

© П.К. Ломазов¹, А.В. Павличенко¹, Ю.В. Бучавий¹

¹ Національний технічний університет «Дніпровська політехніка», Дніпро, Україна

МЕТОДИ ПРОГНОЗУВАННЯ ЗАБРУДНЕННЯ АТМОСФЕРНОГО ПОВІТРЯ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

© P. Lomazov¹, A. Pavlychenko¹, Yu. Buchavyi¹

¹ Dnipro University of Technology, Dnipro, Ukraine

METHODS FOR FORECASTING AIR POLLUTION BASED ON MACHINE LEARNING

Мета. Дослідити сучасні методи прогнозування забруднення атмосферного повітря з використанням алгоритмів машинного навчання.

Методика досліджень. Метод моделювання використовувався для розробки алгоритмів прогнозування, які враховують часові та просторові характеристики даних. Для уточнення залежностей між змінними було застосовано метод конкретизації. Також використовувалися методи статистичного узагальнення, що забезпечило очищення даних від шумів і заповнення пропущених значень, а також виявлення трендів і аномалій. Також, у роботі розглядаються алгоритми Random Forest, LSTM, GBM, SVM та MLP, їхні переваги, недоліки та потенціал застосування у задачах екологічного моніторингу.

Результати досліджень. Проведено кореляційний аналіз концентрацій забруднювальних речовин та метеорологічних параметрів зі станції спостереження у м. Дніпро. Обґрунтовано доцільність застосування методів машинного навчання для аналізу і прогнозування часових рядів концентрацій забруднювальних речовин. Підтверджено, що використання алгоритмів машинного навчання дозволяє досягти високої точності прогнозів завдяки обробці великого обсягу даних і врахуванню складних взаємозв'язків між джерелами забруднення та метеорологічними умовами. Алгоритми Random Forest і GBM продемонстрували ефективність у роботі з даними, що характеризуються статичними залежностями, тоді як LSTM виявився оптимальним для моделювання часових рядів.

Наукова новизна. Встановлено закономірності впливу параметрів атмосфери на точність прогнозування забруднення алгоритмами машинного навчання. Виявлено, що моделі Random Forest, LSTM, GBM тощо мають різну чутливість до обсягу й якості даних. Ці залежності сприяють покращенню методології прогнозування, забезпечуючи вищу точність і адаптацію до змінних умов.

Практичне значення. Можливість використання розроблених моделей для міського управління та екологічного моніторингу. Зокрема, моделі можуть бути інтегровані у системи раннього попередження громадян про ризики забруднення, оптимізацію транспортних потоків, а також у мобільні додатки та веб-платформи для надання інформації про стан повітря в режимі реального часу.

Ключові слова: машинне навчання, Random Forest, LSTM, метеорологічні фактори, прогнозування, IoT-технології, моніторинг.

Вступ. Проблема забруднення атмосферного повітря є однією з ключових екологічних та соціально-економічних викликів сучасності. Вплив високого рівня забруднення повітря на здоров'я населення, довкілля та кліматичні зміни обумовлює необхідність створення ефективних методів прогнозування, що дозволять своєчасно виявляти ризики та впроваджувати превентивні заходи.

Забруднення атмосферного повітря є результатом складної взаємодії багатьох факторів, серед яких особливо важливими є джерела забруднення та метеорологічні умови. Основними джерелами забруднення атмосферного повітря є антропогенні та природні фактори, серед яких найвагоміший вплив на концентрацію забруднюючих речовин мають транспорт, промисловість, енергетика, сільське господарство та будівельна діяльність. Транспортний сектор, зокрема автомобільний, є одним з основних джерел викидів шкідливих газів та часток, таких як діоксид азоту (NO_2), вуглекислий газ (CO_2), сажу та пил. Викиди від транспортних засобів мають сильний вплив на міське забруднення, оскільки найбільша частина транспортної інфраструктури сконцентрована в урбанізованих районах.

Промислові підприємства, зокрема енергетичні та хімічні заводи, викидають в атмосферне повітря велику кількість токсичних газів, таких як сірчаний ангідрид (SO_2), аміак (NH_3) та оксиди азоту. Викиди з цих джерел можуть бути стабільними або змінюватися залежно від виробничих циклів, технологічних процесів та сезонних змін. Технологічні інновації, що сприяють зменшенню викидів, є важливими для зниження рівня забруднення, проте значна частина промислових підприємств продовжує мати високий рівень екологічної небезпеки.

Природні джерела забруднення, такі як виверження вулканів, лісові пожежі та пилові бурі, можуть також суттєво впливати на якість атмосферного повітря. Однак, на відміну від антропогенних джерел, ці фактори здебільшого характеризуються сезонністю та невизначеністю, що ускладнює їх прогнозування [1].

Сучасні підходи до прогнозування забруднення базуються на поєднанні екологічного моделювання та методів аналізу даних. Використання алгоритмів машинного навчання, зокрема Random Forest і Long Short-Term Memory (LSTM), відкриває нові можливості для врахування складних взаємозв'язків між джерелами забруднення, атмосферними умовами та їх часовою динамікою.

Особливого значення набуває дослідження залежності між концентрацією забруднюючих речовин і погодними факторами, такими як температура, вологість, швидкість вітру тощо. Такий підхід забезпечує можливість прогнозувати коливання рівнів забруднення з високою точністю, що є критично важливим для урбаністичного планування, моніторингу екологічної ситуації та попередження громадян про ризики для їхнього здоров'я.

Основні частина. Слід зазначити, що збір даних про забруднення атмосферного повітря є надзвичайно важливим етапом для ефективного моделювання та прогнозування екологічних показників. Існує кілька основних джерел даних, що дозволяють отримати різноманітні, точні та своєчасні показники забруднення.

Станції моніторингу. Стаціонарні та мобільні станції екологічного моніторингу забруднення повітря є основними джерелами даних. Вони обладнані сучасними сенсорами, здатними вимірювати концентрації різноманітних

забруднюючих речовин, таких як діоксид азоту (NO_2), оксид вуглецю (CO), сірчаний ангідрид (SO_2), озон (O_3) та частки пилу (PM_{10} , $\text{PM}_{2.5}$). Станції моніторингу зазвичай належать державним або місцевим екологічним органам та іноді приватним підприємствам і дозволяють отримувати дані з конкретних територій, де встановлена станція. Однак для створення більш повної картини забруднення варто використовувати додаткові джерела інформації.

Супутникові знімки. В умовах обмеженої кількості наземних станцій моніторингу супутникові знімки надають можливість отримувати широкомасштабну інформацію про стан атмосфери на глобальному рівні. Супутникові платформи, такі як NASA Earth Observing System (EOS), European Space Agency (ESA) та інші, здійснюють моніторинг забруднення повітря через спектральні канали, що дозволяють визначити рівень забруднення на великих територіях. Технології дистанційного зондування сприяють вивченню таких параметрів, як: концентрація пилу, дим, озон, а також парникові гази. Додавання даних з супутників значно покращує точність прогнозів, дозволяючи охопити території, де немає наземних стаціонарних або мобільних станцій.

Громадські датчики. В останні роки з'явилися нові джерела даних – громадські датчики моніторингу. Це пристрої, розташовані в громадських місцях або приватних об'єктах, що вимірюють рівень забруднення повітря та передають інформацію через інтернет. Вони зазвичай дешевші і розповсюджені в більших кількостях, надаючи локалізовані дані про стан повітря в конкретному місці. Проте дані з цих датчиків можуть бути менш точними порівняно з професійними постами, що вимагає додаткової валідації і калібрування [2].

Для ефективного використання даних в алгоритмах машинного навчання необхідно провести їх ретельну попередню обробку, оскільки первісні дані часто містять шум, відсутні значення та інші проблеми, які можуть значно вплинути на точність прогнозів.

Обробка шумів. У процесі збору даних можуть спостерігатися випадкові або систематичні помилки, що викликають появу шумів. Це може бути через технічні несправності сенсорів, атмосферні умови або інтерференцією від інших джерел сигналу. Для зменшення шуму використовуються методи фільтрації, такі як фільтри Калмана, методи згладжування та фільтрація за допомогою статистичних алгоритмів (наприклад, медіанний фільтр). Це дозволяє зменшити вплив випадкових коливань та ефективно покращити якість даних, що використовуються для навчання моделей.

Відсутні значення. Проблема відсутніх значень є поширеною в наборі даних, отриманих з різних джерел. Відсутність даних може виникнути через технічну несправність сенсора, відсутність вимірювань у певний період часу або недосконалість інфраструктури збору даних. Найпопулярніші методи обробки відсутніх значень включають заповнення відсутніх значень через інтерполяцію (наприклад, лінійну або сплайн-інтерполяцію), використання статистичних методів (наприклад, заміни відсутніх значень середнім або медіанним значенням по колонці), а також складніші техніки, як прогнозування відсутніх значень за допомогою машинного навчання.

Нормалізація даних. Оскільки дані про забруднення атмосферного повітря можуть бути зібрані з різних джерел і мати різні масштаби, нормалізація є важливою для приведення їх до одного масштабу. Для цього зазвичай застосовують методи мінімаксної нормалізації або стандартизації, які забезпечують однаковий масштаб значень для всіх ознак. Це важливо для алгоритмів, таких як нейронні мережі чи SVM, оскільки різні масштаби ознак можуть негативно вплинути на ефективність моделей [3].

Для того, аби практично реалізувати алгоритми, варто почати з опису та аналізу реальних концентрацій забруднюючих речовин в атмосферному повітрі. Наприклад, на рисунку 1 наведено дані концентрацій, взятих з постів моніторингу КП «Центр екологічного моніторингу «ДОР» у м. Дніпро.

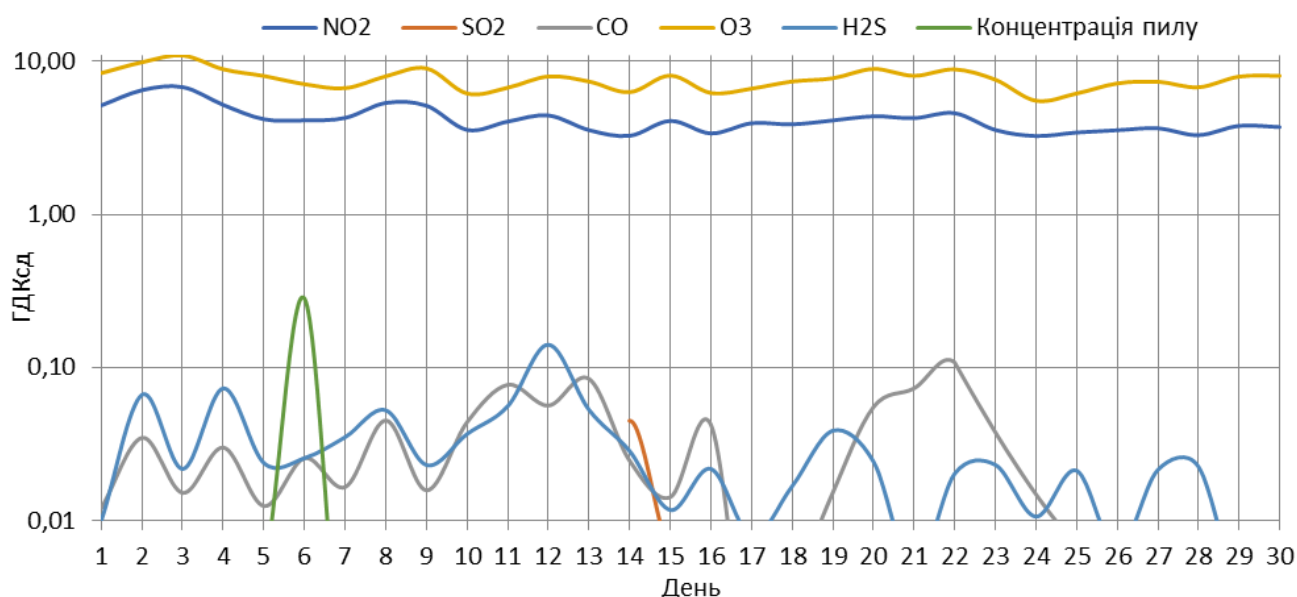


Рис. 1. Динаміка концентрацій пріоритетних забруднювальних речовин у м. Дніпро, 2018 (червень)

Масив даних, отриманих за допомогою сучасних газоаналізаторів безперервної дії та приладів для вимірювання метеорологічних параметрів що реєструються кожні 2 хвилини, містить 21492 записи за червень 2018 р., за наступними показниками: 1 – двоокис азоту (ppm), 2 – двоокис сірки (ppm), 3 – окис вуглецю (ppb), 4 – озон, 5 – сірководень (ppm), 6 – тверді речовини (мкг/м³), 7 – відносна вологість (%), 8 – напрямок вітру (град.), 9 – атмосферний тиск (Па), 10 – тип опадів, 11 – сонячна радіація (Вт/м²), 12 – температура повітря (°C), 13 – швидкість вітру (м/с), 14 – інтенсивність опадів (мм/год). Для встановлення взаємозв'язків між зазначеними моніторинговими даними було проведено кореляційний аналіз, табл. 1.

Отримані результати є неоднозначними, оскільки з одного боку ілюструють як сильні зв'язки (двоокис азоту – озон, сірководень – окис вуглецю) так і слабкі (сірководень – двоокис сірки, двоокис сірки – окис вуглецю) або навіть негативні (двоокис азоту – двоокис сірки, окис вуглецю – двоокис сірки) зв'язки між концентраціями різних забруднювальних речовин. З іншого боку було отримано

несподівано значні коефіцієнти зв'язок між концентраціями деяких забруднювальних речовин та метеорологічних показників (сірководень – швидкість вітру, окис вуглецю – швидкість вітру, окис азоту – відносна вологість).

Таблиця 1

Кореляційний аналіз концентрацій забруднювальних речовин в атмосферному повітрі та метеорологічних параметрів

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1,00													
2	-0,30	1,00												
3	0,14	0,06	1,00											
4	0,86	-0,33	0,17	1,00										
5	0,28	0,01	0,40	0,15	1,00									
6	-0,02	-0,06	-0,02	-0,09	-0,03	1,00								
7	-0,64	0,27	-0,25	-0,40	-0,30	0,02	1,00							
8	-0,29	-0,21	-0,52	-0,21	-0,38	0,22	0,50	1,00						
9	0,37	-0,21	0,05	0,04	0,10	-0,06	-0,66	-0,22	1,00					
10	-0,27	-0,04	-0,12	0,02	0,03	-0,06	0,65	0,27	-0,59	1,00				
11	0,26	-0,13	0,04	0,05	-0,19	-0,08	-0,51	-0,22	0,48	-0,81	1,00			
12	-0,35	0,16	0,13	0,04	-0,29	-0,13	0,22	-0,09	-0,51	0,22	-0,25	1,00		
13	-0,33	-0,12	-0,49	-0,38	-0,49	0,03	0,16	0,55	0,11	-0,11	0,25	-0,11	1,00	
14	-0,26	-0,04	-0,19	0,02	-0,17	-0,01	0,65	0,29	-0,56	0,93	-0,63	0,22	-0,03	1,00

Такий характер коефіцієнтів кореляції та динаміки моніторингових даних через їх значну кількість ускладнює ефективність використання «традиційних» регресійних моделей (лінійна, логарифмічна, поліноміальна згладжування даних), які можна було б застосувати для прогнозування та відновлення втрачених параметрів, що потребує пошуку альтернативних методів прогнозування забруднення, на основі машинного навчання.

Важливо зазначити, що метеорологічні фактори, зокрема температура, вологість, швидкість вітру та атмосферний тиск, мають вирішальне значення для прогнозування забруднення атмосферного повітря. Їх взаємодія з забруднювачами може значною мірою впливати на рівень забруднення в різних регіонах. Відображення впливу та взаємодія метеорологічних факторів відображено в таблиці 2.

Ці фактори є критичними для прогнозування забруднення повітря, оскільки взаємодія метеорологічних умов з забруднювачами змінюється в залежності від часу доби, сезону та географічного розташування. Тому, для точного моделювання та прогнозування, важливо інтегрувати метеорологічні фактори в процес обробки даних, що дозволяє створити більш ефективні системи попередження та моніторингу забруднення [4].

Моделювання забруднення атмосфери з урахуванням численних змінних, таких як джерела забруднення та метеорологічні фактори, вимагає використання складних багатовимірних моделей, здатних враховувати нелінійні взаємодії між

цими змінними. Важливим кроком є інтеграція взаємодій між змінними, що дозволяє більш точно прогнозувати рівень забруднення.

Таблиця 2

Вплив метеорологічних факторів на рівень забруднення повітря

Метеорологічний фактор	Вплив на забруднення повітря
Температура	Вища температура сприяє хімічним реакціям, що збільшують концентрацію озону та інших речовин. Впливає на змішування повітря.
Вологість	Висока вологість сприяє осадженню дрібних часток, зменшуючи їх концентрацію у повітрі.
Швидкість вітру	Допомагає розсіювати або транспортувати забруднюючі речовини, знижуючи або підвищуючи рівень забруднення.
Атмосферний тиск	Умови температурної інверсії можуть утримувати забруднення біля земної поверхні, сприяючи його накопиченню.

Одним з основних підходів є використання багатфакторних моделей, що включають як прямі впливи (наприклад, вплив температури на концентрацію озону), так і складніші взаємодії між змінними (наприклад, взаємодія між рівнем забруднення та швидкістю вітру). Для цього використовуються різноманітні методи машинного навчання, зокрема методи ансамблів, які дозволяють комбінувати результати кількох моделей для покращення точності прогнозів.

Інтеграція метеорологічних змінних у моделі машинного навчання є важливою складовою в побудові систем прогнозування, здатних враховувати постійно змінювані умови довкілля. За допомогою таких підходів можна створювати високоточні прогностичні системи, що допомагають у реальному часі визначати рівень забруднення та вжити необхідних заходів для зниження його впливу на здоров'я населення [3].

У контексті прогнозування забруднення повітря застосовуються різноманітні алгоритми машинного навчання, які забезпечують високу точність завдяки обробці великих обсягів даних і врахуванню складних залежностей між змінними. Найпоширеніші підходи включають алгоритми на основі дерев рішень, такі як Random Forest, нейронні мережі, наприклад, LSTM, а також методи Gradient Boosting Machines (GBM), Support Vector Machines (SVM) та Multilayer Perceptron (MLP). Кожен із цих алгоритмів має свої переваги, недоліки та специфіку застосування.

Алгоритм Random Forest (RF) є ансамблевим методом, що використовує множину дерев рішень для підвищення точності прогнозів. Основною перевагою RF є стійкість до переобчислення завдяки використанню техніки беггінгу. Проте

RF має певні недоліки, серед яких чутливість до розмірності даних, що може впливати на обчислювальну складність, а також неспроможність моделювати часові залежності. Цей алгоритм добре підходить для задач прогнозування, що базуються на статичних ознаках, таких як географічні дані або середньодобові показники забруднення [5].

Long Short-Term Memory (LSTM) є різновидом рекурентних нейронних мереж, який спеціалізується на аналізі часових рядів. LSTM може моделювати довготривалі часові залежності, що робить її ідеальним інструментом для прогнозування динамічних процесів, таких як щогодинні зміни концентрації забруднюючих речовин. Однак LSTM має і недоліки, зокрема високу обчислювальну складність і потребу в значних обсягах даних для ефективного навчання. Крім того, оптимізація її параметрів є складним завданням, що вимагає спеціальних навичок [6].

Gradient Boosting Machines (GBM), до яких належать XGBoost, LightGBM і CatBoost, є потужними інструментами для роботи з великими наборами даних. GBM демонструють високу точність завдяки поетапному навчанню, що дозволяє зменшувати помилки попередніх ітерацій. Вони також відзначаються гнучкістю в налаштуванні гіперпараметрів і здатністю ефективно обробляти великі обсяги ознак. Серед недоліків GBM відзначається схильність до переобчислення на великих наборах даних і вразливість до шумів, якщо не було виконано попередньої обробки [7].

Алгоритм Support Vector Machines (SVM) використовується для задач класифікації та регресії, особливо на малих наборах даних. SVM забезпечує хорошу узагальнюючу здатність і підходить для задач, де кількість змінних обмежена. Однак цей метод не завжди ефективний для великих наборів даних через високу обчислювальну складність. Крім того, його точність значною мірою залежить від вибору ядра та гіперпараметрів, що вимагає ретельної оптимізації [8].

Multilayer Perceptron (MLP) є базовим типом нейронних мереж, який може виконувати нелінійне моделювання. Цей алгоритм є менш ресурсомістким, ніж LSTM, і легше реалізується. Проте його ефективність є нижчою для задач, пов'язаних із часовими рядами. Для досягнення високої точності в складних задачах MLP може потребувати великої кількості шарів, що ускладнює його навчання та оптимізацію.

Для апробації алгоритмів машинного навчання на основі зібраних моніторингових даних про концентрації та метеорологічні параметри використовувався програмний комплекс XLSTAT, рис. 2.

Значний масив вихідних моніторингових даних дозволив сформулювати як навчальні так і тестові вибірки, що необхідні для машинного навчання за наведеними вище алгоритмами. Точність прогнозування змінювалася (залежно від алгоритму, обсягу вибірки даних та налаштувань додаткових параметрів розрахунку) в достатньо широкому діапазоні від 32 до 76%, періодично досягала понад 80% на окремих (коротких) ділянках часових рядів.

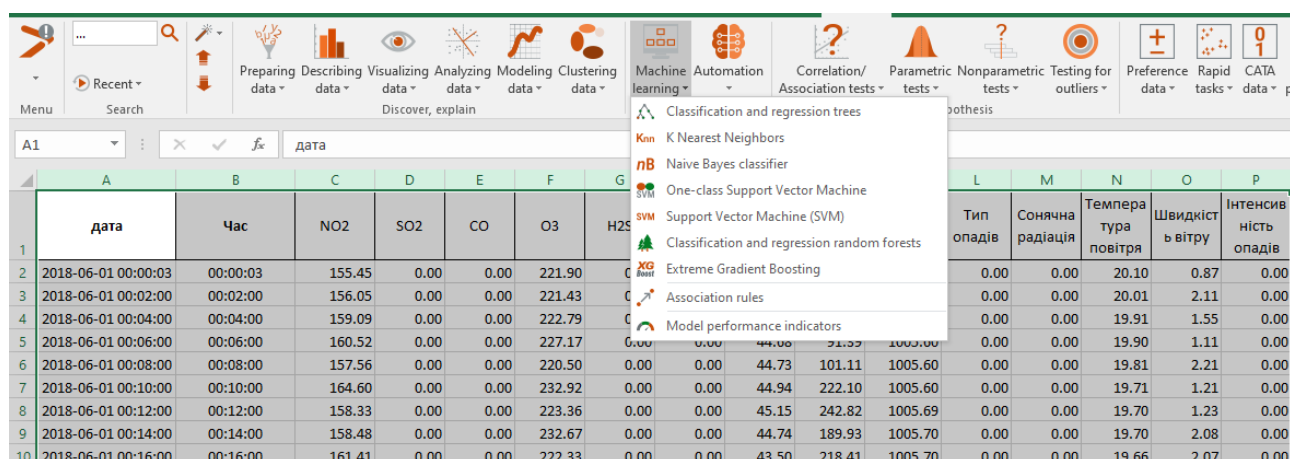


Рис. 2. Застосування методів машинного навчання для прогнозування концентрацій забруднювальних речовин в атмосферному повітрі (у програмному середовищі XLSTAT)

Узагальнене порівняння алгоритмів машинного навчання в узагальненому вигляді можна переглянути в таблиці 3.

Таблиця 3

Порівняння алгоритмів машинного навчання

Алгоритм	Часові ряди	Точність	Швидкість навчання	Інтерпретація	Обчислювальна складність
Random Forest	✓	Висока	Швидка	Висока	Середня
LSTM	✓	Висока	Повільна	Низька	Висока
GBM	✗	Висока	Середня	Середня	Середня
SVM	✗	Середня	Повільна	Висока	Висока
MLP	✗	Середня	Швидка	Низька	Середня

Для задач прогнозування забруднення атмосферного повітря вибір алгоритму залежить від типу даних. Random Forest і GBM підходять для аналізу статичних даних, тоді як LSTM є оптимальним для прогнозування на основі часових рядів. У випадках обмежених ресурсів або невеликих наборів даних можна використовувати SVM або MLP. Поєднання кількох методів, таких як GBM для вибору ознак і LSTM для аналізу часових залежностей, може забезпечити найкращі результати.

Одним з перспективних підходів у побудові моделей прогнозування є комбінування алгоритмів Random Forest та LSTM, оскільки це дозволяє поєднати сильні сторони кожного з них. Random Forest може бути використаний для попередньої обробки даних, виділення важливих ознак і побудови моделей на основі ансамблю дерев, тоді як LSTM може обробляти часові залежності між даними та надавати більш точні прогнози для майбутнього [9].

Комбінація цих методів дозволяє створити потужні прогностичні системи, які можуть враховувати як негайний вплив метеорологічних змінних і джерел

забруднення, так і довгострокові патерни і цикли, що спостерігаються в атмосфері. Наприклад, можна застосувати Random Forest для виділення найбільш значущих факторів впливу, а потім передавати ці ознаки як вхідні дані до LSTM, щоб забезпечити прогнози, які враховують часові тренди та взаємозалежності між різними змінними.

Тож, вибір алгоритму машинного навчання для прогнозування забруднення повітря залежить від специфіки задачі, обсягу доступних даних та вимог до точності й інтерпретації результатів. Комбінація різних методів може забезпечити оптимальні результати, поєднуючи їхні сильні сторони та мінімізуючи недоліки [9].

Налаштування гіперпараметрів є ключовим етапом оптимізації моделей машинного навчання, оскільки воно значно впливає на точність та здатність моделей до узагальнення. Для алгоритму Random Forest основними параметрами є кількість дерев (*n_estimators*), максимальна глибина дерев (*max_depth*), мінімальна кількість вибірок для поділу вузлів (*min_samples_split*), мінімальна кількість вибірок у листі (*min_samples_leaf*), та кількість ознак для кожного дерева (*max_features*). Оптимальне налаштування цих параметрів дозволяє уникнути перенавчання та забезпечити високу продуктивність моделі [5].

У випадку LSTM гіперпараметри, такі як кількість нейронів у прихованому шарі (*units*), кількість шарів (*layers*), розмір пакету (*batch_size*), швидкість навчання (*learning_rate*) та параметри регуляризації (*dropout*, *recurrent_dropout*), визначають здатність моделі до вивчення складних залежностей у часових рядах. Оптимізація здійснюється за допомогою методів, таких як пошук по сітці, випадковий пошук або байєсівська оптимізація. Грамотне налаштування гіперпараметрів забезпечує баланс між ефективністю моделі та її здатністю працювати в реальних умовах [6].

У сучасному світі зростання урбанізації та індустріалізації створює додаткові виклики для екологічного управління, зокрема у сфері контролю якості повітря. Використання моделей машинного навчання для прогнозування рівня забруднення атмосфери відкриває нові можливості для прийняття обґрунтованих рішень щодо управління довкіллям. Розробка таких моделей не лише покращує точність передбачень, але й сприяє інтеграції наукових підходів до реальних практичних задач. Наступним важливим етапом є розгляд практичних аспектів застосування цих моделей у контексті міського управління якістю повітря.

Моделі прогнозування забруднення повітря, створені на основі алгоритмів машинного навчання, можуть бути впроваджені в широкий спектр завдань міського управління. Однією з ключових переваг є можливість використання цих моделей для оперативного моніторингу та прогнозування концентрацій забруднюючих речовин, що дозволяє приймати своєчасні рішення щодо зниження екологічних ризиків.

Завдяки точному прогнозуванню концентрації забруднюючих речовин, моделі можуть стати основою для створення систем раннього попередження. Такі системи інформують громадськість про можливі екологічні загрози, спричинені підвищенням рівня забруднення повітря, що дозволяє громадянам коригувати

свою поведінку (наприклад, уникати перебування на відкритому повітрі в періоди пікового забруднення). Наприклад, попередження про високу концентрацію дрібнодисперсних частинок РМ_{2.5} може спонукати міських мешканців користуватися індивідуальними засобами захисту дихальних шляхів [10].

Моделі машинного навчання можуть бути інтегровані у системи міського управління транспортом для мінімізації викидів забруднюючих речовин. Прогнозування забруднення дозволяє визначити ділянки з високим рівнем концентрації шкідливих речовин та оптимізувати транспортні потоки для зниження екологічного навантаження. Наприклад, моделі можуть рекомендувати зміни в організації руху транспорту або впровадження зон з обмеженням викидів у періоди високої забрудненості.

На основі прогнозів можна формувати довгострокові екологічні стратегії та регуляторні заходи, спрямовані на зниження рівня забруднення атмосфери. Наприклад, аналіз джерел забруднення в поєднанні з прогнозами дозволяє визначати найбільш значущі фактори впливу на якість повітря та розробляти політики, спрямовані на їх зниження, зокрема обмеження викидів для промислових підприємств або стимулювання використання електротранспорту.

Дані про забруднення та прогнози його динаміки можуть бути використані при плануванні міського розвитку. Врахування таких даних дозволяє більш раціонально розмішувати житлові зони, промислові об'єкти, зелені насадження та інші міські елементи з метою зменшення впливу забруднення на населення. Наприклад, створення «зелених коридорів» у міських районах із високим рівнем забруднення сприяє покращенню якості повітря.

Прогностичні моделі можуть бути інтегровані у структуру розумних міст, забезпечуючи автоматизований збір і обробку даних з екологічних датчиків, транспортних систем та метеорологічних станцій. Це сприяє створенню єдиного інформаційного простору, в якому прогнозування рівня забруднення стає невід'ємною частиною екологічного моніторингу. Наприклад, система управління розумного міста може автоматично адаптувати режими роботи вентиляційних систем або обмежувати доступ транспорту до певних зон у разі підвищеного забруднення [4].

Використання моделей машинного навчання для прогнозування та подальшої інтеграції цих прогнозів у мобільні додатки та веб-платформи відкриває нові горизонти у формуванні екологічно обґрунтованої поведінки населення.

Системи оповіщення громадян, побудовані на основі прогнозів моделей машинного навчання, забезпечують оперативне інформування населення про зміну якості повітря. Це особливо важливо для попередження потенційно небезпечних ситуацій, пов'язаних із високими концентраціями забруднюючих речовин.

Прогностичні моделі, такі як Random Forest і LSTM, можуть бути використані для створення точних і релевантних прогнозів, які враховують часові та просторові характеристики забруднення. Дані про поточний стан атмосферного повітря, доповнені метеорологічними факторами, проходять через обробку і аналіз для генерації прогнозів про короткострокові та довгострокові зміни концентрацій шкідливих речовин.

Системи оповіщення мають бути багаторівневими та забезпечувати диференційоване інформування громадян. Наприклад, на рівні екстремальних ситуацій можуть бути реалізовані сповіщення через SMS-повідомлення або push-сповіщення в мобільних додатках із зазначенням конкретних рекомендацій щодо поведінки (обмеження перебування на відкритому повітрі, використання засобів індивідуального захисту тощо). Більш загальна інформація про прогноз якості повітря може бути доступною у форматі дашбордів у веб-платформах або мобільних додатках [10].

Оцінка та валідація моделей машинного навчання є також невід'ємною частиною їхньої розробки, оскільки вони забезпечують обґрунтованість і точність прогнозів у реальних умовах. Зокрема, перевірка стійкості моделей, порівняння їх результатів з фактичними даними та дослідження потенціалу масштабування на інші регіони мають ключове значення для їх ефективного використання в екологічному моніторингу.

Крос-валідація виступає фундаментальним підходом для оцінки стійкості моделей машинного навчання до змін у вибірках даних. У контексті прогнозування забруднення атмосфери, крос-валідація дозволяє оцінити здатність моделей узагальнювати знання на нових наборах даних, які не використовувалися в процесі навчання.

Найпоширенішим методом є *k-fold крос-валідація*, за якої початковий набір даних розбивається на k підмножин. На кожному етапі $k - 1$ підмножини використовуються для навчання, а залишкова підмножина – для тестування. Циклічний процес забезпечує рівномірне використання всіх даних як для навчання, так і для перевірки, що мінімізує ризики перенавчання та дозволяє об'єктивно оцінити продуктивність моделей [3].

Ефективність моделей оцінюється шляхом порівняння прогнозів із фактичними даними, отриманими зі станцій моніторингу якості повітря, супутникових вимірювань або даних громадських екологічних датчиків. Цей процес передбачає використання метрик, таких як:

1. Середня абсолютна помилка (MAE) – забезпечує оцінку середнього відхилення між прогнозованими та реальними значеннями, що дозволяє виявити загальний рівень помилок моделі.

2. Корінь середньоквадратичної помилки (RMSE) – враховує квадратичне відхилення, акцентуючи увагу на більших помилках.

3. Коефіцієнт детермінації (R^2) – визначає, наскільки добре модель пояснює варіацію залежної змінної, демонструючи її здатність відображати взаємозв'язки між параметрами.

Аналіз результатів дозволяє не лише оцінити точність моделей, але й виявити області для їх вдосконалення, наприклад, шляхом оптимізації вибору ознак чи налаштування гіперпараметрів [6].

Розвиток систем прогнозування забруднення атмосферного повітря має широкий спектр перспектив, які охоплюють інтеграцію новітніх технологій, удосконалення алгоритмічних підходів та впровадження моделей у процес формування екологічної політики. Ці напрями сприяють не лише підвищенню точності

та ефективності прогнозів, а й зміцненню зв'язку між наукою, технологіями та управлінськими рішеннями.

Інтеграція технологій Інтернету речей (IoT) відкриває нові горизонти для збору, аналізу та використання даних про забруднення у реальному часі. Розумні датчики, розташовані в різних частинах міста або поблизу промислових об'єктів, дозволяють фіксувати концентрацію забруднюючих речовин та метеорологічні параметри з високою точністю. Це не лише розширює базу даних для моделей, але й забезпечує адаптивність прогнозів, враховуючи динамічні зміни умов навколишнього середовища. Зростання кількості доступних датчиків сприяє створенню прогнозів із високою просторовою деталізацією, що є важливим для міського управління, зокрема, для розробки локалізованих заходів зі зниження забруднення [11].

Важливим напрямом вдосконалення є використання більш складних алгоритмічних архітектур, зокрема трансформерів, які вже зарекомендували себе в багатьох галузях. Завдяки своїй здатності обробляти довгострокові залежності у часових рядах, трансформери можуть значно покращити якість прогнозування динаміки забруднення. Їхня масштабованість дозволяє працювати з великими наборами даних, які генеруються IoT-пристроями, а мультизадачність дає змогу одночасно прогнозувати кілька параметрів, таких як рівень забруднення повітря та метеорологічні умови, що впливають на його розсіювання.

Ще одним стратегічним напрямом є впровадження моделей у формування політик, спрямованих на зниження рівня забруднення. На основі прогнозів можна розробляти заходи, що враховують специфіку забруднення у різних районах. Наприклад, обмеження транспортного руху в зонах із підвищеним рівнем забруднення у певний час або коригування роботи промислових підприємств залежно від погодних умов, що сприяють накопиченню шкідливих речовин. Такі заходи, доповнені інформуванням громадян про ризики та рекомендаціями щодо поведінки, можуть суттєво знизити техногенний вплив на довкілля.

Таким чином, перспективи розвитку моделей прогнозування забруднення атмосферного повітря базуються на інтеграції передових технологій, інноваційних алгоритмів та їхньому практичному використанні у процесі прийняття екологічних рішень. Це забезпечить не лише підвищення точності прогнозів, але й сприятиме побудові сталих систем управління якістю повітря, що враховують як технічні, так і соціальні аспекти [12].

Висновки. Отже, результати дослідження підтверджують високу ефективність алгоритмів машинного навчання для прогнозування забруднення атмосферного повітря. Використання таких методів, як Random Forest, LSTM, GBM, SVM та MLP, забезпечує обробку великих обсягів даних і врахування складних залежностей між змінними. Алгоритми Random Forest і GBM оптимально підходять для аналізу статичних даних, тоді як LSTM демонструє високу ефективність у прогнозуванні часових рядів. Поєднання цих методів дозволяє досягти синергетичного ефекту, що сприяє покращенню точності прогнозів.

Використання моделей у реальних умовах доводить їх значний потенціал для підтримки прийняття рішень у міському управлінні, зокрема для розробки

систем раннього попередження, оптимізації транспортних потоків та формування довгострокових екологічних стратегій. Інтеграція з IoT-технологіями забезпечує можливість отримання даних у реальному часі, підвищуючи адаптивність та точність прогнозування.

Подальші дослідження можуть бути спрямовані на розробку нових гібридних алгоритмів, що поєднують сильні сторони сучасних нейронних мереж та традиційних моделей машинного навчання. Крім того, інтеграція моделей з розумними міськими платформами та їх адаптація до специфіки різних регіонів потребують додаткового аналізу. Важливим напрямом залишається розробка екологічних політик, заснованих на прогнозах, та створення систем, які активно інформуватимуть громадян про ризики забруднення.

Перелік посилань

1. Міщук, О.С. (2019). Підвищення точності прогнозування параметрів забруднення повітря. У *Комп'ютерне моделювання та оптимізація складних систем (КМОСС2019): матеріали V-ї міжнародної науково-технічної конференції*, 129–130.
2. Вишневський, А.В., & Потапова, Н.А. (2024). Методи аналізу даних в інформаційній системі прогнозування показників забруднення повітря. *Прикладні аспекти сучасних міждисциплінарних досліджень*, 156–157.
3. Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. *IEEE Access*, 7, 128325–128338. <https://doi.org/10.1109/access.2019.2925082>
4. Запорожець, А. О. (2017). Аналіз засобів моніторингу забруднення повітря навколишнього середовища. *Science-Based Technologies*, 35(3), 242–252.
5. Renugadevi, R., Vyshnavi, T., Prasanna, T. R., & Lahari, P. S. (2023). Air quality prediction using random forest algorithm. *International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMK-MATE)*, 1–6. <https://doi.org/10.1109/RMKMATE59243.2023.10369180>
6. Hayasouk, T., Lee, H., & Lee, G. (2020). Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability*, 12(6), 2570. <https://doi.org/10.3390/su12062570>
7. Su, Y. (2020). Prediction of air quality based on gradient boosting machine method. *International Conference on Big Data and Informatization Education (ICBDIE)*, Zhangjiajie, China, 395–397. <https://doi.org/10.1109/ICBDIE50010.2020.00099>
8. Widyarini, L., & Purnomo, H. (2024). Air quality prediction using the support vector machine algorithm. *Journal of Information Systems and Informatics*, 6(2), 652–661. <https://doi.org/10.51519/journalisi.v6i2.705>
9. Surya, R., Jeevananthan, P., Muniyappan, A., Prakash, P. K. B., & Saveeth, P. (2024). Air Quality Monitoring and Prediction for Industries. *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, 1–6. <https://doi.org/10.1109/icstem61137.2024.10560996>
10. Widyarini, L., & Purnomo, H. (2024). Air quality prediction using the support vector machine algorithm. *Journal of Information Systems and Informatics*, 6(2), 652–661. <https://doi.org/10.51519/journalisi.v6i2.705>
11. Осадчук, О., Осадчук, Я., & Довгун, В. (2024). Досягнення та виклики хмарних технологій моніторингу якості повітря. *Measuring and Computing Devices in Technological Processes*, (2), 18–25. <https://doi.org/10.31891/2219-9365-2024-78-2>
12. Freeman, B. S., Taylor, G., Gharabaghi, B. J., & Thé, J. (2018). Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8), 866–886. <https://doi.org/10.1080/10962247.2018.1459956>

ABSTRACT

Goal. To investigate modern methods for predicting air pollution using machine learning algorithms.

Methodology. A modeling approach was used to develop forecasting algorithms that consider the temporal and spatial characteristics of the data. The method of specification was applied to refine the dependencies between variables. Statistical generalization methods were also employed to remove noise, fill in missing values, and identify trends and anomalies. Additionally, this study examines the Random Forest, LSTM, GBM, SVM, and MLP algorithms, discussing their advantages, limitations, and potential applications in environmental monitoring.

Research results. A correlation analysis of pollutant concentrations and meteorological parameters was conducted using data from an observation station in Dnipro. The feasibility of applying machine learning methods for analyzing and forecasting time series of pollutant concentrations was substantiated. The results confirm that machine learning algorithms enable high forecasting accuracy by processing large datasets and capturing complex relationships between pollution sources and meteorological conditions. The Random Forest and GBM algorithms proved effective for data with static dependencies, whereas LSTM was optimal for time series modeling.

Scientific novelty. The patterns of the influence of atmospheric parameters on the accuracy of pollution prediction by machine learning algorithms have been established. It has been found that models such as Random Forest, LSTM, GBM, etc., exhibit different sensitivity to the volume and quality of input data. These dependencies contribute to improving forecasting methodology, ensuring higher accuracy and adaptation to changing conditions.

Practical significance. The developed models can be applied in urban management and environmental monitoring. Specifically, they can be integrated into early warning systems for informing citizens about pollution risks, optimizing traffic flows, and supporting mobile applications and web platforms that provide real-time air quality information.

Keywords: *machine learning, Random Forest, LSTM, meteorological factors, forecasting, IoT technologies, monitoring.*