано прогнози трьох моделей-базисів: ARIMA(2,1,2), Pure LightGBM (без ознаки $\hat{y}_p(t)$) та Prophet. Результати порівняльного аналізу показано на рис. 7—9.

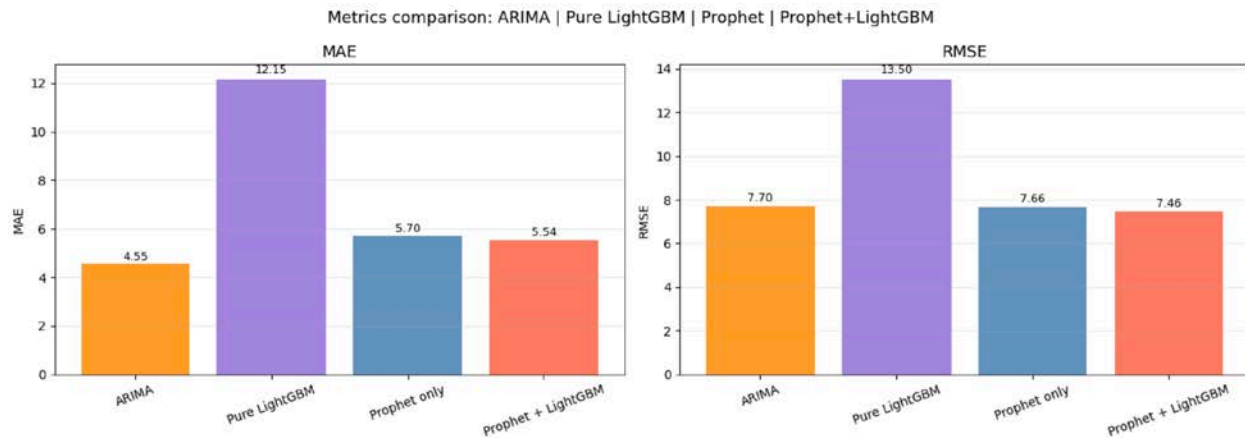


Рис. 7. Порівняння точності моделей за метриками MAE та RMSE

З аналізу рисунків випливає, що гібридна модель демонструє найкращі значення обох метрик серед усіх порівнюваних моделей. Загальна різниця в точності між гібридною моделлю та ARIMA є статистично незначущою за DM-тестом ($p = 0,090$), що пояснюється різною спеціалізацією моделей: ARIMA переважає на нормальних рівнях концентрації, тоді як гібридна модель показує кращі результати на пікових значеннях.

Model	MAE	RMSE	RMSE_improvement (%)
ARIMA	4.5520	7.7009	42.9757
Pure LightGBM	12.1512	13.5045	0.0000
Prophet only	5.7006	7.6645	43.2451
Prophet + LightGBM	5.5414	7.4639	44.7304

Рис. 8. Зведена таблиця метрик точності моделей

Hybrid vs Pure LGBM: stat=-36.2598, p=0.0000
 Hybrid vs ARIMA: stat=-1.6951, p=0.0903
 ARIMA vs Pure LGBM: stat=-22.5968, p=0.0000

Рис. 9. Результати критерію Діболда–Маріано для попарного порівняння моделей

На рис. 10 показано накопичену середню MAE по кроку горизонту прогнозу для всіх моделей.

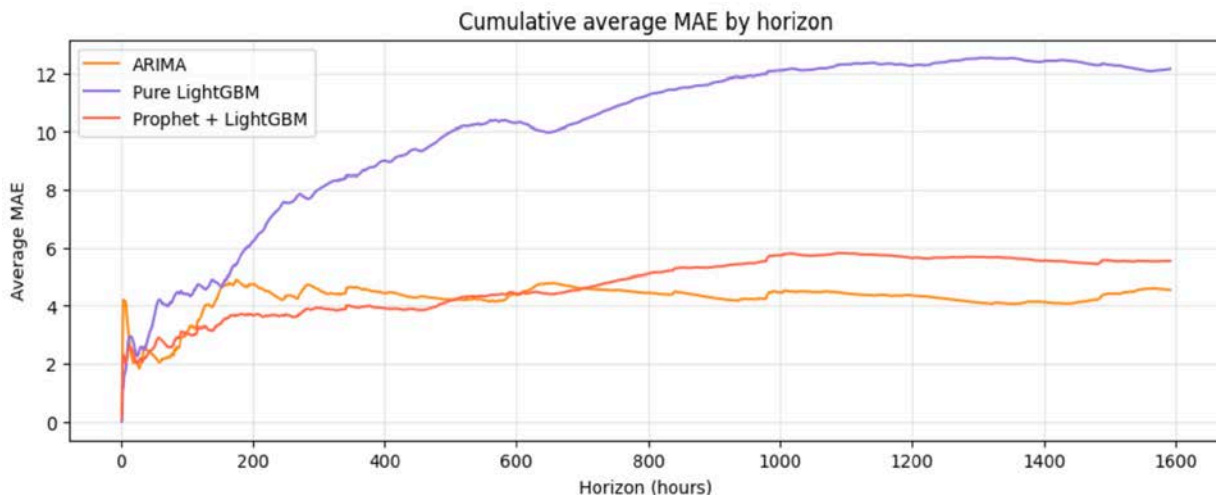


Рис. 10. Накопичена середня MAE по горизонту прогнозування

Спостерігається стрімке зростання помилки Pure LightGBM після ~100 годин прогнозу — накопичена MAE досягає 12 проти 5—6 для гібридної моделі та ARIMA. Це підтверджує, що включення детермінованої ознаки $\hat{y}_p(t)$ від Prophet зменшує швидкість накопичення рекурсивної помилки на довгих горизонтах. Хоча практичний інтерес найчастіше становлять прогнози на

горизонті 24...48 годин, аналіз довших горизонтів дозволяє оцінити стійкість моделі до накопичення рекурсивної похибки та її поведінку в умовах багаторазового використання власних прогнозів.

Для глибшого розуміння поведінки моделей проведено сегментний аналіз точності за двома режимами: нормальні концентрації (нижче 75-го перцентилу тренувального ряду) та пікові концентрації (вище 75-го перцентилу). Результати показано на рис. 11.

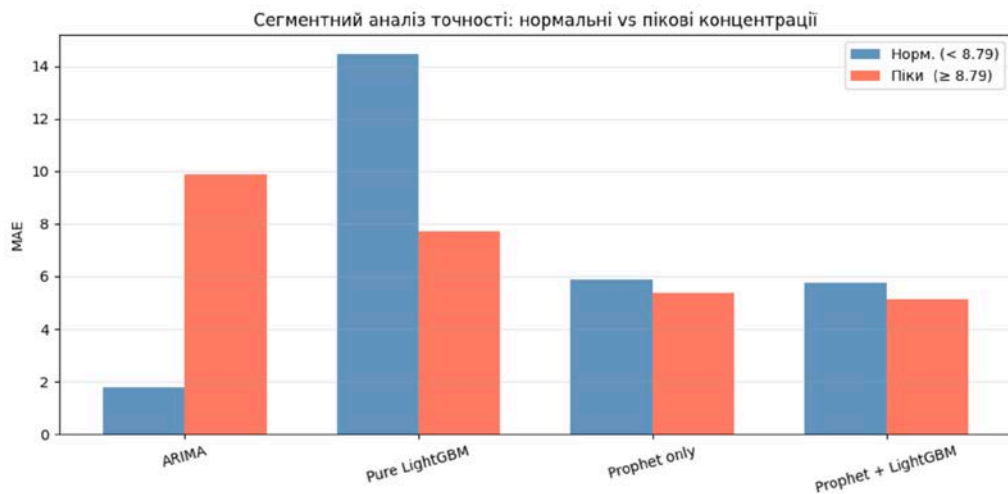


Рис. 11. Сегментний аналіз точності моделей

З рис. 11 випливає, що ARIMA демонструє найкращу точність на нормальних рівнях концентрації, проте суттєво поступається на пікових значеннях. Pure LightGBM задовільно відтворює піки, але суттєво втрачає точність на нормальних рівнях внаслідок накопичення рекурсивної помилки. Запропонована гібридна модель демонструє найзбалансованішу точність між нормальними та піковими режимами концентрації.

Висновки

Розроблено стекінгову гібридну архітектуру прогнозування погодинної концентрації PM_{2.5} на основі моделей Prophet та LightGBM, адаптовану для задач екологічного моніторингу з вираженими піковими концентраціями. На відміну від традиційного підходу корекції залишків, у запропонованому методі прогноз моделі Prophet використовується як детермінована сезонна ознака для алгоритму градієнтного бустингу LightGBM разом з авторегресійними лагами, ковзними статистиками та календарними ознаками, на основі яких LightGBM безпосередньо передбачає цільове значення концентрації. Такий підхід дозволяє LightGBM самостійно визначати міру довіри до структурного прогнозу в кожний момент часу та ефективно поєднувати переваги статистичного моделювання з нелінійною авторегресією.

Для тестування використовувались реальні дані погодинних вимірювань показника PM_{2.5} зі станції № 650 мережі громадського моніторингу якості атмосферного повітря EcoCity спільно з міжнародною програмою «Чисте повітря для України» за 2021—2024 роки.

Тестування показало, що запропонована гібридна модель забезпечує найкращі значення метрик точності серед усіх порівнюваних моделей (MAE = 5,54, RMSE = 7,46) та статистично значуще перевершує Pure LightGBM за критерієм Діболда-Маріано ($p < 0,001$, статистика $-36,26$). Різниця між гібридною моделлю та ARIMA в загальній точності статистично незначуща ($p = 0,090$), проте гібридна модель перевершує ARIMA за обома метриками з суттєво вищою стабільністю на довгих горизонтах прогнозу.

Аналіз деградації по горизонту прогнозування показав, що накопичена середня помилка Pure LightGBM монотонно зростає після ~100 годин рекурсивного прогнозу і досягає 12+, тоді як гібридна модель залишається стабільною протягом усього тестового горизонту (~1754 години). Це підтверджує, що включення детермінованої ознаки $\hat{y}_p(t)$ від Prophet зменшує швидкість накопичення рекурсивної помилки, оскільки сезонний контекст є відомим наперед для будь-якого майбутнього моменту і не залежить від точності попередніх кроків прогнозу.

До того ж сегментний аналіз точності виявив принципово різну спеціалізацію моделей: ARIMA демонструє найкращу точність на нормальних рівнях концентрації (MAE = 1,84), але суттєво поступається на пікових значеннях (MAE = 9,88 — у 1,9 рази гірше за гібрид); Pure LightGBM задовільно відтворює піки (MAE = 7,68), але демонструє суттєве зниження точності на нормальних рівнях (MAE = 14,55) внаслідок накопичення рекурсивної помилки. Запропонована гібридна модель продемонструвала найкращий результат роботи на пікових концентраціях у порівнянні з різними базовими моделями.

Отримані результати підтверджують доцільність використання запропонованої гібридної моделі в системах раннього попередження про небезпечні рівні забруднення атмосферного повітря, де точне прогнозування пікових концентрацій є критично важливим для захисту здоров'я населення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] S. J. Taylor, and B. Letham, "Forecasting at scale," *PeerJ Preprints*, vol. 5, e3190v2, 2017. <https://doi.org/10.7287/peerj.preprints.3190v2>.
- [2] Д. О. Шмундяк, «Дослідження впливу параметрів моделі Prophet на точність прогнозування якості атмосферного повітря,» у матеріали ЛП Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації, Вінниця, Україна, 2023. [Електронний ресурс]. Режим доступу: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18210/15056>.
- [3] V. I. Kontopoulou, A. D. Panagopoulos, I. Kakkos, and G. K. Matsopoulos, "A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks," *Future Internet*, vol. 15, no. 255, 2023. <https://doi.org/10.3390/fi15080255>.
- [4] О. Б. Мокін, В. Б. Мокін, і Б. І. Мокін, «Алгоритм методу ідентифікації моделі авторегресії — ковзного середнього, який узагальнює методику Юла–Уокера, та його програмна Python-реалізація,» *Вісник Вінницького політехнічного інституту*, № 4, с. 41-55, 2022. <https://doi.org/10.31649/1997-9266-2022-163-4-41-55>.
- [5] P. Lara-Benítez, M. Carranza-García, and J. C. Riquelme, "An Experimental Review on Deep Learning Architectures for Time Series Forecasting," *International Journal of Neural Systems*, vol. 31, no. 3, 2130001, 2021. <https://doi.org/10.1142/S0129065721300011>.
- [6] G. Ke, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 3146-3154. [Online]. Available: <https://proceedings.neurips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>.
- [7] J. Oblitas, and E. J. Escobedo Cárdenas, "Prediction of PM2.5 and PM10 Concentrations Using XGBoost and LightGBM Algorithms: A Case Study in Lima, Peru," *Interfases*, pp. 183-206, 2024. <https://doi.org/10.26439/interfases2024.n020.7417>.
- [8] Eco-City, «Громадський моніторинг стану якості повітря» [Електронний ресурс]. Режим доступу: <https://reborn.eco-city.org.ua/>.
- [9] F. A. Mohammed, and M. A. Mousa, "Applying Diebold–Mariano Test for Performance Evaluation Between Individual and Hybrid Time-Series Models for Modeling Bivariate Time-Series Data and Forecasting the Unemployment Rate in the USA," *Theory and Applications of Time Series Analysis. ITISE 2019, Contributions to Statistics*, O. Valenzuela, F. Rojas, L. J. Herrera, H. Pomares, and I. Rojas, Ed. Cham, Switzerland: Springer, 2020. https://doi.org/10.1007/978-3-030-56219-9_29.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Дата надходження 29.04.2026

Дата прийняття після доопрацювання 4.05.2026

Дата прийняття до друку після рецензування 12.05.2026

Дата публікації 2.10.2026

Шмундяк Дмитро Олександрович — д-р філософії, асистент кафедри системного аналізу та інформаційних технологій, e-mail: dmytro.shmundiak@vntu.edu.ua . <https://orcid.org/0009-0000-3299-7915>;

Варчук Ілона В'ячеславівна — канд. техн. наук, доцент, викладач кафедри системного аналізу та інформаційних технологій, e-mail: ilonavarchuk@gmail.com . <https://orcid.org/0009-0007-2317-0108>

D. O. Shmundiak¹I. V. Varchuk¹

Stacking Hybrid Model for Time Series Forecasting

¹Vinnytsia National Technical University

As environmental risks continue to grow, the task of forecasting pollution levels is becoming increasingly important. Traditional statistical models such as ARIMA demonstrate high accuracy at background pollution levels but lose forecast quality at peak concentrations that are critical from a public health perspective. Machine learning models, in particular gradient boosting, conversely, perform better at reproducing nonlinear peaks, but degrade over long forecast horizons due to recursive error accumulation. These limitations necessitate the development of hybrid approaches capable of combining the advantages of both model classes.

A hybrid stacking architecture for hourly time series forecasting has been developed and investigated, it combines the structural Prophet model and the LightGBM gradient boosting algorithm. Unlike the traditional residual correction approach, in the proposed architecture the Prophet forecast is used as a deterministic seasonal feature alongside autoregressive and calendar features, on the basis of which LightGBM directly predicts the target value.

The implementation and testing of the proposed model were carried out using real air quality monitoring data obtained from the EcoCity public air quality monitoring network. The statistical significance of the results was confirmed by the Diebold-Mariano test. Comparison with baseline models showed that the hybrid model statistically significantly outperforms the baseline models and demonstrates no critical accuracy failures across all concentration regimes. The results obtained confirm the feasibility of using the proposed model in early warning systems for hazardous air pollution levels.

Keywords: time series forecasting, air quality, PM2.5, hybrid model, Prophet, LightGBM, gradient boosting, stacking, machine learning, EcoCity.

Shmundiak Dmytro O. — PhD, Assistant of the Chair of System Analysis and Information Technologies, e-mail: dmytro.shmundiak@vntu.edu.ua . <https://orcid.org/0009-0000-3299-7915>;

Varchuk Ilona V. — Cand. Sc., Associate Professor, Lecturer of the Chair of System Analysis and Information Technologies, e-mail: ilonavarchuk@gmail.com . <https://orcid.org/0009-0007-2317-0108>