ensure this requirement and are not designed to work with a random container, which is most frequently used in practice. The aim of this work is to increase the efficiency of a steganographic system when using an arbitrary steganographic method and a random container. This is achieved through the development of a method for selecting container blocks from a digital image to embed additional information within them. The criterion for efficiency is the reliability of perception of the generated steganogram, quantitatively evaluated using PSNR. The goal was achieved by investigating the properties of formal parameters of digital image blocks – singular values, considering the correspondence between singular triplets and the frequency components of the block. The most important result of this work is the theoretically substantiated definition of a block parameter that provides an integral quantitative characteristic of the relative distribution of its frequency components – the normalized separability of the maximum singular value of the block. The practical significance of the obtained results lies in the development of a method for selecting image blocks and a method for identifying images undesirable for use as containers. The algorithmic implementations of these methods have significantly improved the efficiency of the steganographic system when applied.

***Keywords**: steganography, digital image, reliability of perception of the steganogram, container block selection, container selection, normalized separability of the singular value*

UDC 004.623
DOI https://doi.org/10.36059/978-966-397-538-2-5

ІНТЕЛЕКТУАЛЬНИЙ ГІБРИДНИЙ ПРОЄКТ СИСТЕМИ ДЛЯ АНАЛІЗУ РИЗИКІВ ШКОДИ ЗДОРОВ'Ю ПРИ ВЕЛИКИХ ОБСЯГАХ ДАНИХ

Ph.D. М. Рудніченко[1][0000-0002-7343-8076], **Ph.D. Н. Шибаєва**[1][0000-0002-7869-9953], **Ph.D. Т. Отрадська**[2][0000-0002-5808-5647], **Д. Шведов**[1][0009-0002-4823-8782], **Ph.D. І. Шпінарева**[1][0000-0001-9208-4923], **Dr.Sci. І.Петров**[1][0000-0002-8740-6198]

[1]Національний університет «Одеська Політехніка», Україна,
[2]Одеський коледж комп'ютерних технологій «СЕРВЕР», Україна
EMAIL: nickolay.rud@gmail.com, nati.sh@gmail.com, tv_61@ukr.net, frumle@ukr.net, iryna.shpinareva@onu.edu.ua, firmn@gmail.com

***Анотація.** У статті представлено всебічне дослідження щодо розробки проекту гібридної інтелектуальної системи для аналізу ризиків шкоди здоров'ю населення на великих обсягах екологічних даних, спричиненої забрудненням повітря, із використанням гібридної архітектури на основі моделей машинного та глибинного навчання. Актуальність дослідження обґрунтована зростаючими загрозами для здоров'я, пов'язаними з атмосферними забруднювачами, такими як PM2.5, PM10, NO_2, SO_2, CO та*

O_3, особливо у густонаселених міських середовищах. Стаття підкреслює нагальну потребу використання інтелектуальних систем на основі даних для надання точних, масштабованих та інтерпретованих оцінок ризиків для здоров'я населення на рівні популяції. Проведено детальний аналіз існуючих наукових підходів, виділено обмеження класичних статистичних моделей та підкреслено переваги використання нейронних мереж і методів ансамблевого навчання для моделювання складних нелінійних залежностей та просторово-часової динаміки у гетерогенних екологічних даних. Запропонована система реалізує шестиступеневу модельну структуру, що включає дерева рішень, метод опорних векторів, випадкові ліси, XGBoost, згорткові нейронні мережі та моделі довготривалої короткочасної пам'яті (LSTM). Кожен компонент сприяє формуванню структурованого конвеєра для попередньої обробки даних, класифікації, прогнозування та стратифікації ризиків. У дослідженні описано використані набори даних для навчання та валідації, переважно Air Pollution Image Dataset з Індії та Непалу, доповнені структурованими наборами даних якості повітря. Архітектура системи реалізована за моделлю клієнт-сервер, де бекенд розроблено на Python та Flask, а фронтенд — на JavaScript. Компоненти програмного забезпечення та взаємодії класів представлені через UML-діаграми, а аналітична панель системи обладнана модулями для візуалізації даних, прогнозування моделей та порівняння їхньої ефективності. Експериментальні результати підтверджують ефективність запропонованого підходу з використанням ансамблевих та глибинних моделей. Проведено порівняльну оцінку всіх моделей за показниками точності (accuracy), точності прогнозу (precision), повноти (recall), F1-міри та середнього часу навчання.

Ключові слова: *інтелектуальний аналіз даних, аналіз здоров'я, великі дані, гібридний інтелектуальний аналіз, аналіз даних, нейронні мережі, машинне навчання*

Вступ

У останні десятиліття швидка індустріалізація, зростання міст і посилення антропогенної діяльності призвели до значного погіршення якості атмосферного повітря, особливо в густонаселених районах. Наслідки для здоров'я від тривалого впливу забруднювачів повітря, таких як дрібні частинки ($PM_{2.5}$ та PM_{10}), двоокис азоту (NO_2), двоокис сірки (SO_2), озон (O_3) та чадний газ (CO), дедалі більше визнаються світовою науковою спільнотою. Ці забруднювачі пов'язані з низкою гострих і хронічних захворювань, включаючи респіраторні та серцево-судинні хвороби, негативні наслідки для новонароджених, порушення нейропсихічного розвитку та підвищену смертність.

У результаті органи охорони громадського здоров'я, екологічні агентства та міські планувальники стикаються з зростаючим тиском щодо впровадження науково обґрунтованих і технологічно ефективних стратегій зменшення ризиків для здоров'я населення, пов'язаних із забрудненням повітря [1].

Незважаючи на зростаючий обсяг епідеміологічних досліджень, що висвітлюють причинно-наслідкові зв'язки між забруднювачами повітря та станом здоров'я, впровадження інтелектуальних систем, здатних оцінювати ризики для здоров'я на рівні популяції, залишається недостатньо розвиненим. Існує критичний розрив між теоретичними моделями взаємодії здоров'я та навколишнього середовища та практичними застосуваннями, здатними обробляти гетерогенні дані в режимі реального часу, враховувати просторові та часові варіації й надавати персоналізовані або спільнотні оцінки ризику. Відсутність інтегрованих платформ, здатних агрегувати, аналізувати та інтерпретувати великі обсяги екологічних та медичних даних, ускладнює проактивне планування громадського здоров'я, комунікацію ризиків та цільові втручання [2].

У науковій літературі все частіше підкреслюється, що традиційні методи оцінки впливу забруднювачів на здоров'я населення, зокрема класичні статистичні моделі, обмежені у можливості враховувати складні нелінійні взаємозв'язки, просторово-часову динаміку забруднення та різнорідність даних про стан здоров'я. Класичні регресійні моделі часто не здатні коректно моделювати взаємодію між багатьма змінними, що обмежує точність прогнозів та адекватність рекомендацій для політики громадського здоров'я.

Однією з фундаментальних проблем у цій сфері є відсутність стандартизованих, високоякісних та відкритих наборів даних, які одночасно містять детальну інформацію про умови навколишнього середовища та параметри здоров'я на індивідуальному або популяційному рівні. Хоча існують окремі системи моніторингу навколишнього середовища та реєстри здоров'я, їх інтеграція часто ускладнюється різницею у структурі даних, роздільній здатності, доступності та обмеженнях конфіденційності. Крім того, багато наборів даних мають проблеми, такі як відсутні значення, непослідовні вимірювання, тимчасові невідповідності та обмежене географічне охоплення. Ці обмеження знижують здатність створювати надійні моделі, які могли б узагальнюватися на різні регіони та популяції, що в кінцевому результаті зменшує ефективність оцінки ризиків для здоров'я та прогнозів [3].

Ще однією великою перешкодою є складність збору, попередньої обробки та агрегування даних з кількох джерел. Дані про якість повітря можуть отримуватися з супутникових зображень, наземних моніторингових станцій або мобільних сенсорів, кожне з яких має різну просторову та часову роздільну здатність. Дані про здоров'я можуть надходити з лікарень, клінік, громадських опитувань або носимих пристроїв, часто у несумісних форматах і з різним ступенем надійності. Процес уніфікації цих потоків даних у узгоджені, аналізовані формати потребує значних обчислювальних та методологічних зусиль, включаючи очищення даних, створення ознак, нормалізацію та імпутацію. Крім того, етичні та правові аспекти, пов'язані з конфіденційністю даних та анонімізацією, ще більше ускладнюють отримання та використання даних про здоров'я на індивідуальному рівні у таких дослідженнях [4].

Сучасні дослідження демонструють значний потенціал застосування методів машинного та глибинного навчання для прогнозування ризиків, пов'язаних із забрудненням повітря. Нейронні мережі, ансамблеві моделі (Random Forest, XGBoost) та моделі послідовностей (LSTM) дозволяють ефективно інтегрувати великі масиви гетерогенних даних, моделювати складні залежності та враховувати як просторові, так і тимчасові варіації забруднення. Наприклад, у дослідженнях останніх років показано, що ансамблеві методи значно підвищують точність прогнозів смертності та захворюваності в порівнянні зі статистичними моделями, а згорткові нейронні мережі ефективні для аналізу зображень супутникового моніторингу та дистанційного оцінювання якості повітря.

Слід зазначити, що існує кілька ключових проблем, які залишаються невирішеними та ускладнюють впровадження інтелектуальних систем для оцінки ризиків для здоров'я:

– відсутність стандартизованих та відкритих наборів даних, що одночасно містять детальну інформацію про екологічні умови та параметри здоров'я населення. Незважаючи на наявність численних систем моніторингу повітря та медичних реєстрів, їх інтеграція часто утруднена через різницю у форматах, роздільній здатності, часових масштабах та обмеження конфіденційності. Це значно ускладнює створення узагальнюваних моделей, здатних працювати на різних регіональних рівнях;

– низька якість та неповнота даних. Багато наборів даних містять пропущені значення, непослідовні або несумісні вимірювання, тимчасові зсуви та обмежене географічне охоплення. Це призводить до необхідності складної попередньої обробки даних, включно з очищенням, нормалізацією, імпутацією та інженерією ознак;

– складність інтеграції даних із різних джерел. Дані про якість повітря можуть надходити з супутників, наземних станцій або мобільних сенсорів, а дані про здоров'я – із лікарень, клінік, опитувань або носимих пристроїв. Об'єднання цих потоків у єдиний формат для аналізу потребує значних обчислювальних ресурсів та методологічної компетенції;

– етичні та правові обмеження. Використання індивідуальних даних про стан здоров'я потребує дотримання правил конфіденційності та анонімізації, що додатково ускладнює проведення досліджень на рівні населення.

Незважаючи на ці проблеми, останні дослідження [3, 4, 5] демонструють, що інтелектуальні системи мають значний потенціал для підвищення ефективності оцінки ризиків, пов'язаних із забрудненням повітря. Вони можуть стати основою для:

– раннього попередження населення про небезпечні рівні забруднення;

– персоналізованого моніторингу впливу забруднення на здоров'я конкретних груп населення;

– прийняття рішень у сфері міського планування та екологічного регулювання;

– підвищення ефективності цільових медичних та соціальних втручань.

Інтелектуальні системи також можуть інтегруватися в концепцію «розумного міста», поєднуючи технологічні інновації з цілями забезпечення рівного доступу до здоров'я та сталого розвитку.

Таким чином, перспективи використання інтелектуальних систем для революціонізації оцінки ризиків для громадського здоров'я, пов'язаних із забрудненням повітря, є значними. Зі збільшенням доступності даних та розвитком методів моделювання ці системи можуть забезпечувати механізми раннього попередження, полегшувати персоналізоване відстеження експозиції та сприяти адаптивним стратегіям міського планування, надання медичної допомоги та екологічного регулювання.

Актуальність даного дослідження обумовлена поєднанням кількох факторів: зростанням глобальної урбанізації та антропогенного навантаження, високим рівнем забруднення повітря в міських агломераціях, недостатнім розвитком інтегрованих платформ для оцінки ризиків для здоров'я, а також потребою у впровадженні сучасних методів машинного та глибинного навчання для обробки гетерогенних даних. Проведений огляд наукових досліджень підтверджує, що створення інтелектуальної системи для оцінки ризиків здоров'я на основі даних про забруднення повітря є важливим і актуальним напрямом сучасної науки та практики громадського здоров'я.

У підсумку такі інтелектуальні платформи можуть стати ключовими компонентами інфраструктури «розумних міст», поєднуючи технологічні інновації з цілями забезпечення здоров'я та сталого розвитку. У цьому контексті метою даного дослідження є внесок у цю сферу шляхом розробки інтелектуальної системи, призначеної для оцінки рівнів ризику для здоров'я населення на основі впливу забруднення повітря [5].

1. Постановка проблеми та аналіз існуючих підходів та публікацій

З огляду на ці виклики, розробка та впровадження інтелектуальних систем для оцінки ризиків для здоров'я населення є не лише необхідністю, а й складним технічним завданням. Останні досягнення у сфері машинного навчання (ML) та глибинного навчання (DL) [6] відкривають перспективні можливості для вирішення складнощів, що виникають при моделюванні впливу забруднення повітря на здоров'я. Ці підходи здатні навчатися складним нелінійним взаємозв'язкам між вхідними змінними та результатами для здоров'я, інтегрувати різноманітні типи даних і адаптуватися до динамічних умов [7]. На відміну від традиційних статистичних методів, які часто потребують сильних апріорних припущень та обмежених взаємодій змінних, моделі машинного навчання можуть виявляти приховані закономірності у даних, автоматично визначати впливові ознаки та ефективно масштабуватися для роботи з високовимірними наборами даних.

Зокрема, глибинні нейронні мережі, архітектури згорткових та рекурентних мереж, механізми уваги та ансамблеві методи навчання [8] продемонстрували високі результати у задачах моделювання впливу

забруднення на здоров'я. Це включає прогнозування рівнів концентрацій забруднювачів, оцінку ризику експозиції, моделювання захворюваності залежно від екологічних факторів та навіть імітацію довгострокових ефектів політичних втручань. При інтеграції з геопросторовими даними, часовими рядами та соціодемографічними показниками такі моделі можуть надавати високоточні та контекстно-залежні оцінки вразливості населення та розподілу ризиків. Крім того, методи пояснюваного штучного інтелекту (XAI) все частіше застосовуються у цих системах для забезпечення прозорості та підтримки науково обґрунтованого прийняття рішень органами охорони громадського здоров'я.

Незважаючи на технологічні досягнення, залишаються значні перешкоди. Багато застосувань ML [9] у цій сфері досі перебувають на експериментальному рівні та часто базуються на обмежених або синтетичних наборах даних. Для реального впровадження потрібна надійна валідація, адаптація моделей до конкретних умов та узгодження з місцевими політичними та медичними системами.

Також необхідна міждисциплінарна співпраця між дата-сайентистами, епідеміологами, інженерами-екологами та фахівцями з громадського здоров'я, щоб моделі були не лише технічно досконалими, а й контекстно релевантними та етично обґрунтованими. Крім того, необхідно ретельно оцінювати інтерпретованість та справедливість AI-оцінок здоров'я, щоб уникнути потенційних упереджень та небажаних наслідків у розподілі ресурсів або комунікації ризику.

Наприклад, деякі дослідження [10] демонструють високу точність класифікації рівнів забруднення повітря за допомогою ансамблевих ML-моделей та нейронних мереж. Особливо відзначається застосування XGBoost та LSTM, які досягли високої точності прогнозування рівнів AQI (понад 90%). Сильними сторонами цих робіт є акцент на інтерпретованості моделей та порівняльна оцінка алгоритмів. Проте, незважаючи на технічну досконалість, такі дослідження обмежуються прогнозуванням якості повітря та не включають прямої оцінки наслідків для здоров'я. Насправді ж дослідження можуть зосереджуватися не лише на прогнозуванні рівнів забруднення, а й на розробці моделей, що безпосередньо оцінюють ризики для здоров'я населення.

Робота [11] пропонує більш прикладний підхід, пов'язуючи концентрації $PM_{2.5}$ та PM_{10} з кількістю амбулаторних звернень через респіраторні захворювання. Використовуючи багатошаровий перцептрон, автори демонструють можливість побудови надійних прогнозних моделей для конкретних вікових груп, що є важливим кроком у напрямку персоналізованої оцінки ризику. Сильними сторонами цього дослідження є використання реальних медичних даних із сезонною та демографічною стратифікацією. Водночас модель обмежена своєю «мілкою» архітектурою та обмеженим географічним охопленням. Інтелектуальна система може подолати ці обмеження шляхом впровадження більш просунутих архітектур, таких як

LSTM та механізми уваги, а також розширення географічного покриття та інтеграції додаткових прогнозних показників.

За результатами досліджень [12] можна відзначити гібридний підхід глибинного навчання, який поєднує просторові згорткові нейронні мережі, часові рекурентні модулі та механізми уваги для прогнозування концентрацій забруднювачів. Ключовим внеском цього дослідження є здатність моделі працювати з неповними даними та враховувати просторово-часові залежності. Водночас її застосування обмежується прогнозуванням забруднення без прямого зв'язку з наслідками для здоров'я.

У наступній роботі [13] запропоновано інноваційне застосування нейронних мереж U-Net як сурогатної моделі для швидкої оцінки впливу короткострокової експозиції $PM_{2.5}$ на здоров'я. Модель замінює ресурсоємні атмосферні симуляції умовною моделлю глибинного навчання, що забезпечує високошвидкісні обчислення та масштабованість. Основною перевагою цього підходу є ефективність та потенціал для використання в реальному часі. Проте дослідження фокусується на імітації фізичної моделі, а не на створенні комплексного конвеєра оцінки ризику для здоров'я. Тому, хоча метод технічно обґрунтований, він не інтегрує епідеміологічні або клінічні дані.

Інший підхід [14] використовує класичні GIS-методи для картографування забруднення повітря та оцінки ризику для здоров'я, застосовуючи супутникові дані $PM_{2.5}$. Завдяки просторовій інтерполяції та використанню функцій «експозиція–відповідь» автори створюють детальну карту експозиції для міських районів із рідкісними наземними сенсорами. Хоча перевагою цього методу є висока просторова роздільна здатність, він не враховує часові динаміки та не використовує методи ML для прогнозного моделювання. У сучасних інтелектуальних системах можна інтегрувати як просторові, так і часові виміри експозиції, а також застосовувати мультимодальні дані для моделювання ризиків для здоров'я населення за допомогою передових методів глибинного навчання [15].

Таким чином, можна підтвердити, що сучасні аналітичні системи широко використовують методи ML для попередньої обробки даних, виділення ознак та прогнозного моделювання. Моделі навчаються та валідуються на інтегрованих наборах даних, що поєднують дані екологічних сенсорів та індикатори здоров'я населення. Наша мета – побудувати масштабовану та інтерпретовану платформу, яка не лише прогнозує ризики для здоров'я з високою точністю, але й допомагає у формуванні цільових втручань та прийнятті політичних рішень [16].

Наукова новизна цієї роботи полягає у розробці end-to-end конвеєра для оцінки ризиків для здоров'я на основі даних, який вирішує проблеми гетерогенності, обмеженості та складності даних. Практичне значення визначається потенційною інтеграцією системи у реальне моніторинг громадського здоров'я та управління міським середовищем. Завдяки систематичній оцінці продуктивності та інтерпретованості моделей запропонований підхід прагне створити основу для ширшого застосування

інтелектуальних систем у сфері оцінки впливу навколишнього середовища на здоров'я та підтримки прийняття рішень у політиці.

Узагальнюючи аналіз наведених джерел слід зазначити, що останні дослідження у сфері інтелектуальних систем для оцінки впливу забруднення повітря на здоров'я демонструють значне різноманіття методологічних підходів та технічних рішень, що відображає розвиток машинного та глибинного навчання у цій галузі. Одним із перспективних напрямів є гібридні моделі глибинного навчання, які поєднують просторові згорткові нейронні мережі, часові рекурентні модулі та механізми уваги для прогнозування концентрацій забруднювачів. Ці моделі вирізняються здатністю працювати з неповними або частково відсутніми даними та одночасно враховувати просторово-часові залежності, що підвищує точність прогнозів і дозволяє моделювати динамічні процеси забруднення у складних урбанізованих середовищах. Водночас такі підходи залишаються обмеженими у зв'язку з відсутністю прямого зв'язку між прогнозованими рівнями забруднення та конкретними наслідками для здоров'я населення, що обмежує їхню практичну цінність у контексті громадського здоров'я.

Іншим напрямом є застосування нейронних мереж U-Net у якості сурогатних моделей для швидкої оцінки впливу короткострокової експозиції $PM_{2.5}$. Цей підхід дозволяє замінити ресурсоємні атмосферні симуляції умовною моделлю глибинного навчання, що забезпечує значне скорочення часу обчислень і масштабованість системи. Основним досягненням таких моделей є ефективність у реальному часі, що відкриває перспективи для оперативного моніторингу ризиків та швидкого реагування на зміну рівнів забруднення. Проте даний метод фокусується на імітації фізичної моделі забруднення, не інтегруючи безпосередньо епідеміологічні чи клінічні дані, що обмежує можливості комплексної оцінки ризиків для здоров'я на рівні популяції.

У більш класичних підходах, заснованих на геоінформаційних системах, для картографування забруднення та оцінки ризиків використовуються супутникові дані $PM_{2.5}$ у поєднанні з просторовою інтерполяцією та функціями «експозиція–відповідь». Ці методи дозволяють отримувати деталізовані карти експозиції для міських районів із обмеженою кількістю наземних сенсорів. Перевагою такого підходу є висока просторова роздільна здатність, що забезпечує локалізоване відображення рівнів забруднення та потенційного ризику. Разом з тим, метод не враховує часові динаміки та не застосовує алгоритми машинного навчання для прогнозного моделювання, що обмежує його здатність до передбачення майбутніх змін у забрудненні та відповідних наслідків для здоров'я.

Сучасні інтелектуальні системи намагаються поєднувати переваги цих різних підходів, інтегруючи просторові та часові виміри експозиції, а також мультимодальні дані про стан навколишнього середовища та індикатори здоров'я населення. Це дозволяє створювати прогностичні моделі високої точності, які здатні оцінювати ризики для здоров'я з урахуванням

комплексних взаємозв'язків між екологічними та соціальними факторами. Крім того, системи останнього покоління приділяють значну увагу інтерпретованості моделей, використовуючи методи пояснюваного штучного інтелекту, що забезпечує прозорість прийняття рішень та підтримує науково обґрунтоване формулювання політики громадського здоров'я.

Порівняльний аналіз робіт останніх років дозволяє визначити кілька ключових тенденцій та обмежень. Гібридні DL-підходи демонструють високі можливості для прогнозування концентрацій забруднювачів із врахуванням просторово-часових залежностей, але потребують інтеграції з даними про здоров'я для практичної оцінки ризику.

Сурогатні моделі U-Net забезпечують швидкі та масштабовані розрахунки, проте обмежені у зв'язку з відсутністю комплексного конвеєра оцінки здоров'я. Класичні GIS-підходи відзначаються високою просторовою роздільною здатністю та здатністю працювати з обмеженими даними сенсорів, але не враховують часову динаміку та не використовують ML для прогнозування. Інтеграція цих аспектів у сучасних інтелектуальних системах дозволяє формувати моделі, що одночасно забезпечують високоточне прогнозування забруднення, оцінку ризику для здоров'я та підтримку цільових заходів та політичних рішень.

Високий потенціал сучасних систем полягає у створенні масштабованих та інтерпретованих платформ, здатних працювати з інтегрованими наборами даних, що об'єднують інформацію з сенсорів навколишнього середовища та показники здоров'я населення. Наукова новизна таких підходів полягає у розробці end-to-end конвеєрів для оцінки ризиків здоров'я на основі даних, які враховують проблеми гетерогенності, обмеженості та складності інформації. Практичне значення визначається потенційною інтеграцією цих систем у реальні процеси моніторингу громадського здоров'я та управління міським середовищем, а також їхньою здатністю підтримувати обґрунтоване прийняття рішень на рівні політики.

В результаті аналізу джерел створена порівняльна таблиця основних підходів із декількома критеріями оцінки відображає переваги та обмеження різних методологій у контексті прогнозування забруднення та оцінки ризику для здоров'я, яка наведена нижче (таблиця 1).

Таким чином, сучасні наукові дослідження демонструють тенденцію до поєднання переваг різних методологій, прагнучи до створення систем, які забезпечують високоточну оцінку ризиків для здоров'я населення на основі комплексних даних про забруднення та соціально-демографічні фактори. Ключовим завданням залишається інтеграція прогнозування екологічних показників з безпосередньою оцінкою впливу на здоров'я та розробка інтерпретованих і масштабованих платформ, здатних підтримувати прийняття рішень у сфері громадського здоров'я та управління міським середовищем.

Таблиця 1

Порівняння основних підходів із декількома критеріями оцінки

Підхід	**Тип моделі**	**Облік просторово-часових залежностей**	**Прогнозування в реальному часі**	**Основні переваги**	**Основні обмеження**
Гібридні DL-моделі	CNN + RNN + Attention	Так	Обмежено	Робота з неповними даними, точне прогнозування концентрацій	Відсутність прямого зв'язку з наслідками для здоров'я
U-Net сурогат-ні моделі	Conditional DL	Обмежено	Так	Висока швидкість, масштабованість	Фокус на фізичній моделі, відсутність інтеграції медичних даних
GIS-підхід	Просторові інтерполяції + функції «експозиція–відповідь»	Частково	Ні	Висока просторова роздільна здатність, робота з обмеженими сенсорними даними	Не враховує часові зміни, відсутність ML-прогнозування
Сучасні інтегро-вані системи	ML + DL, мультимодальні дані	Так	Так	Висока точність прогнозів, інтеграція даних, інтерпретованість	Вимагає складної попередньої обробки та міждисциплінарної співпраці

2. Розробка концепції проекту

Запропонована архітектура інтегрує декілька моделей машинного (ML) та глибинного навчання (DL), кожна з яких виконує спеціалізовану функцію в межах ієрархічного конвеєра оцінки ризиків [17]. Такий підхід забезпечує не лише високу точність прогнозування, але й підвищену інтерпретованість результатів і адаптивність до різних умов даних та екологічних контекстів.

Початковий етап передбачає попередню фільтрацію та категоризацію даних за допомогою моделі дерева рішень (DT). Завдяки своїй інтерпретованості та ієрархічній логіці ця модель дозволяє ідентифікувати ключові закономірності та порогові значення показників забруднювачів, які

справляють першочерговий вплив на стан здоров'я. Цей етап також сприяє класифікації регіонів чи часових сегментів за рівнем забруднення, формуючи основу для диференційованої обробки на наступних етапах.

Другий етап зосереджений на виявленні складних нелінійних взаємозв'язків між показниками якості повітря та медико-демографічними характеристиками населення. Це завдання реалізується через машину опорних векторів (SVM), яка демонструє високу ефективність при роботі з гетерогенними ознаками та здатна будувати точні роздільні поверхні у багатовимірному просторі ознак. Вихідним результатом цього етапу є окреслення потенційних зон ризику та формалізація критичних умов, за яких вплив забруднювачів стає суттєво шкідливим для здоров'я.

На третьому етапі застосовується ансамблеве моделювання з використанням алгоритму випадкового лісу (RF), метою якого є агрегування попередніх результатів та уточнення структури факторів ризику шляхом повторної вибірки та побудови множини дерев рішень. Це підвищує стійкість системи до викидів та шумів, а також покращує надійність і відтворюваність оцінок ризику.

Четвертий етап передбачає використання моделі градієнтного бустингу XGBoost для посилення прогностичної спроможності системи. Інтегруючи додаткові часові, просторові та соціально-економічні змінні, раніше не враховані, ця модель поступово підвищує точність оцінки ризику завдяки ітеративній корекції помилок і ефективному опрацюванню складних міжфакторних залежностей.

На п'ятому етапі інтегрується згорткова нейронна мережа (CNN), яка аналізує просторово-часові закономірності, особливо у випадках, коли доступні геопросторові шари чи структури сенсорних мереж. Ця модель дозволяє системі враховувати географічну кореляцію забруднювачів та ідентифікувати просторові кластери ризиків для здоров'я, що є надзвичайно важливим для цільових втручань у сфері громадського здоров'я та стратегічного планування політики.

Фінальний етап реалізується через модель довгої короткострокової пам'яті (LSTM), яка відображає часову динаміку впливу забруднення на здоров'я. Завдяки здатності навчатися довгостроковим залежностям модель LSTM дозволяє системі прогнозувати відстрочені ефекти для здоров'я та моделювати сценарії ризику за різних стратегій екологічної політики.

Інтеграція цих моделей у багатоступеневу архітектуру формує комплексну, багатовимірну та стійку до помилок систему, здатну не лише точно оцінювати поточні ризики, але й прогнозувати їхню еволюцію за різних сценаріїв. Запропонована гібридна концепція вносить наукову новизну завдяки структурованому поєднанню інтерпретованих та глибинних моделей у єдину аналітичну систему, тим самим розширюючи можливості моніторингу, управління та зменшення ризиків для здоров'я населення, пов'язаних із забрудненням повітря.

Основним набором даних, обраним для нашого дослідження, є Air Pollution Image Dataset з Індії та Непалу [18]. Цей набір містить колекцію анотованих зображень, що відображають різні рівні забруднення повітря у міських і сільських середовищах. Його релевантність до поставленого завдання визначається багатим візуальним представленням умов забруднення, яке дозволяє навчати DL-моделі, зокрема CNN, для виявлення та класифікації рівня тяжкості забруднення. Це добре узгоджується з метою інтеграції просторово-контекстних характеристик у конвеєр оцінки ризику.

Для підвищення надійності та узагальнюваності вихідних результатів модель була додатково розширена та валідована за допомогою структурованих екологічних вимірювань із двох авторитетних джерел [19, 20]. Ці додаткові джерела дали змогу здійснити крос-валідацію рівнів забруднення, покращити калібрування порогів прогнозування та забезпечити узгодженість між ознаками, отриманими зі зображень, та числовими показниками забруднення. Це зміцнило цілісність та практичну застосовність запропонованої мультимодельної системи для оцінки ризиків здоров'я.

Проблема нашого дослідження може бути сформульована як завдання багатокласової класифікації. У ML під багатокласовою класифікацією розуміють задачу прогнозного моделювання, у якій алгоритм повинен віднести спостережуваний об'єкт до одного з кількох попередньо визначених класів. У контексті оцінки ризику для здоров'я, зумовленого забрудненням повітря, це завдання може бути математично формалізоване відповідним чином.

Вхідний набір даних у нашому випадку можна представити наступним чином: $X = \{x_1, x_2, ..., x_n\}$, де кожен об'єкт x_i є вектором ознак з простору $\Re^d$. Для кожного об'єкта X_i мітка класу збігається $y_i \in \{1,2,...,K\}$, де K — кількість класів (6 класів у рамках нашого завдання). Потрібно побудувати функцію $f : \Re^d \to \{1,2,...,K\}$, який для будь-якого вхідного об'єкта x передбачить мітку класу y (ризик для громадського здоров'я). Модель машинного навчання побудована з використанням навчальної вибірки $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, яка може бути сформована з інформативних вхідних ознак, головне завдання полягає в знаходженні наближення функції f на основі цих даних. Якщо $P(y = k \mid x)$ - ймовірність належності об'єкта x класу k, тоді функція $f(x)$ прогнозує клас з апостеріорною ймовірністю

$$f(x) = \arg \max_{k \in \{1,2,...,K\}} P(y = k \mid x)$$

Для навчання моделі використовується функція втрат, яка кількісно визначає розбіжність між передбачуваними та справжніми мітками класів.

Однією з найчастіше використовуваних функцій втрат для цієї мети є втрата перехресної ентропії:

$$L(y, y') = -\sum_{k=1}^{K} y_k \log(y'_k)$$

де y_k — двійковий індикатор (0 або 1), що вказує, чи належить об'єкт до класу k, $y' = P(y = k \mid x)$ - ймовірність того, що об'єкт належить до класу k, передбачено моделлю. Модель оптимізується шляхом мінімізації функції втрат L використання методів оптимізації, таких як градієнтний спуск.

Фінальне прогнозування здійснюється шляхом вибору класу з найвищою ймовірністю на основі навченої моделі. Загальний конвеєр системи, використаний у дослідженні, представлений на рис. 1. Імпортовані набори даних завантажуються за допомогою функцій бібліотеки Pandas та зберігаються як об’єкти DataFrame. Далі виконується етап попередньої обробки, який включає виявлення аномалій і викидів, а також усунення дисбалансу класів у вихідній змінній.

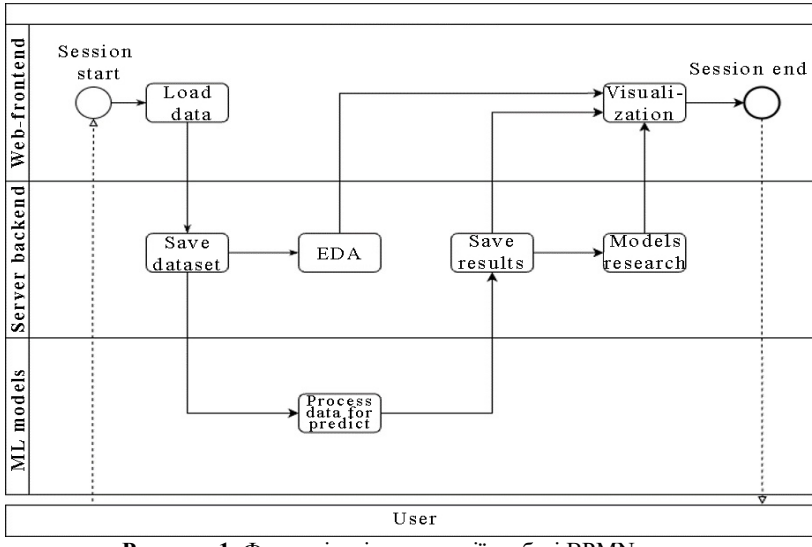


Рисунок 1. Формалізація концепції на базі BPMN схеми

Після попередньої обробки здійснюється комплекс процедур розвідувального аналізу даних (EDA). До них належать статистичний аналіз,

аналіз кореляцій, факторний аналіз, а також за необхідності - зменшення розмірності у випадках, коли кількість вхідних ознак є занадто великою.

На етапі моделювання будуються окремі ML-моделі, формується ансамбль моделей та розробляються глибокі повнозв'язні штучні нейронні мережі, навчені на тренувальних і тестових підмножинах даних. Ці підмножини формуються шляхом розподілу даних і підлягають процедурам крос-валідації для забезпечення надійності моделей.

Дисбаланс класів розв'язується за допомогою підходу зважування класів, коли ваги обчислюються як обернена частота появи класу у наборі даних. Це ефективно підвищує «штраф» за неправильну класифікацію недостатньо представлених класів, покращуючи чутливість моделі до міноритарних категорій. При цьому генерація синтетичних даних не застосовується, що позитивно впливає на надійність і валідність результатів класифікації.

Навчені моделі оцінюються за допомогою вибраних метрик продуктивності (наприклад, accuracy), тестуються на невідомих даних і серіалізуються у окремі файли для подальшого завантаження та застосування до нових наборів даних з метою оцінки рівня ризику для громадського здоров'я. Перед обробкою даних імпортуються необхідні бібліотеки. Серед них NumPy та Pandas для роботи зі структурами даних і попередньої обробки, Matplotlib та Seaborn для візуалізації, а також різні модулі бібліотеки scikit-learn (sklearn). Останні використовуються для таких завдань, як кодування категоріальних (рядкових чи текстових) ознак у числові формати, нормалізація даних, оцінка метрик продуктивності та ініціалізація моделей (наприклад, DecisionTreeClassifier).

У діаграмі класів (рис. 2) клас ExploratoryDataAnalysis відповідає за виконання розвідувального аналізу даних. Він містить атрибути для зберігання шляхів до директорій для збереження графічних результатів, а також набір методів для створення візуалізацій.

До таких методів належать generateDataDistributionPlot, generateEmissionIndexPlot, generateCorrelationMatrixPlot, generateZScorePlot, generatePairplotPlot та generateClassDistributionPlot.

Центральний файл server.py містить основну операційну логіку системи. У ньому визначені головні API-ендпоінти, включно з get_session (ініціалізація сесії), upload_file (завантаження даних), start_eda (запуск розвідувального аналізу), predict (інференція моделі) та compare (порівняння продуктивності моделей). Клас Regression відповідає за виконання регресійних моделей. Він містить атрибути для зберігання результатів прогнозування, а також серіалізованих моделей, серед яких decisionTreeModel.pkl, randomForestModel.pkl, XGBoostModel.pkl, mlpModel.keras, lstmModel.keras та cnnModel.keras. Клас забезпечує методи для масштабування даних (scaleData), генерації прогнозів (predict, predictProba) та створення різноманітних візуалізацій, зокрема generateAQIByLocationPlot, generateCorrelationMatrixPlot та generateAQIByTimePlot.

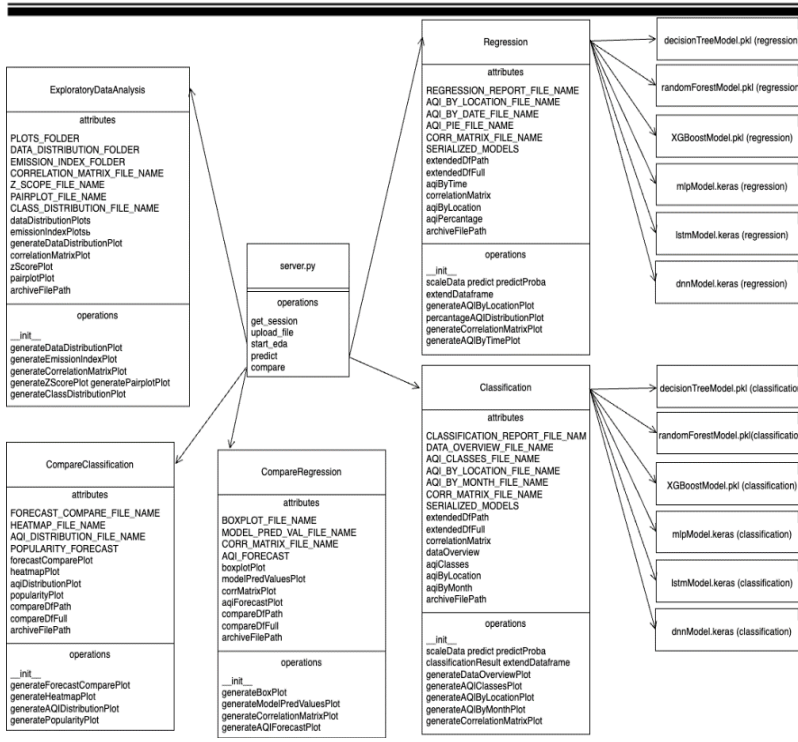


Рисунок 2. Діаграма головних класів проекту

Клас Classification, призначений для класифікаційних завдань, має подібну структуру до Regression, але зосереджується на віднесенні екологічних даних до наперед визначених класів. Він керує результатами класифікації та дозволяє розширювати вихідний датафрейм (classificationResult, extendDataframe). Для класифікації використовується той самий набір архітектур моделей: decisionTreeModel.pkl, randomForestModel.pkl, XGBoostModel.pkl, mlpModel.keras, lstmModel.keras та cnnModel.keras.

Класи CompareClassification та CompareRegression відповідають за порівняльний аналіз результатів класифікаційних та регресійних моделей відповідно. Вони надають широкий набір методів візуалізації, серед яких forecastComparePlot, heatmapPlot, compareDFPlot, boxplot, corrMatrixPlot та aqiForecastPlot, що забезпечують інтерпретованість результатів роботи моделей і дають змогу проводити міжмодельне бенчмаркінг-порівняння.

Інтелектуальна система побудована на клієнт-серверній архітектурі, де бекенд реалізовано на Python із використанням мікрофреймворку Flask, а фронтенд розроблено мовою JavaScript із застосуванням бібліотеки React.

Основний функціонал системи охоплює розвідувальний аналіз даних, прогнозне моделювання із використанням методів класифікації та регресії, а також порівняльну оцінку продуктивності моделей.

Бекенд базується на Flask, який забезпечує легковаговий і гнучкий каркас для побудови RESTful API, що дозволяє реалізувати безшовну комунікацію між фронтендом і серверними компонентами. Flask обробляє HTTP-маршрутизацію, яка відповідає за отримання, обробку та збереження екологічних даних. Структурна організація бекенд-проєкту представлена в наступному розділі (рис. 3).

Фронтенд системи реалізовано за допомогою фреймворку React, що забезпечує динамічний та зручний для користувача інтерфейс. Основні компоненти інтерфейсу включають сторінку завантаження файлів, модулі для відображення результатів розвідувального аналізу даних, сторінку прогнозування для запуску моделей та спеціальний розділ для порівняння моделей. Взаємодія з бекендом здійснюється через бібліотеку Axios, яка дозволяє виконувати HTTP-запити та забезпечує ефективну комунікацію з серверною частиною застосунку.

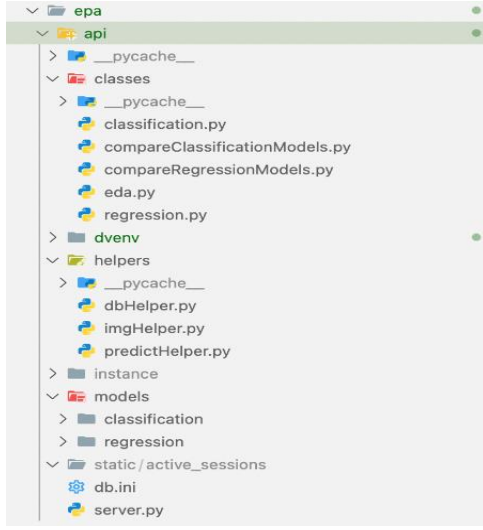


Рисунок 3. Структура проекту у середовищі розробки

3. Дослідження роботи та обчислювальні експерименти

На початковому етапі дані були імпортовані та збережені у вигляді структурованих об’єктів DataFrame за допомогою бібліотеки Pandas. Для оптимізації простору ознак було видалено непотрібні або неінформативні

атрибути, зокрема назви файлів, ідентифікатори року та місяця. До категоріальних змінних, таких як локація та година, було застосовано кодування міток (Label Encoding) з використанням LabelEncoder для перетворення текстових даних у числовий формат, сумісний із алгоритмами машинного навчання.

Далі датасет було піддано низці діагностичних процедур на наявність пропусків і аномалій. Записи з високою часткою відсутніх значень було вилучено, тоді як часткові пропуски заповнювалися методами середньої підстановки на основі атрибутної агрегації. Розподіли ключових забруднювачів ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO та O_3) були візуалізовані за допомогою гістограм та оцінок ядерної щільності для аналізу асиметрії, дисперсії та модальності. Також були використані boxplot-графіки для ідентифікації викидів у розрізі класів індексу якості повітря (AQI).

Для подолання проблеми дисбалансу класів у цільовій змінній (AQI_Class) було застосовано методи оверсемплінгу. Використовувалися алгоритми SMOTE та ADASYN, які синтетично збільшували кількість прикладів міноритарних класів, утворюючи збалансовані вибірки для навчання класифікаційних моделей. Отримані розширені датасети були експортовані для подальшого використання.

Було проведено кореляційний аналіз із використанням метрик Пірсона та Спірмена для оцінки лінійних і монотонних зв'язків між ознаками. Кореляційні матриці візуалізувалися у вигляді теплових карт, а значущість асоціацій додатково перевірялася статистичними тестами. Серед них t-тест Велча для порівняння значень AQI між парами класів, однофакторний дисперсійний аналіз (ANOVA) для багатокласових порівнянь і тест χ^2 для оцінки залежності між категоріями AQI та бінаризованими рівнями забруднення (наприклад, високий проти низького $PM_{2.5}$).

Таким чином, було виконано кореляційний аналіз ознак, а результат його візуалізації у вигляді теплової карти представлено на рис. 4. Для доповнення цих аналізів проведено виявлення викидів із використанням двох незалежних підходів = методу Z-оцінки та методу інтерквартильного розмаху (IQR). Теплові карти на основі Z-оцінки та boxplot-графіки на основі IQR надали додаткову інформацію про аномальні розподіли числових характеристик. Додатково було досліджено вимірну структуру датасету за допомогою факторного аналізу та методів зниження розмірності.

Факторний аналіз із використанням заданої трьохкомпонентної моделі виявив латентні конструкти, що лежать в основі даних про забруднювачі та значення ІСЯ (AQI). Завантаження (loadings) були візуалізовані у вигляді теплової карти для інтерпретації внеску змінних у різні фактори. Для зменшення простору ознак до двох вимірів було застосовано метод головних компонент (PCA) та t-розподілене стохастичне сусіднє вбудовування (t-SNE). Трансформовані уявлення ознак були візуалізовані для оцінки відокремлюваності класів та внутрішньої структури даних.

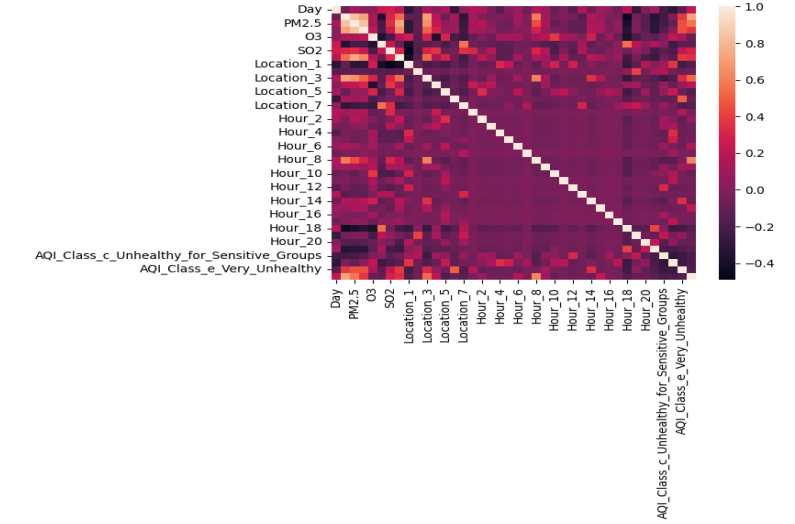


Рисунок 4. Теплова карта кореляційних оцінок між вхідними ознаками набору даних

Як було виявлено під час кореляційного аналізу, окрім неінформативної ознаки, що представляє ім'я файлу зображення, високі значення кореляції особливо спостерігаються серед атрибутів, що характеризують забруднення повітря шкідливими речовинами, зокрема $PM_{2.5}$ та PM_{10}. Це пояснюється природою їх вимірювання та схожістю інструментів і методологій, використаних для їхнього виявлення. У рамках попереднього аналізу даних, їх очищення та попередньої обробки було прийнято рішення видалити атрибут Filename із DataFrame, оскільки він не несе аналітичної цінності в контексті досліджуваного завдання.

Під час подальшого дослідження було встановлено, що ознаки Month і Year демонструють сильну міжкореляцію та роблять непослідовний внесок у структуру даних. У зв'язку з цим їх було виключено з фінального набору ознак, щоб уникнути потенційних проблем мультиколінеарності та зменшити надлишковість. Під час реалізації аналізу пропущених значень за допомогою методу isnull() було виявлено, що ознаки O_3, CO, SO_2 та NO_2 містять понад 2000 пропущених записів. Для розв'язання цієї проблеми було застосовано метод імпутації шляхом обчислення середнього значення відповідних змінних і заміни пропусків за допомогою функції mean(), що дозволило відновити цілісність набору даних і мінімізувати втрати інформації.

Розглянемо процес EDA та аналізу даних у межах локально розгорнутої системи з точки зору кінцевого користувача (рис. 5).

Розділ завантаження даних надає короткий опис базових вимог до вхідного файлу, який користувач може завантажити, а також містить

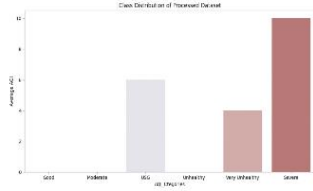
інтерактивну кнопку «Upload file». Після завантаження користувачеві відображається назва та характеристики обраного файлу для перевірки його коректності. Після підтвердження шляхом натискання кнопки «Confirm data» інтерфейс переходить до секції аналізу даних.

Location	Year	Month	Day	Hour	PM2.5	PM10	O3	CO	SO2	NO2	AQI	AQI_Ctegories
0	2023	2	23	6	43.0	78.0	26.0	258.0	10.0	17.0	5	Severe
2	2023	2	22	3	500.0	480.0	91.0	78.0	17.0	47.0	5	Severe
5	2022	10	2	0	185.0	199.0	10.0	52.0	12.0	26.0	4	Very Unhealthy
2	2023	2	16	3	401.0	325.0	73.0	88.0	16.0	57.66157894736842	5	Severe
4	2023	3	23	5	14.0	41.0	35.0	6.0	5.0	7.0	2	USG

Download CSV report

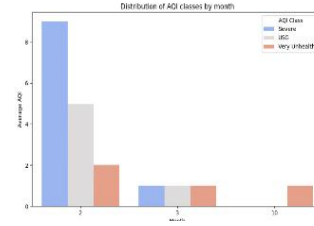
AQI Class Distribution

An AQI Class Distribution plot visualizes the frequency or proportion of different Air Quality Index (AQI) categories (e.g., Good, Moderate, Unhealthy) in a dataset, helping to understand air quality patterns.



Distribution of AQI classes by time

The Distribution of AQI Classes by Month plot shows how air quality index (AQI) categories (e.g., Good, Moderate, Unhealthy) vary across different months, helping to identify seasonal trends and pollution patterns.

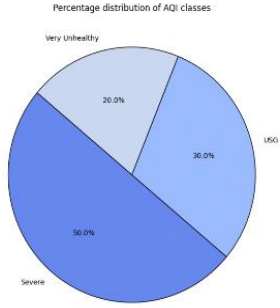


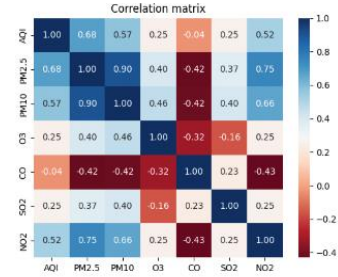
Grouping data by AQI class

Grouping data by AQI class means categorizing air quality data into predefined AQI ranges (e.g., Good, Moderate, Unhealthy) to analyze trends, compare distributions, and extract meaningful insights.

Correlation between variables

A Correlation Between Variables plot visually represents the strength and direction of relationships between multiple variables, often using a heatmap or scatter plots to identify patterns and dependencies in the data.





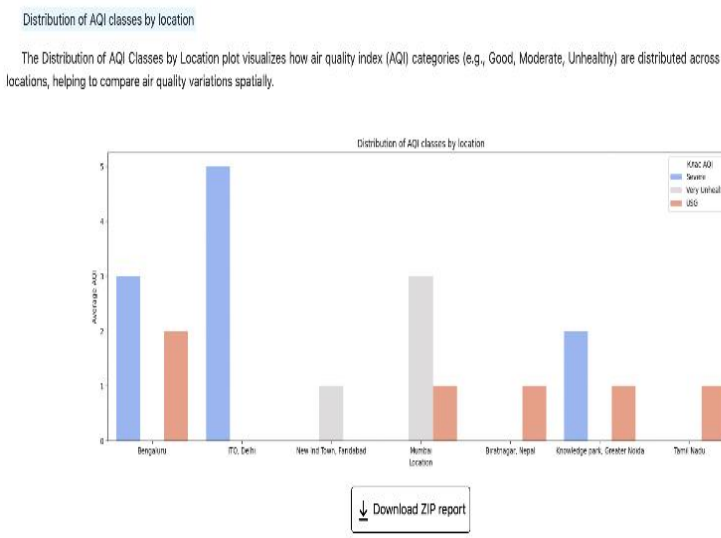


Рисунок 5. Інтерфейс користувача аналітичної панелі EDA системи

На стороні бекенду інтегровано спеціалізований модуль розвідувального аналізу даних (EDA) для обробки завантажених вхідних даних. Цей модуль забезпечує візуалізацію статистичних характеристик набору даних, побудову кореляційних матриць, графіків розподілу та реалізацію інших методів EDA. Для його імплементації використано відомі бібліотеки, зокрема Pandas, Matplotlib, Keras та Seaborn. REST API ендпоінти системи розроблені таким чином, щоб повертати аналітичні результати у вигляді зображень, які потім відтворюються у фронтенд-інтерфейсі для інтерпретації користувачем.

Для прогнозного моделювання система включає як класифікаційні, так і регресійні алгоритми, зокрема DT, SVM, RF, XGBoost, LSTM, CNN. Ці моделі попередньо навчені та збережені у серіалізованому форматі (.pkl або .keras) у файловій системі сервера. Під час ініціювання запиту бекенд-застосунок на базі Flask десеріалізує відповідну модель, обробляє підготовлені вхідні дані та повертає результат прогнозування у вигляді структурованої JSON-відповіді.

Для запуску прогнозування користувач спочатку обирає бажану модель та її тип - класифікація чи регресія. Після вибору стає доступною кнопка «Start Prediction». При її активації формується запит до бекенду для запуску обраного прогнозного конвеєра. Після генерації результатів система надсилає відповідь до фронтенду, що містить як візуальні результати (наприклад, графіки), так і детальний CSV-звіт. Зміст результатів залежить від обраного типу моделі, що відображає відмінності у логіці алгоритмів та структурі вихідних даних.

Під час оцінювання моделей машинного навчання набір даних було розділено на навчальну та тестову вибірки у співвідношенні 75:25 для підтримки процесу розробки моделей та оцінки їх продуктивності. Для структурованого логування ефективності моделей обчислені значення метрик зберігалися у спеціалізованих змінних.

Для забезпечення комплексного порівняльного аналізу ефективності класифікацій за точністю застосовувалися візуалізації матриць плутанини, згенеровані за допомогою бібліотеки Seaborn. Отримані матриці плутанини для всіх реалізованих моделей ML наведено на рис. 6. Для зручності інтерпретації вихідні класи ризику були чисельно закодовані у послідовному порядку від 0 до 5.

Узагальнені результати продемонстрували високу точність класифікації для всіх моделей, причому найвищу ефективність за показником точності класифікації продемонструвала модель XGBoost. Для подальшого порівняльного аналізу була створена візуалізація, яка ілюструє точність класифікації всіх моделей у багатокласовому середовищі. Усереднені профілі ROC-кривих для моделей наведено окремо на рис. 7.

Представлена ROC-крива демонструє продуктивність класифікації різних моделей у багатокласовому середовищі з використанням підходу one-vs-rest. Графік відображає співвідношення між показником True Positive Rate (TPR) та False Positive Rate (FPR) для кожного класифікатора. Чим ближче крива розташована до верхнього лівого кута, тим ефективніше модель розрізняє цільові класи. Серед усіх оцінених моделей найбільш сприятливу поведінку демонструє класифікатор XGBoost, ROC-крива якого максимально наближена до оптимальної межі. Це свідчить про стабільно високий рівень чутливості та специфічності в усьому діапазоні порогів прийняття рішень.

Модель LSTM також показала майже ідеальні результати, демонструючи плавну криву, що залишається близькою до верхнього лівого кута графіка, що свідчить про її сильні можливості до узагальнення. Модель CNN працює порівняно добре, хоча її крива демонструє незначні відхилення від оптимальної форми на початковому діапазоні FPR. Проте її траєкторія загалом підтверджує надійну якість класифікації. Модель RF показала помірно якісний ROC-профіль, однак її крива має більш східчастий характер, що можна пояснити її чутливістю до взаємодії ознак та розподілу даних.

Як видно, форма ROC-кривих варіюється в різних ділянках розподілу балів, проте ансамблеві моделі - насамперед Random Forest та XGBoost - виявляють найбільш наближені до оптимальних характеристики з плавними кривими, що тяжіють до значень, близьких до 1.

У порівнянні з цим, модель DT демонструє помітно нижчу ефективність. Її крива характеризується різкими змінами і розташована значно нижче кривих більш просунутих моделей, особливо у ділянках, де False Positive Rate (FPR) перевищує 0,2. Такий результат відображає обмеження базових структур дерев рішень у розпізнаванні складних меж класів у завданнях багатокласової класифікації.

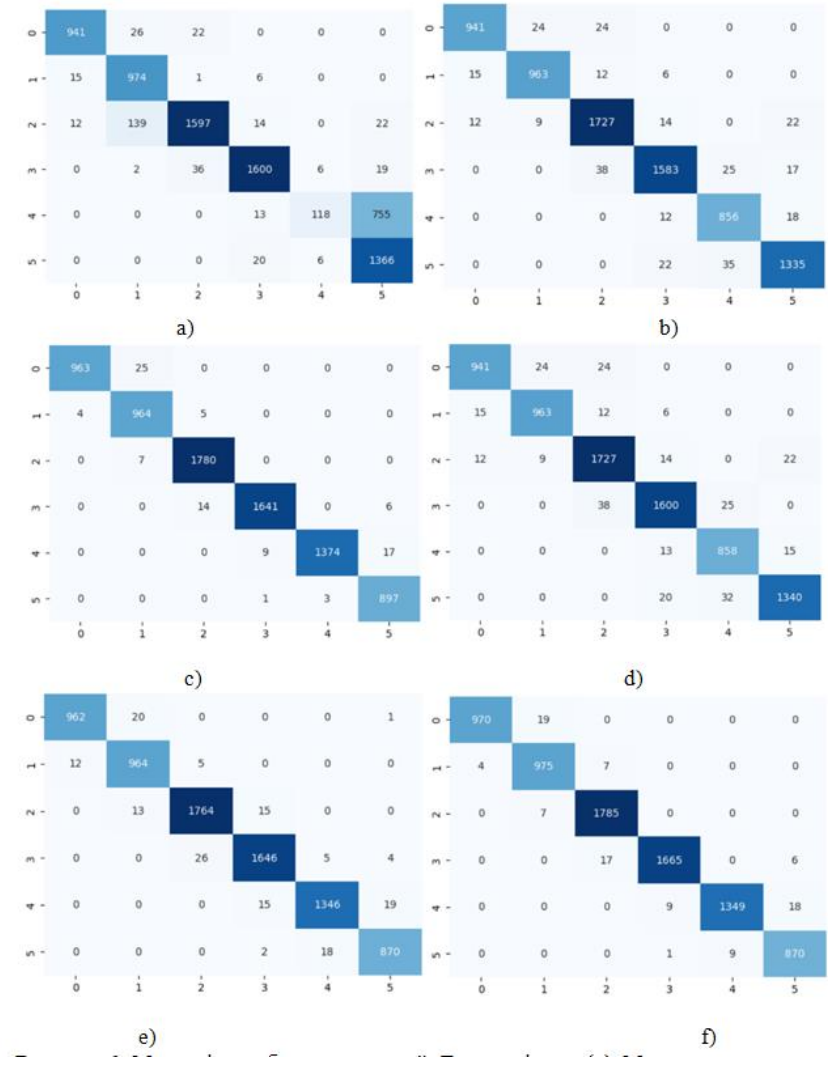


Рисунок 6. Матриці похибок для моделей: Дерево рішень (a), Метод опорних векторів (b), Випадковий ліс (c), XGBoost (d), Рекурентна нейронна мережа (e), Згорткова нейронна мережа (f)

Модель SVM показує нестабільний ROC-профіль із нерегулярним зростанням TPR, що свідчить про непослідовну продуктивність класифікації та знижену дискримінаційну здатність порівняно з ансамблевими або глибокими нейронними методами. Загалом результати підтверджують, що ансамблеві моделі, зокрема XGBoost, та глибокі нейронні архітектури, такі як LSTM та CNN, перевершують класичні методи за оцінкою на основі ROC-кривих. Ці моделі демонструють вищу здатність до узагальнення та підтримують збалансовану продуктивність між класами. Результати порівняльного аналізу метрик створених моделей в інтелектуальній системі представлені у таблиці 1.

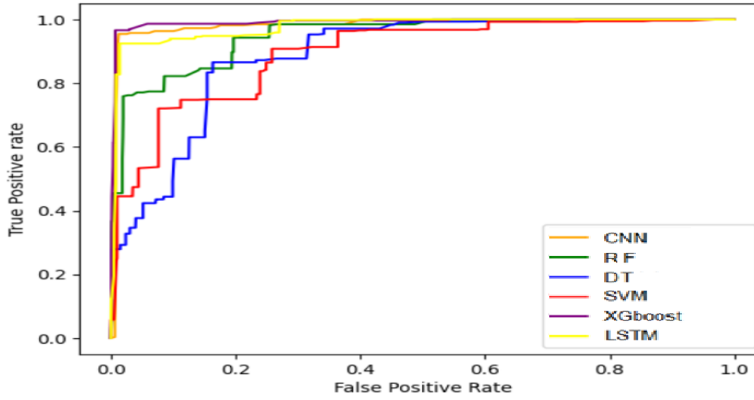


Рисунок 7. Узагальнена ROC-крива для створених моделей

Таблиця 2

Результати порівняльного аналізу метрик створених моделей

Metric	DT	SVM	RF	XGB	CNN	LSTM
Accuracy	0,879	0,961	0,973	0,993	0,987	0,981
Precision	0,861	0,957	0,967	0,991	0,981	0,973
Recall	0,856	0,958	0,969	0,989	0,977	0,969
F1-score	0,864	0,952	0,967	0,988	0,979	0,977
t_{avg}	3	18	7	8	45	67

Загалом, нейронні мережі демонструють високу точність класифікації. Модель LSTM досягає рівня точності, близького до 0,98, приблизно до 15-ї

епохи, тоді як CNN досягає аналогічного рівня лише після 25-ї епохи. Після цього подальше покращення точності значно сповільнюється, а іноді спостерігаються незначні коливання. Такі динаміки вказують на потенційний ризик перенавчання; однак обрані параметри регуляризації ефективно зменшують цю проблему. На початковому етапі навчання CNN демонструє більші значення помилки порівняно з LSTM. Незважаючи на це, її швидкість збіжності значно вища, що свідчить про більш ефективну траєкторію оптимізації, незважаючи на початкову варіативність.

Був проведений порівняльний аналіз моделей не лише за стандартними метриками оцінки, а й з урахуванням обчислювальної продуктивності. Крім точності класифікації, розглядався аспект обчислювальної складності шляхом вимірювання часу навчання моделей. Для цього було введено додатковий показник (tavg), що відображає середній час навчання у секундах.

4. Висновки

У результаті дослідження можна зазначити, що порівняльна оцінка реалізованих моделей продемонструвала явну перевагу ансамблевих та глибоких методів навчання. Класифікатор XGBoost досяг найвищої загальної точності — 0,993, з відповідними значеннями precision 0,991, recall 0,989 та F1-score 0,988. Серед моделей глибокого навчання CNN і LSTM також показали високий рівень продуктивності: CNN досягла точності 0,987, а LSTM - 0,981. Ці результати свідчать про здатність системи правильно класифікувати рівень ризику для здоров'я у понад 98% випадків при використанні найефективніших архітектур.

Окрім точності класифікації, була оцінена обчислювальна ефективність через середній час навчання моделей. Простіші моделі, такі як DT та RF, продемонстрували нижчий час навчання (відповідно 3 та 7 секунд), тоді як моделі глибокого навчання потребували більше ресурсів: CNN у середньому 45 секунд, а LSTM — 67 секунд. Ця різниця підкреслює типовий компроміс між складністю моделі та швидкістю виконання. Проте вища прогностична продуктивність моделей глибокого навчання виправдовує їх застосування, особливо у сценаріях, де критично важливі точність та надійність.

Аналіз надійності за допомогою матриць плутанини та ROC-кривих підтвердив стабільність прогнозів для всіх шести класів ризику. ROC-крива XGBoost послідовно наближалася до оптимальної форми, що відображає високу чутливість і специфічність. Подібно до цього, моделі LSTM та CNN демонстрували майже ідеальну поведінку з мінімальними відхиленнями. У той же час традиційні моделі, такі як DT та SVM, виявили обмеження продуктивності, особливо при обробці складних меж класів та підтриманні послідовності у різних ділянках розподілу балів.

З точки зору системної архітектури, поетапна структура значною мірою сприяла підвищенню надійності та інтерпретованості всього конвеєра. Використання інтерпретованих моделей на ранніх етапах дозволяло ефективно сегментувати дані та оцінювати порогові значення, тоді як глибокі

архітектури на пізніших етапах забезпечували захоплення просторово-часових залежностей і віддалених ефектів забруднювачів на здоров'я. Залучення зовнішніх наборів даних гарантувало вищу узагальнюваність і послідовність прогнозів моделей.

Перспективи подальшого розвитку включають інтеграцію механізмів уваги та трансформерів для покращення моделювання послідовностей, включення потоків даних із сенсорів у режимі реального часу та застосування федеративного навчання для оцінки ризику здоров'я з урахуванням приватності. Крім того, розширення функціоналу системи для персоналізованої оцінки ризику на рівні окремої особи та моделювання довгострокових епідеміологічних тенденцій ще більше підвищить її практичну цінність для планування громадського здоров'я та підтримки політики.

5. Література

1. Zhang, X., Liu, Y., Yang, H., & Sun, W. (2020). DeepAir: A hybrid deep learning framework for air quality forecasting. *Environmental Modelling & Software, 134***,** 104844. https://doi.org/10.1016/j.envsoft.2020.104844

2. Yao, S., Shen, M., Chen, Q., & Wang, Y. (2022). Deep learning for the prediction of population exposure to PM2.5 using remotely sensed data. *Environment International, 158***,** 106899. https://doi.org/10.1016/j.envint.2021.106899

3. Li, J., Yuan, J., & Wang, Z. (2020). Urban air pollution mapping using a deep learning approach and street-level images**.** *Environmental Pollution, 266***,** 115251. https://doi.org/10.1016/j.envpol.2020.115251

4. Fang, L., Yang, L., & Wang, C. (2019). Air quality index forecasting using a deep learning model based on 1D convolutional neural network and bidirectional long short-term memory. *Science of the Total Environment, 646,* 400–409. https://doi.org/10.1016/j.scitotenv.2018.07.400

5. Chen, F., Cao, S., Wang, Y., & Anwar, M. (2021). An interpretable machine learning framework for urban air quality analysis: A case study of PM2.5 in China. *Ecological Indicators, 124***,** 107447. https://doi.org/10.1016/j.ecolind.2021.107447

6. Rudnichenko, N., Vychuzhanin, V., Shibaeva, N., Petrov, I., & Otradskya, T. (2023). Intelligent data clustering system for searching hidden regularities in financial transactions. In *Proceedings of the 11th International Conference "Information Control Systems & Technologies" (ICST-2023)* (Vol. 3513, pp. 163–176). CEUR-WS.

7. Vychuzhanin, V., Rudnichenko, N., Sagova, Z., Smieszek, M., Cherniavskyi, V. V., Golovan, A. I., & Volodarets, M. V. (2019). Analysis and structuring diagnostic large volume data of technical condition of complex equipment in transport. *IOP Conference Series: Materials Science and Engineering, 776,* 012049. https://doi.org/10.1088/1757-899X/776/1/012049

8. Rudnichenko, N., Vychuzhanin, V., Petrov, I., & Shibaev, D. (2020). Decision support system for the machine learning methods selection in big data mining. In *Proceedings of the Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020)* (Vol. 2608, pp. 872–885). CEUR-WS.

9. Rudnichenko, N., Vychuzhanin, V., Otradskya, T., Petrov, I., & Shpinareva, I. (2023). Hybrid intelligent system for recognizing biometric personal data. In *Proceedings of the 3rd International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk 2022), co-located with the XXII International Scientific and Technical Conference on Information Technologies in Education and Management (ITEM 2022)* (Vol. 3422, pp. 74–85). CEUR-WS.

10. Lipatova, E., & Potapchenko, I. (2025). Machine learning models for air pollution health risk assessment: A comparative study. *Procedia Computer Science, 230*, 1–9. https://doi.org/10.1016/j.procs.2025.01.005

11. Chen, C.-H., Wang, C.-Y., & Wu, M.-Y. (2018). Machine learning techniques for short-term prediction of outpatient visits for respiratory diseases based on air pollution data. *Atmospheric Environment, 182*, 200–208. https://doi.org/10.1016/j.atmosenv.2018.04.051

12. Wang, Y., Lu, L., & Zhang, Q. (2018). Deep inferential spatial–temporal network for air pollution forecasting. *Neurocomputing, 322*, 204–213. https://doi.org/10.1016/j.neucom.2018.09.016

13. Lee, H., Kwon, S., & Kim, S. (2024). Rapid PM2.5-induced health impact assessment using a conditional U-Net CMAQ surrogate model. *Environmental Modelling & Software, 170*, 105643. https://doi.org/10.1016/j.envsoft.2023.105643

14. Zhang, J., Xu, Y., & Chen, L. (2022). PM2.5 exposure and health risk assessment using remote sensing and GIS: A case study in Shanghai. *Ecological Indicators, 140*, 109064. https://doi.org/10.1016/j.ecolind.2022.109064

15. Yao, S., Shen, M., Chen, Q., & Wang, Y. (2022). Deep learning for the prediction of population exposure to PM2.5 using remotely sensed data. *Environment International, 158*, 106899. https://doi.org/10.1016/j.envint.2021.106899

16. Bai, X., Zhang, N., Cao, X., & Chen, W. (2024). Prediction of PM2.5 concentration based on a CNN-LSTM neural network algorithm. *PeerJ, 12*, e17811. https://doi.org/10.7717/peerj.17811

17. Koçak, E. (2025). Comprehensive evaluation of machine learning models for real-world air quality prediction and health risk assessment by AirQ+. *Earth Science Informatics, 18*, 447–463. https://doi.org/10.1007/s12145-025-01941-7

18. Kaggle. (n.d.). *Air pollution image dataset from India and Nepal***.** Retrieved from https://www.kaggle.com/datasets/adarshrouniyar/air-pollution-image-dataset-from-india-and-nepal

19. U.S. Environmental Protection Agency. (n.d.). *Air Quality System (AQS)***.** ***Retrieved from https://www.epa.gov/aqs***

20. Figshare. (n.d.). *Air pollution data for Seoul***.** Retrieved from https://figshare.com/articles/dataset/CO_csv/12805643

INTELLIGENT HYBRID SYSTEM PROJECT FOR BIG DATA VOLUMES HEALTH HARM RISKS ANALYSIS

Ph.D. M. Rudnichenko[1[0000-0002-7343-8076]], **Ph.D. N. Shibaeva**[1[0000-0002-7869-9953]],
Ph.D. T. Otradskya[2[0000-0002-5808-5647]], **D. Shvedov**[1[0009-0002-4823-8782]],
Ph.D. I. Shpinareva[1[0000-0001-9208-4923]], **Dr.Sci. I.Petrov**[3[0000-0002-8740-6198]]
[1]Odessa Polytechnic National University, Ukraine,
[2]Odesa College of Computer Technologies "SERVER", Ukraine,
[3]National University "Odessa Maritime Academy", Ukraine
EMAIL: nickolay.rud@gmail.com, nati.sh@gmail.com, tv_61@ukr.net, firmn@gmail.com, frumle@ukr.net, iryna.shpinareva@onu.edu.ua

This paper presents a comprehensive study on the design and implementation of an intelligent system for analyzing and assessing population health risks associated with large-scale environmental data, with a focus on air pollution. The system leverages a hybrid architecture integrating machine learning and deep learning models to process heterogeneous datasets, including high-volume sensor measurements and satellite-derived indicators. The relevance of this research is underscored by the growing health threats posed by atmospheric pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, particularly in densely populated urban areas. The study emphasizes the necessity of employing intelligent data-driven systems to deliver accurate, scalable, and interpretable assessments of health risks at the population level. A critical review of existing methodologies highlights the limitations of traditional statistical approaches while demonstrating the advantages of neural networks and ensemble learning techniques in modeling complex nonlinear relationships and spatiotemporal dependencies within large, heterogeneous datasets. The proposed system incorporates a six-stage modeling framework, comprising decision trees, support vector machines, random forests, XGBoost, convolutional neural networks, and long short-term memory networks. Each stage contributes to a structured pipeline for data preprocessing, classification, prediction, and risk stratification. Training and validation were performed using the Air Pollution Image Dataset from India and Nepal, complemented by structured air quality datasets to enhance model generalizability. The system architecture follows a client-server design, with a Python and Flask backend and a JavaScript-based frontend, featuring an interactive analytical dashboard for data visualization, model inference, and comparative evaluation. Experimental results demonstrate the efficacy of ensemble and deep learning models, with a comparative analysis across accuracy, precision, recall, F1-score, and average training time confirming their superior performance in large-scale health risk assessment.

Keywords: *Intelligent data analysis, health analysis, big data, hybrid data mining, data analysis, neural networks, machine learning*