УДК 004.8

Dmytro Shevchenko[1], PhD, Assistant, Department of Computer Science
ORCID ID: https://orcid.org/0009-0001-7736-8263
e-mail: dimashevchenko10021999@gmail.com

Bella Holub[1], Candidate of Engineering Sciences, Associate Professor, Head of the Department of Computer Science
ORCID ID: https://orcid.org/0000-0002-1256-6138
e-mail: bellalg@nubip.edu.ua

Taras Trysnyuk[2], Candidate o f Technical Sciences, Senior Researcher
ORCID ID: https://orcid.org/0000-0002-3672-8242
e-mail: taras24t@gmail.com

[1]National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine
[2]Institute of Telecommunications and Global Information Space of NAS of Ukraine, Kyiv, Ukraine

RESTORATION OF MISSING ENVIRONMENTAL DATA IN AN AIR QUALITY MONITORING SYSTEM BASED ON A NAIVE BAYES CLASSIFIER

Abstract. *The article examines an approach to restoring missing environmental data in an air quality monitoring system based on a Naive Bayes classifier. The relevance of the study is determined by the fact that missing values in observational time series reduce the reliability of air quality index calculations, complicate the interpretation of environmental conditions, and weaken the analytical support of managerial decision-making. The study is a logical continuation of previous research in which stable and representative monitoring stations suitable for forming a high-quality training dataset were identified using cluster analysis. In contrast to approaches that use the entire set of available measurements without considering their reliability, the proposed method involves training the model only on data from selected stations characterized by higher completeness, stability, and credibility of time series.*

The study forms a feature space based on concentrations of major pollutants and accompanying meteorological parameters, uses CAQI categories as the target variable, and implements a procedure for restoring missing values according to the most probable air quality class. The obtained results confirmed the acceptable quality of the constructed model: the overall classification accuracy reached 0.71, which indicates the suitability of the approach for basic air quality assessment and its further use in intelligent data restoration tasks.

The practical value of the proposed approach lies in its potential integration into environmental information and analytical systems in order to improve data completeness, enhance the quality of air quality index calculations, and provide more reliable analytical support for decision-making.

Keywords: *Data Mining, K-Means, data restoration, information and analytical system, intelligent technology, information technologies, data quality, data reliability and validity.*

Д.В. Шевченко[1], Б.Л. Голуб[1], Т.В. Триснюк[2]

[1]Національний університет біоресурсів і природокористування України, м. Київ, Україна
[2]Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ, Україна

ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ЕКОЛОГІЧНИХ ДАНИХ У СИСТЕМІ МОНІТОРИНГУ ЯКОСТІ АТМОСФЕРНОГО ПОВІТРЯ НА ОСНОВІ НАЇВНОГО БАЙЄСІВСЬКОГО КЛАСИФІКАТОРА

Анотація. *У статті розглянуто підхід до відновлення пропущених екологічних даних у системі моніторингу якості атмосферного повітря на основі наївного байєсівського класифікатора. Актуальність дослідження зумовлена тим, що пропуски в часових рядах спостережень знижують достовірність розрахунку індексів якості повітря, ускладнюють інтерпретацію екологічної ситуації та погіршують аналітичну підтримку управлінських рішень. Дослідження є логічним продовженням попередньої роботи, у якій за допомогою кластерного аналізу було виявлено стабільні та репрезентативні станції моніторингу, придатні для формування якісної навчальної вибірки. На відміну від підходів, що використовують усю сукупність доступних вимірювань без урахування їх надійності, запропонований метод передбачає навчання моделі лише на даних відібраних станцій із вищим рівнем повноти, стабільності та достовірності часових рядів.*

У роботі сформовано ознаковий простір на основі концентрацій основних забруднювачів і супровідних метеорологічних параметрів, використано категорії індексу CAQI як цільову змінну та реалізовано процедуру відновлення пропущених значень за найбільш імовірним класом стану повітря. Отримані результати підтвердили прийнятну якість побудованої моделі: загальна точність класифікації становила 0.71, що свідчить про придатність підходу для базового аналізу стану атмосферного повітря та подальшого використання у задачах інтелектуального відновлення даних.

Практична цінність запропонованого підходу полягає у можливості його інтеграції в інформаційно-аналітичні системи екологічного моніторингу для підвищення повноти, узгодженості та достовірності даних. Його застосування дає змогу покращити якість розрахунку індексів стану атмосферного повітря, зменшити вплив пропусків і спотворень у часових рядах на результати аналітики, а також забезпечити більш надійну інформаційну основу для підтримки управлінських рішень у сфері екологічного контролю та оцінювання стану довкілля.

Ключові слова: *Data Mining, Naive Bayes, відновлення даних, інформаційно-аналітична система, інтелектуальна технологія, інформаційні технології, якість даних, надійність та достовірність даних.*

https://doi.org/10.32347/2411-4049.2026.2.262-273

Вступ

Відновлення пропущених екологічних даних у системах моніторингу якості атмосферного повітря є не другорядною технічною процедурою, а необхідною умовою коректного розрахунку індексів якості повітря, надійного інформування населення та обґрунтованої підтримки управлінських рішень. Як логічне продовження попередньої роботи авторів [1], у якій кластеризацією

станцій було виділено стабільні та проблемні канали вимірювання, поточне дослідження закономірно зміщує акцент із виявлення нестабільності на відновлення пропусків саме на основі надійних станцій. За оцінками Всесвітньої організації охорони здоров'я, тягар захворюваності, пов’язаний із забрудненням повітря, є співмірним з іншими ключовими глобальними факторами ризику [2], а за даними Європейського агентства з довкілля, 96% міського населення ЄС усе ще зазнає впливу небезпечних концентрацій $PM_{2.5}$ [3]. Водночас індекси якості повітря використовуються для комунікації короткострокових концентрацій забруднювачів і пов’язаних із ними ризиків для здоров’я [4], тому пропуски в часових рядах безпосередньо знижують достовірність індексної оцінки та можуть спричиняти викривлення у подальшій аналітиці. Проблема посилюється тим, що якість даних сенсорних мереж є високоваріативною, а без стандартизованої оцінки продуктивності складно зіставляти сенсорні вимірювання з еталонним моніторингом [5].

У такій постановці методологічно виправданим є підхід, орієнтований на якість даних: модель відновлення доцільно навчати не на всіх доступних вимірюваннях, а на попередньо відібраних станціях із вищою повнотою, стабільністю та достовірністю часових рядів. Європейська практика аналізу повітряних спостережень показує, що перевірка повноти та відсів занадто коротких або неповних рядів підвищують якість підсумкової аналітики, навіть якщо це зменшує обсяг вибірки [6]. Сучасні фреймворки для мереж низьковартісних сенсорів також розглядають імпутацію не ізольовано, а як частину ширшого конвеєра забезпечення якості даних – поряд із калібруванням, згладжуванням, виявленням аномалій та мережевою обробкою [7]. Саме тому для поставленої задачі доцільним є поєднання попереднього відбору надійних станцій із наївним байєсівським класифікатором як легкою, інтерпретованою та придатною до інтеграції моделлю, у якій відновлення відсутніх значень виконується через визначення найбільш імовірного класу стану повітря.

Аналіз останніх досліджень і публікацій. Огляд сучасних праць свідчить, що імпутація в екологічних часових рядах розвивається від простих статистичних процедур до просторово-часових і гібридних моделей, однак універсально найкращого методу для всіх типів пропусків і всіх конфігурацій моніторингових мереж досі не встановлено. Для короткочасних прогалин у даних забруднювачів показано, що одновимірні методи можуть бути практичним рішенням, тоді як багатовимірні підходи не завжди дають кращий результат [8]. Водночас порівняльні дослідження mean, median, kNNI, MICE, SAITS, BRITS, MRNN і Transformer підтверджують, що якість імпутації безпосередньо впливає на подальші задачі класифікації та прогнозування, а перевага конкретного методу залежить від структури даних і частки пропусків [9]. На великомасштабних наборах екологічного моніторингу також встановлено, що оптимальний імпутер має dataset-specific характер, хоча методи, що враховують просторові залежності, часто перевищують базові підстановки [10]. Для $PM_{2.5}$-вимірювань окремо показано стабільність KNN на різних часових інтервалах пропусків.

Другий виразний напрям – моделі, що поєднують часові, просторові та гібридні залежності для складніших сценаріїв, насамперед блокових або довгих пропусків. У дослідженні Wang et al. BRITS-ALSTM перевищив Mean, KNN, MF, MICE, M-RNN, BRITS і BRITS-LSTM для шести забруднювачів, що

автори пов’язують із використанням двоспрямованого моделювання, структури та механізму уваги [11]. Подальший розвиток цього напряму демонструє робота Lee et al., де GA-BiLSTM у міській мережі Чикаго перевершив XGBoost і KNN, особливо для тривалих відключень до десяти днів [12]. Для довгих послідовних прогалин у даних атмосферного забруднення також запропоновано гібридний decomposition-based підхід із transfer learning, який статистично перевищив низку альтернатив за MAE та MAPE [13].

На цьому тлі Наївний Байєсівський підхід займає окрему нішу. Література з прямої Naive Bayes імпутації істотно менша за корпус досліджень з BRITS, LSTM чи Transformer, проте вона наявна й демонструє життєздатність ймовірнісної логіки для неповних даних. Так, Khotimah et al. ще у 2019 р. показали, що Naïve Bayes Imputation перевищує прості підстановки mean/mode в задачах класифікації з пропусками [14]. Щодо саме моніторингу повітря, у сучасних порівняльних роботах Naive Bayes розглядається як повноцінна класична базова модель. Imam et al. на AQI-категоріях відзначають високу результативність Random Forest і Naive Bayes серед традиційних алгоритмів, а в роботі Natarajan et al. Gaussian Naive Bayes у конкретному порівнянні для прогнозування AQI навіть перевершив інші популярні моделі. Отже, мова про його практичну цінність як легкої, швидкої та інтерпретованої базової моделі, особливо там, де потрібна інтеграція в реальний інформаційно-аналітичний контур і відсутні надвеликі навчальні масиви.

Окремий блок останніх публікацій стосується не безпосередньо імпутації, а формування надійної навчальної основи для неї. Агентство з охорони довкілля США прямо зазначає, що якість сенсорних даних є високоваріативною, а без узгоджених протоколів тестування складно зрозуміти, наскільки такі дані придатні для порівняння з регуляторним моніторингом. У європейських оцінках перевірка повноти використовується як формальний механізм відсіву коротких і неповних рядів перед трендовим аналізом. За нашим узагальненням, у більшості нових праць імпутація оптимізується вже після етапу очищення й контролю якості, проте формалізований зв’язок між попереднім відбором надійних станцій і подальшим навчанням легкої ймовірнісної моделі висвітлено значно слабше. Саме цю прогалину і заповнює запропонована постановка: використати результати попередньої кластеризації станцій як підґрунтя для наївно-байєсівського відновлення пропущених екологічних даних.

Метою статті є розроблення та дослідження підходу до відновлення пропущених екологічних даних у системі моніторингу якості атмосферного повітря на основі наївного байєсівського класифікатора з використанням попередньо відібраних надійних станцій моніторингу для формування якісної навчальної вибірки.

Виклад основного матеріалу дослідження

У межах дослідження розглянуто вхідні дані системи моніторингу якості атмосферного повітря, підходи до формування якісної навчальної вибірки та методи інтелектуального аналізу даних, використані для відновлення пропущених екологічних значень.

Вхідні дані та якість навчальної вибірки. У дослідженні використано екологічні дані системи моніторингу якості атмосферного повітря, що містять концентрації основних забруднювачів та супровідні метеорологічні параметри. До ознакового простору включено показники $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO, а також температуру повітря, відносну вологість і швидкість вітру. Для подальшого аналізу дані було агреговано у формат, придатний для побудови класифікаційної моделі: кожен запис відповідає визначеному часовому інтервалу спостереження, а стовпці містять значення екологічних і метеорологічних параметрів.

Оскільки якість навчальної вибірки безпосередньо впливає на точність моделі відновлення пропущених даних, у роботі використано попередньо відібрані станції моніторингу, які за результатами кластерного аналізу були визначені як найбільш стабільні та репрезентативні. Такий підхід дозволяє зменшити вплив шуму, технічних збоїв, аномальних значень і нерепрезентативних часових рядів на процес навчання. Отже, навчальна вибірка формувалася не з усієї множини доступних станцій, а лише з тих, що характеризуються вищим рівнем повноти, стабільності та достовірності даних.

Наївний Байєс (Naive Bayes). Методи Data Mining є важливими інструментами для аналізу великих масивів даних у системах моніторингу атмосферного повітря. Вони дають змогу не лише виявляти приховані закономірності у часових рядах спостережень, а й будувати моделі, що пояснюють формування забруднення та підтримують подальше прогнозування. На відміну від традиційних статистичних підходів, Data Mining орієнтований на пошук нетривіальних залежностей і автоматизовану обробку багатовимірних даних, що є особливо важливим для екологічних систем, де вимірювання надходять безперервно, мають різний рівень повноти та можуть містити технічні викривлення.

Класифікатор належить до методів, які ґрунтуються на теоремі Байєса та припущенні незалежності ознак [15]. У версії Gaussian Naive Bayes робиться додаткове припущення, що числові параметри розподілені за нормальним законом, що добре узгоджується з природою багатьох екологічних змінних параметрів. У контексті моніторингу атмосферного повітря багато параметрів, що вимірюються, часто демонструють близькі до нормального розподіли, особливо коли розглядаються усереднені добові значення.

Важливо зазначити, що навіть якщо розподіл не є ідеально нормальним, алгоритм зберігає прийнятну точність. Саме тому він широко застосовується як базова модель (baseline) у дослідженнях: швидкий у реалізації, обчислювально ефективний і достатньо надійний для порівняння з більш складними методами. Модель використовує наступне формальне представлення (1):

$$P(C_k \mid x) = \frac{P(C_k)\,\Pi_{i=1}^{n}\,P(x_i \mid C_k)}{P(x)}, \tag{1}$$

де:

- C_k – клас (наприклад, категорія якості повітря),
- $x = (x_1, \ldots, x_n)$ – вектор ознак (екологічні параметри),
- $P(x_i \mid C_k)$ – ймовірність значення ознаки за умови належності до класу, яка апроксимується гаусівським розподілом (2):

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\delta_k^2}} exp\left(-\frac{(x_i - \mu_k)^2}{\sqrt{2\pi\delta_k^2}}\right), \quad (2)$$

де:

- μ_k – середнє значення,
- δ_k – стандартне відхилення ознаки в класі C_k.

Цей алгоритм є доцільним для класифікації поточного стану атмосфери за багатьма параметрами одночасно. Як цільову змінну можна брати категорію індексу якості повітря, а як предиктори – концентрації забруднювачів ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO) та метеорологічні фактори (температура, вологість, швидкість вітру). Завдяки цьому модель здатна оперативно класифікувати ситуацію (якість повітря).

Результати дослідження та їх обговорення

На першому етапі аналізу побудовано теплову карту статистичних характеристик екологічних показників у розрізі категорій якості повітря за індексом CAQI. Візуалізація відображає мінімальні, максимальні та середні значення ключових параметрів (рис. 1). Такий підхід дозволяє виявити загальні закономірності зміни показників залежно від переходу між класами якості повітря.

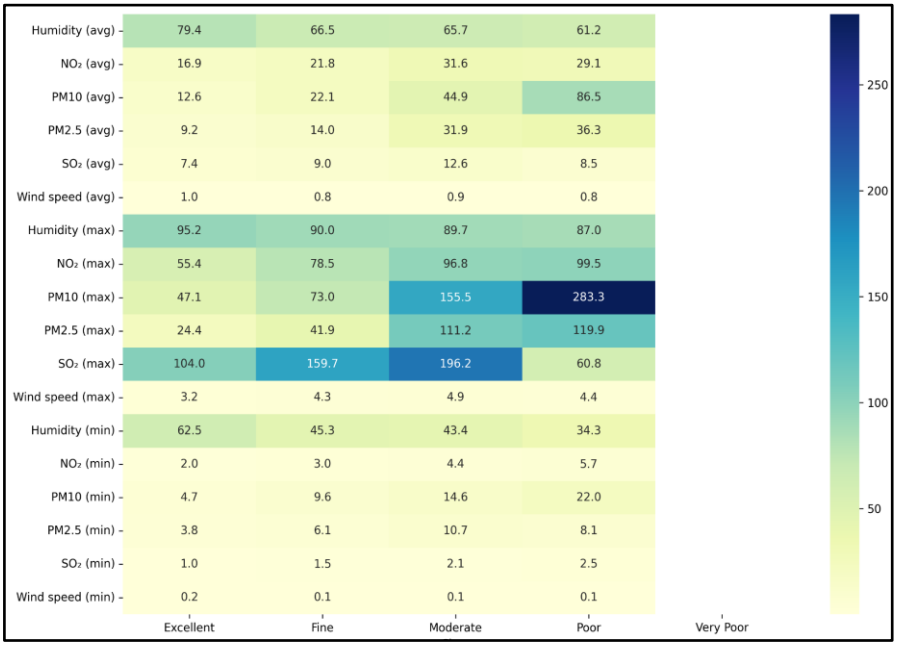

Рисунок 1. Теплова карта статистичних характеристик екологічних показників

За результатами аналізу встановлено чітку тенденцію до зростання концентрацій твердих часток зі збільшенням рівня забруднення. Зокрема, значення PM_{10} зростають від приблизно 47 у категорії «Excellent» до понад 280 у категоріях «Poor/Very Poor», а $PM_{2.5}$ – до близько 220 у найгірших класах. Для SO_2 зафіксовано локальний сплеск у категорії «Moderate», що може свідчити про епізодичні викиди. Водночас вологість повітря має тенденцію до

зниження зі зростанням рівня забруднення, тоді як швидкість вітру не демонструє сталої залежності, хоча в окремих випадках спостерігаються короткочасні підвищення, пов'язані з процесами розсіювання домішок.

Отримані залежності підтверджують, що використані ознаки є інформативними для класифікації стану атмосферного повітря та можуть слугувати основою для побудови моделі відновлення пропущених значень.

Матриця помилок. Перед аналізом результатів класифікації доцільно розглянути розподіл даних за класами цільової змінної, оскільки він суттєво впливає на ефективність будь-якої моделі. На рисунку 2 подано візуалізацію структури даних, що демонструє нерівномірність представлення різних категорій індексу якості повітря (CAQI).

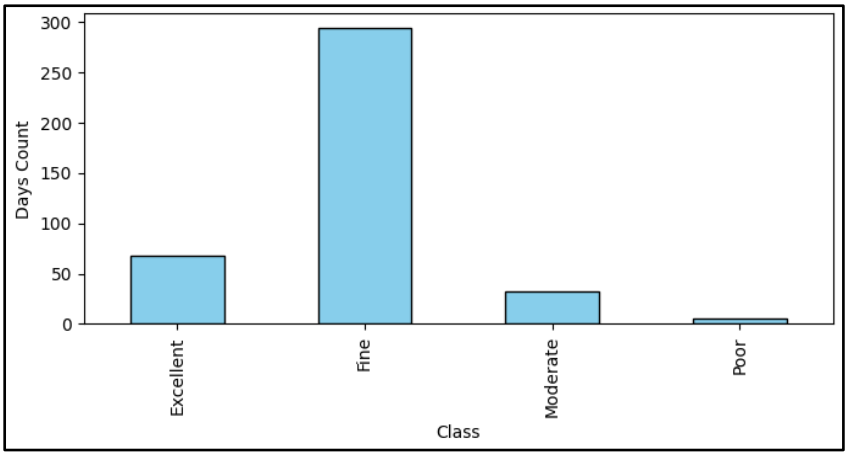

Рисунок 2. Розподіл кількості даних по класах цільової змінної

Помітно, що окремі класи мають значно більшу кількість спостережень, тоді як інші представлені обмежено. Такий дисбаланс класів призводить до того, що модель краще розпізнає більш численні категорії, забезпечуючи для них вищі значення точності. Це необхідно враховувати під час інтерпретації подальших результатів класифікації та планування можливих методів балансування вибірки у майбутніх дослідженнях.

Для глибшого аналізу результатів побудовано матрицю помилок, яка демонструє розподіл правильних і помилкових передбачень між реальними та прогнозованими класами. Найкраще модель розпізнає клас «Fine», для якого зафіксовано найбільшу кількість правильних класифікацій. Частина спостережень цього класу була віднесена до категорії «Excellent», що свідчить про часткове перекриття сусідніх класів. Аналогічна зворотна тенденція спостерігається і для класу «Excellent». Класи «Moderate» та «Poor» також демонструють прийнятну якість розпізнавання, однак для останнього слід враховувати малу кількість прикладів у вибірці, що обмежує стійкість оцінювання.

Таким чином, матриця помилок підтверджує, що основні помилки моделі зосереджені переважно між близькими за змістом категоріями якості повітря, що є логічним з огляду на подібність їхніх граничних значень і часткове перекриття ознак.

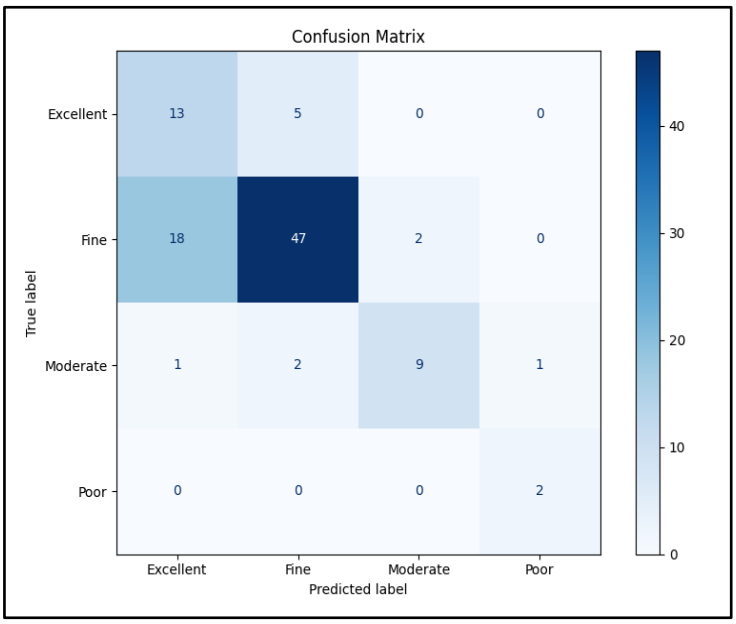

Рисунок 3. Матриця помилок

Метрики класифікаційної моделі. Після аналізу розподілу прикладів доцільно перейти до оцінки якості самої моделі. Таблиця 1 містить основні метрики класифікаційної моделі, побудованої на основі алгоритму Наївного Байєса. Показники точності (precision), повноти (recall) та F1-міри (середнє між precision і recall) для кожного класу дозволяють оцінити, наскільки добре модель розрізняє різні рівні якості повітря.

Таблиця 1. Метрики класифікаційної моделі

Клас	**Precision**	**Recall**	**F1-score**	**Support**
Excellent	0.41	0.72	0.52	18
Fine	0.87	0.70	0.78	67
Moderate	0.82	0.69	0.75	13
Poor	0.67	1.00	0.80	2
	–	–	–	–
Accuracy	–	–	0.71	100
Macro avg	0.69	0.78	0.71	100
Weighted avg	0.78	0.71	0.73	100

Загальна точність моделі становить 0.71, тобто правильно класифіковано близько 71% об’єктів тестової вибірки. Середні значення метрик також залишаються прийнятними: macro avg дорівнює 0.71, a weighted avg – 0.73. Це свідчить про загалом збалансовану ефективність моделі, навіть попри наявність дисбалансу між класами.

Отримані результати підтверджують, що наївний байєсівський класифікатор є придатною базовою моделлю для класифікації стану атмосферного повітря та може бути використаний як основа для подальшого відновлення пропущених екологічних даних.

Висновки

У статті запропоновано підхід до відновлення пропущених екологічних даних у системі моніторингу якості атмосферного повітря на основі наївного байєсівського класифікатора. Особливістю підходу є формування моделі на підмножині попередньо відібраних надійних станцій моніторингу, що дало змогу зменшити вплив нестабільних, неповних і спотворених часових рядів на якість навчальної вибірки. У роботі обґрунтовано доцільність застосування методів Data Mining для обробки екологічних часових рядів, сформовано набір інформативних ознак на основі концентрацій основних забруднювачів і метеорологічних параметрів та використано категорії індексу CAQI як цільову змінну для класифікації стану атмосферного повітря. Отримані результати підтвердили прийнятну якість моделі, загальна точність якої становить 0,71 відсотка.

Практична цінність запропонованого підходу полягає у можливості його інтеграції в інформаційно-аналітичні системи екологічного моніторингу для підвищення повноти, узгодженості та достовірності даних. Використання моделі дає змогу не лише класифікувати поточний стан атмосфери, а й відновлювати відсутні або спотворені показники на основі найбільш імовірного класу якості повітря, що покращує подальші розрахунки екологічних індексів і підтримку прийняття рішень. Перспективи подальших досліджень пов’язані з порівнянням запропонованого підходу з іншими методами імпутації, аналізом різних механізмів виникнення пропусків, перевіркою стійкості методу в різних сезонних умовах і просторових конфігураціях мережі, а також із розробленням гібридних моделей відновлення даних для складніших сценаріїв моніторингу.

СПИСОК ЛІТЕРАТУРИ

1. Шевченко, Д. В., Голуб, Б. Л., Бородкіна, І. Л. (2026). Кластеризація станцій для виявлення нестабільності даних у мережі моніторингу якості атмосферного повітря. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка»*, 4 (32), 1054–1064.

2. World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. WHO. https://www.who.int/publications/i/item/9789240034228

3. European Environment Agency. (2024). *Europe’s air quality status 2024* (Briefing No. 06/2024). Publications Office of the European Union. https://doi.org/10.2800/5970

4. World Health Organization. (2026). *Air quality indexes: Key considerations and roadmaps for best practices*. WHO. https://www.who.int/publications/i/item/9789289062701
5. U.S. Environmental Protection Agency. (2026). *Air sensor performance targets and testing protocols*. EPA. https://www.epa.gov/air-sensor-toolbox/air-sensor-performance-targets-and-testing-protocols
6. Gbangou, T., Colette, A., Soares, J., & González Ortiz, A. (2023). *ETC HE report 2023/8: Long-term trends of air pollutants at European and national level 2005–2021*. European Topic Centre on Human Health and the Environment. https://www.eionet.europa.eu/etcs/etc-he/products/etc-he-products/etc-he-reports/etc-he-report-2023-8-long-term-trends-of-air-pollutants-at-european-and-national-level-2005-2021
7. Ferrer-Cid, P., Paredes Ahumada, J. A., Allka, X., Guerrero Zapata, M., Barceló Ordinas, J. M., & García Vidal, J. (2024). A data-driven framework for air quality sensor networks. *IEEE Internet of Things Magazine, 7*(1), 128–134. https://doi.org/10.1109/IOTM.001.2300112
8. Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment, 730*, Article 139140. https://doi.org/10.1016/j.scitotenv.2020.139140
9. Hua, V., Nguyen, T., Dao, M.-S., Nguyen, H. D., & Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLoS ONE, 19*(9), Article e0306303. https://doi.org/10.1371/journal.pone.0306303
10. Decorte, T., Mortier, S., Lembrechts, J. J., Meysman, F. J. R., Latré, S., Mannens, E., & Verdonck, T. (2024). Missing value imputation of wireless sensor data for environmental monitoring. *Sensors, 24*(8), Article 2416. https://doi.org/10.3390/s24082416
11. Wang, Y., Liu, K., He, Y., Fu, Q., Luo, W., Li, W., Liu, X., Wang, P., & Xiao, S. (2023). Research on missing value imputation to improve the validity of air quality data evaluation on the Qinghai-Tibetan Plateau. *Atmosphere, 14*(12), Article 1821. https://doi.org/10.3390/atmos14121821
12. Lee, J., Berkelhammer, M., O'Brien, J., McNicol, G., Vincent, A. E. S., Grover, M., Packman, A. I., Kaludi, B., Cho, A., & Gonzalez-Meler, M. (2026). Imputation of urban environmental sensor data using gated attention bidirectional long short-term memory (GA-BiLSTM): Methods, performance, and implications. *Environmental Monitoring and Assessment, 198*, Article 262. https://doi.org/10.1007/s10661-026-15112-8
13. Wei, X., Meng, H., Shao, L., Fu, D., Ma, L., & Zhang, D. (2025). A decomposition-based imputation algorithm for long consecutive missing atmospheric pollution data and its application. *Journal of Computational Science, 92*, Article 102697. https://doi.org/10.1016/j.jocs.2025.102697
14. Khotimah, B. K., Miswanto, & Suprajitno, H. (2019). Modeling naïve Bayes imputation classification for missing data. *IOP Conference Series: Earth and Environmental Science, 243*(1), Article 012111. https://doi.org/10.1088/1755-1315/243/1/012111
15. *Naive Bayes classifier with Scikit-learn tutorial.* (n.d.). DataCamp. https://www.datacamp.com/tutorial/naive-bayes-scikit-learn
16. Шевченко, Д. В., Голуб, Б. Л. (2025). Моніторинг якості повітря в реальному часі. *Науковий журнал «Математичні машини і системи»*, 1, 103–112.
17. Trysnyuk, T., Trysnyuk, V., Okhariev, V., & Shumeiko, A. (2018). Cartographic model of Dniester River basin probable flooding. *Series D: Geology and Environmental Engineering, 32*(1), 51–55. https://doi.org/10.37193/SBSD.2018.1.07
18. Zaitsev, S., Vasylenko, V., Trysnyuk, V., & Trysnyuk, T. (2023). Adaptive method for assessing information reliability under uncertainty for 5G and IoT systems. In *Proceedings of the 3rd International Workshop on Information Technologies: Theoretical and Applied Problems* (CEUR Workshop Proceedings). CEUR-WS. https://ceur-ws.org/Vol-3628/paper2.pdf

19. Трофимчук, О. М., Триснюк, В. М., Анпілова, Є. С., Бутенко, О. С., Вишняков, В. Ю., Загородня, С. А., Клименко, В. І., Красовська, І. Г., Крета, Д. Л., Миронцов, М. Л., Охарєв, В. О., Попова, М. А., Радчук, І. В., Триснюк, Т. В., Шевякіна, Н. А., Шумейко, В. О. (2022). *Геоінформаційні дослідження водних екосистем України: моніторинг та прогнозування*. Івано-Франківськ: Супрун В. П. ISBN 978-617-7468-53-9.

Стаття надійшла до редакції 13.01.2026, надійшла після рецензування 23.02.2026, прийнята 11.03.2026

REFERENCES

1. Shevchenko, D. V., Holub, B. L., & Borodkina, I. L. (2026). Station clustering for detecting data instability in an air quality monitoring network. *Cybersecurity: Education, Science, Technique, 4*(32), 1054–1064.
2. World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. WHO. https://www.who.int/publications/i/item/9789240034228
3. European Environment Agency. (2024). *Europe's air quality status 2024* (Briefing No. 06/2024). Publications Office of the European Union. https://doi.org/10.2800/5970
4. World Health Organization. (2026). *Air quality indexes: Key considerations and roadmaps for best practices*. WHO. https://www.who.int/publications/i/item/9789289062701
5. U.S. Environmental Protection Agency. (2026). *Air sensor performance targets and testing protocols*. EPA. https://www.epa.gov/air-sensor-toolbox/air-sensor-performance-targets-and-testing-protocols
6. Gbangou, T., Colette, A., Soares, J., & González Ortiz, A. (2023). *ETC HE report 2023/8: Long-term trends of air pollutants at European and national level 2005–2021*. European Topic Centre on Human Health and the Environment. https://www.eionet.europa.eu/etcs/etc-he/products/etc-he-products/etc-he-reports/etc-he-report-2023-8-long-term-trends-of-air-pollutants-at-european-and-national-level-2005-2021
7. Ferrer-Cid, P., Paredes Ahumada, J. A., Allka, X., Guerrero Zapata, M., Barceló Ordinas, J. M., & García Vidal, J. (2024). A data-driven framework for air quality sensor networks. *IEEE Internet of Things Magazine, 7*(1), 128–134. https://doi.org/10.1109/IOTM.001.2300112
8. Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment, 730*, Article 139140. https://doi.org/10.1016/j.scitotenv.2020.139140
9. Hua, V., Nguyen, T., Dao, M.-S., Nguyen, H. D., & Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLoS ONE, 19*(9), Article e0306303. https://doi.org/10.1371/journal.pone.0306303
10. Decorte, T., Mortier, S., Lembrechts, J. J., Meysman, F. J. R., Latré, S., Mannens, E., & Verdonck, T. (2024). Missing value imputation of wireless sensor data for environmental monitoring. *Sensors, 24*(8), Article 2416. https://doi.org/10.3390/s24082416
11. Wang, Y., Liu, K., He, Y., Fu, Q., Luo, W., Li, W., Liu, X., Wang, P., & Xiao, S. (2023). Research on missing value imputation to improve the validity of air quality data evaluation on the Qinghai-Tibetan Plateau. *Atmosphere, 14*(12), Article 1821. https://doi.org/10.3390/atmos14121821
12. Lee, J., Berkelhammer, M., O'Brien, J., McNicol, G., Vincent, A. E. S., Grover, M., Packman, A. I., Kaludi, B., Cho, A., & Gonzalez-Meler, M. (2026). Imputation of urban environmental sensor data using gated attention bidirectional long short-term memory (GA-BiLSTM): Methods, performance, and implications. *Environmental Monitoring and Assessment, 198*, Article 262. https://doi.org/10.1007/s10661-026-15112-8

13. Wei, X., Meng, H., Shao, L., Fu, D., Ma, L., & Zhang, D. (2025). A decomposition-based imputation algorithm for long consecutive missing atmospheric pollution data and its application. *Journal of Computational Science, 92*, Article 102697. https://doi.org/10.1016/j.jocs.2025.102697
14. Khotimah, B. K., Miswanto, & Suprajitno, H. (2019). Modeling naïve Bayes imputation classification for missing data. *IOP Conference Series: Earth and Environmental Science, 243*(1), Article 012111. https://doi.org/10.1088/1755-1315/243/1/012111
15. *Naive Bayes classifier with Scikit-learn tutorial.* (n.d.). DataCamp. https://www.datacamp.com/tutorial/naive-bayes-scikit-learn
16. Shevchenko, D. V., & Holub, B. L. (2025). *Monitorynh yakosti povitria v realnomu chasi* [Air quality monitoring in real time]. *Matematychni mashyny i systemy*, (1), 103–112.
17. Trysnyuk, T., Trysnyuk, V., Okhariev, V., & Shumeiko, A. (2018). Cartographic model of Dniester River basin probable flooding. *Series D: Geology and Environmental Engineering, 32*(1), 51–55. https://doi.org/10.37193/SBSD.2018.1.07
18. Zaitsev, S., Vasylenko, V., Trysnyuk, V., & Trysnyuk, T. (2023). Adaptive method for assessing information reliability under uncertainty for 5G and IoT systems. In *Proceedings of the 3rd International Workshop on Information Technologies: Theoretical and Applied Problems* (CEUR Workshop Proceedings). CEUR-WS. https://ceur-ws.org/Vol-3628/paper2.pdf
19. Trofymchuk, O. M., Trysnyuk, V. M., Anpilova, E. S., Butenko, O. S., Vyshnyakov, V. Yu., Zagorodnya, S. A., Klymenko, V. I., Krasovska, I. G., Kreta, D. L., Myrontsov, M. L., Okharev, V. O., Popova, M. A., Radchuk, I. V., Trysnyuk, T. V., Shevyakina, N. A., & Shumeiko, V. O. (2022). *Geoinformation studies of aquatic ecosystems of Ukraine: Monitoring and forecasting*. Suprun V. P. Publishers.

The article was received 13.01.2026, received after revision 23.02.2026, accepted 11.03.2026

Шевченко Дмитро Віталійович
доктор філософії, асистент кафедри комп’ютерних наук Національного університету біоресурсів і природокористування України
Адреса робоча: Україна, м. Київ, вулиця Героїв Оборони, 15
ORCID ID: https://orcid.org/0009-0001-7736-8263
e-mail: dimashevchenko10021999@gmail.com

Голуб Белла Львівна
кандидат технічних наук, доцент, завідувач кафедри комп'ютерних наук Національного університету біоресурсів і природокористування України
Адреса робоча: Україна, м. Київ, вулиця Героїв Оборони, 15
ORCID ID: https://orcid.org/0000-0002-1256-6138
e-mail: bellalg@nubip.edu.ua

Триснюк Тарас Васильович
кандидат технічних наук, старший науковий співробітник відділу прикладної інформатики Інституту телекомунікацій і глобального інформаційного простору Національної академії наук України
Адреса робоча: Україна, м. Київ, Чоколівський бульвар, 13
ORCID ID: https://orcid.org/0000-0002-3672-8242
e-mail: taras24t@gmail.com