КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність – 122 «Комп'ютерні науки»,
освітня програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:
«Моделі та методи розробки інтелектуальної системи регіонального моніторингу кліматичних змін»

Студента 2 курсу групи ІАВ-21

Терновцев Віталій Олексійович
(прізвище, ім'я та по батькові)

_____
(підпис студента)

Науковий керівник:

доктор технічних наук, професор, професор кафедри технологій управління
(науковий ступінь, вчене звання)

Хлевна Юлія Леонідівна
(прізвище, ім'я, по батькові)

_____ _____
(дата) (підпис)

Попередній захист:

_____
(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління _____ _____ _____
(підпис) (прізвище, ініціали) (дата)

Київ – 2026

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122 – Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.
_____
«___» _____ 2025 року

ЗАВДАННЯ
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Терновцев Віталій Олексійович

Група ІАВ-21

1. Тема кваліфікаційної роботи

Моделі та методи розробки інтелектуальної системи регіонального моніторингу кліматичних змін

Затверджена наказом від «___» _____ 2026 р. № ___.

2. Строк подання студентом готової роботи – «___» _____ 2026 р.

3. Цільова установка та вихідні дані роботи

Підвищення ефективності регіонального моніторингу кліматичних змін шляхом розробки та дослідження інтелектуальної системи на основі ансамблевих моделей машинного навчання з інтеграцією багатовимірних супутникових даних NASA та глобальних кліматичних макротрендів. Вхідні дані: супутникові продукти місії VIIRS з просторовою роздільною здатністю 500 метрів, часові ряди концентрації діоксиду вуглецю мережі спостережень

NOAA Global Monitoring Laboratory, які охоплюють багаторічний період спостережень, ієрархічні масиви наукових даних у форматах HDF5 та NetCDF.

4. Зміст роботи

Робота містить вступ, аналіз предметної області та сучасних методів дистанційного зондування Землі, огляд існуючих систем моніторингу та обґрунтування концепції «Green AI» як етичного базису екологічних досліджень. Наведено математичну формалізацію вегетаційних індексів, обґрунтування вибору ансамблевих алгоритмів машинного навчання та методику синхронізації різнорідних даних за допомогою кубічної сплайн-інтерполяції 3-го порядку. Описано програмну реалізацію інтелектуального конвеєра обробки даних та результати експериментального дослідження регресійних моделей із глибоким аналізом метрик адекватності. Представлено архітектуру інтелектуальної системи з веб-інтерфейсом на базі фреймворку Streamlit, описано технологію візуалізації просторових аномалій та методи масштабування системи у хмарних середовищах за допомогою Dask. У висновках подано основні результати дослідження, а також список використаних джерел.

5. Перелік графічного матеріалу (слайдів)

Аномалія середньої температури суші та океану на глобальному рівні; результат перенавчання моделі; програмна реалізація методу семантичного пошуку каналів; програмна реалізація уникнення помилки ділення на нуль; програмна реалізація методу лінійної інтерполяції; програмна реалізація методу навчання моделі; обрана територія для проведення першого експерименту; динаміка вегетації у регіоні першого експерименту; прогнозний вплив макротрендів першого експерименту; обрана територія для проведення другого експерименту; динаміка вегетації у регіоні другого експерименту; прогнозний вплив макротрендів другого експерименту; оцінка якості моделі першого експерименту; оцінка якості моделі другого експерименту; блок-схема алгоритму роботи конвеєра даних; програмна

реалізація методу рентгену файлів; макет інтерфейсу; інтерфейс розробленої системи.

6. Перелік графічного матеріалу (слайдів)

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми дипломної роботи	3	01.10.2025	1.10.2025
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	27.12.2025	27.12.2025
3.	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	11	09.01.2026	08.01.2026
4.	Складання розгорнутого плану кваліфікаційної роботи	5	19.01.2026	19.01.2026
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін	4	19.01.2026-20.01.2026	20.01.2026
6.	Підготовка розділу 1 «Аналіз предметної області та сучасних методів моніторингу»	12	11.02.2026	11.02.2026
7.	Підготовка розділу 2 «Формалізація методів аналізу даних у проєкті»	13	05.03.2026	05.03.2026
8.	Підготовка розділу 3 «Інтелектуальний аналіз та предиктивне моделювання кліматичних даних»	14	03.04.2026	03.04.2026
9.	Підготовка розділу 4 «Програмна реалізація та технологія впровадження інтелектуальної системи»	12	18.04.2026	18.04.2026
10.	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	15	30.04.2026	30.04.2026
11.	Передача кваліфікаційної роботи науковому керівникові	2	06.05.2026	06.05.2026
12.	Передача кваліфікаційної роботи рецензенту для рецензування	2	08.05.2026	08.05.2026
13.	Попередній захист кваліфікаційної роботи	5	11.05.2026	11.05.2026

Дата видачі завдання «___» _____ 2026 р.

Керівник роботи Доктор наук, доцент кафедри технологій управління Доктор технічних наук, професор кафедри технологій управління Хлевна Ю. Л.

_____

(підпис)

Завдання прийняв до виконання студент групи ІАВ-21 Терновцев Віталій Олексійович.

_____

(підпис)

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 - Комп'ютерні науки,

освітня програма "Інформаційна аналітика та впливи"

Дипломна робота магістра Терновцева Віталія Олексійовича.

Тема роботи – «Моделі та методи розробки інтелектуальної системи регіонального моніторингу кліматичних змін».

Мета дипломної роботи магістра – підвищення ефективності регіонального моніторингу кліматичних змін шляхом розробки та впровадження інтелектуальної системи та методики, котрі базуються на автоматизованому зборі, обробці та предиктивному моделюванні багатовимірних супутникових даних.

Об'єкт дослідження – процеси автоматизованого дистанційного моніторингу регіональних кліматичних змін та динаміки стану рослинного покриву.

Предмет дослідження – методи, алгоритми інтелектуального аналізу великих масивів даних та регуляризовані моделі машинного навчання, які застосовуються для інтерпретації супутникової спектральної інформації.

Наукова новизна роботи – розроблено концепцію інтелектуального кліматичного моніторингу, яка відрізняється від існуючих підходів впровадженням алгоритму математичної синхронізації супутникових часових рядів із глобальними макротрендами через апарат лінійної інтерполяції, а також застосуванням структурно регуляризованих ансамблевих моделей.

У роботі досліджуються сучасні підходи до використання методів Data Science та дистанційного зондування Землі у задачах дослідження глобального

потепління. Розроблено архітектуру інтелектуального конвеєра для автоматизованої декомпозиції ієрархічних супутникових архівів. Обґрунтовано доцільність застосування ансамблевих алгоритмів машинного навчання для виявлення кореляцій між локальними вегетаційними індексами та глобальними концентраціями CO2. Наводяться рекомендації щодо практичного впровадження системи у сфері державного екологічного моніторингу та точного землеробства.

Дипломна робота складається зі вступу, основної частини, яка включає три розділи, висновків та списку використаних джерел. Всього налічує 103 сторінки та перелік посилань з 70 джерел на 6 сторінках.

Ключові слова: кліматичні зміни, супутникові дані NASA, інтелектуальний аналіз даних, машинне навчання, NDVI, Random Forest, конвеєр даних.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ І СКОРОЧЕНЬ

CART – Classification and Regression Trees

EOS – Earth Observing System

ETL – Extract, Transform, Load

EVI – Enhanced Vegetation Index

GCM – Глобальні кліматичні моделі

GEE – Google Earth Engine

GES DISC – Goddard Earth Sciences Data and Information Services Center

HDF – Hierarchical Data Format

IPCC – Міжурядова група експертів зі зміни клімату

LAI – Leaf Area Index

MAE – Mean Absolute Error

MBE – Mean Bias Error

MDI – Mean Decrease Impurity

MODIS – Moderate Resolution Imaging Spectroradiometer

MTV – Model-Template-View

NDVI – Normalized Difference Vegetation Index

NetCDF – Network Common Data Form

NOAA – Національне управління океанічних і атмосферних досліджень США

OOB – Out-of-Bag

SAVI – Soil Adjusted Vegetation Index

VIIRS – Visible Infrared Imaging Radiometer Suite

ВМО – Всесвітня метеорологічна організація

ГІС – Геоінформаційні технології

ДЗЗ – Дистанційне зондування Землі

ВСТУП

У сучасному світі комплексна проблема глобальних та регіональних змін клімату щодня набуває дедалі більшої актуальності, оскільки її деструктивні наслідки відчутно проявляються у зростанні частоти екстремальних погодних явищ, зміні гідротермічного режиму територій, підвищенні середньої температури приземного шару атмосфери, прогресуючій деградації вразливих природних екосистем. Водночас із бурхливим розвитком методів інтелектуального аналізу даних та аерокосмічних технологій дистанційного зондування Землі відкриваються можливості для отримання високодеталізованої інформації про стан підстильної поверхні в режимі реального часу. Сучасні супутникові місії генерують великі масиви багатовимірної просторово-часової інформації, але складна ієрархічна структура цих бінарних архівів, наявність атмосферних зашумлень та значна просторова розрізненість вимагають розробки та впровадження новітніх підходів інтелектуального аналізу, здатних повністю автоматизувати процеси виявлення прихованих аномалій і прогнозування довгострокових екологічних трендів.

Актуальність теми зумовлена гострою потребою у створенні ефективних, автоматизованих інструментів для дослідження регіональних кліматичних процесів на основі інтеграції супутникових знімків та глобальних макротрендів. У такий період швидких екологічних трансформацій надважливо розробляти аналітичні системи, які можуть не лише фіксувати поточний стан фітоценозів, але й формувати точні прогнози за допомогою алгоритмів машинного навчання, виявляючи складні нелінійні взаємозв’язки між накопиченням парникових газів та реакцією біосфери.

Метою наукової роботи є підвищення ефективності регіонального моніторингу кліматичних змін шляхом розробки та впровадження інтелектуальної системи та методики, котрі базуються на автоматизованому зборі, обробці та предиктивному моделюванні багатовимірних супутникових даних.

Для досягнення поставленої мети вирішуються *наступні завдання*:

- здійснити систематичний аналіз практичної сфери застосування методології Data Science, сучасних моделей машинного навчання в задачах глобального та регіонального моніторингу;
- спроєктувати модульну архітектуру та алгоритмічне забезпечення автоматизованої системи для потокового програмованого збору та декомпозиції багатовимірних супутникових продуктів високої роздільної здатності у межах відкритих програмних інтерфейсів;
- побудувати, оптимізувати та регуляризувати предиктивну модель машинного навчання для точного виявлення та прогнозування довгострокового нелінійного впливу атмосферних макротрендів на регіональні рослинні зони;

Об'єктом дослідження є процеси автоматизованого моніторингу регіональних кліматичних змін, оцінки стану навколишнього середовища та просторово-часової динаміки рослинного покриву за допомогою аерокосмічних супутникових систем.

Предметом дослідження є методи, алгоритми інтелектуального аналізу великих масивів даних та регуляризовані моделі машинного навчання, які застосовуються для автоматизованої обробки та інтерпретації супутникової інформації.

Наукова новизна роботи полягає у розробці та впровадженні комплексного об'єктно-орієнтованого підходу до інтеграції та моделювання різнорідних екологічних даних великого обсягу.

Удосконалено алгоритмічний підхід до попередньої обробки багатовимірних бінарних архівів наукових даних за рахунок розробки алгоритму семантичного рекурсивного обходу вкладених структур ієрархічних файлів. Подальший розвиток застосування методів машинного навчання, архітектуру якого було цілеспрямовано регуляризовано шляхом введення жорстких структурних обмежень на топологію базових дерев рішень.

Практичне значення полягає в тому, що створена інтелектуальна система є повністю завершеним, автономним програмним рішенням, здатним ефективно функціонувати з великими супутниковими масивами в режимі реального часу. Розроблені алгоритми та спроєктований інтерактивний користувацький дашборд можуть бути безпосередньо інтегровані у діяльність аналітичних центрів Міністерства захисту довкілля та природних ресурсів України.

Окрім цього, розроблені програмні модулі просторової агрегації та детекції гідротермального стресу рослинності мають високий потенціал для масштабування та практичного впровадження у комерційний сектор точного землеробства з метою оперативного прогнозування ризиків зниження врожайності агрокультур, а також можуть слугувати об’єктивним цифровим ядром для незалежної верифікації здатності лісових масивів поглинати вуглець.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА СУЧАСНИХ МЕТОДІВ МОНІТОРИНГУ

1.1. Феномен кліматичних змін та стратегії регіонального екологічного моніторингу

У XXI столітті феномен кліматичних змін перетворився із суто наукової проблеми на один з найсуттєвіших викликів для поточного розвитку людства. Сучасні кліматичні процеси можна охарактеризувати не тільки поступовим зростанням середньої глобальної температури, але й значною дестабілізацією кліматичної системи, прояв якої спостерігається у вигляді безпрецедентних кліматичних аномалій. Згідно з Шостим оціночним звітом Міжурядової групи експертів зі зміни клімату (IPCC), антропогенний вплив на атмосферу, океани та біосферу є беззаперечним, а масштаби останніх трансформацій у кліматичній системі є небаченими протягом багатьох століть і навіть тисячоліть [1].

Основною причиною таких аномалій станом на сьогодні залишається безперервне зростання концентрації парникових газів, зокрема діоксиду вуглецю та метану. Глобальні макротренди, які фіксують міжнародні дослідницькі центри, такі як Національне управління океанічних і атмосферних досліджень США (NOAA), вказують на експоненційне зростання рівня CO2, а це безпосередньо корелює з підвищенням теплоємності нижніх шарів атмосфери [2]. Таким чином порушується радіаційний баланс Землі. Як наслідок, виникають просторово-часові зсуви у режимах випадання опадів, збільшується частота та інтенсивність екстремальних погодних явищ: тривалих посух, теплових хвиль, аномальних заморозків та повеней.

Важливою характеристикою сучасних кліматичних аномалій є їхня просторова гетерогенність. Незважаючи на те, що глобальне потепління має планетарний масштаб, його наслідки розподіляються вкрай нерівномірно. Деякі регіони зазнають екстремального опустелювання, тоді як інші навпаки страждають від надмірного зволоження. Глобальні кліматичні моделі (GCM)

здатні прогнозувати загальнопланетарні тренди, проте вони часто не володіють достатньою просторовою роздільною здатністю для точного передбачення локальних екологічних криз. Конкретно з цих причин виникає наукова та практична необхідність у переході від глобальних оцінок до розробки стратегій регіонального екологічного моніторингу, які будуть здатні фіксувати зміни на рівні окремо взятих екосистем.

Регіональний екологічний моніторинг представляє собою комплексну систему спостережень, збору, обробки та аналізу інформації щодо стану довкілля на чітко визначеній території із метою оцінки та прогнозування його майбутніх змін. Традиційні методи моніторингу, які переважно базуються на мережі наземних метеорологічних станцій, виявляються недостатніми в умовах кліматичної нестабільності. Наземні станції хоч і забезпечують високу точність точкових вимірювань, усе одно не можуть достатньою мірою забезпечити суцільне просторове покриття, особливо коли мова йде про важкодоступні регіони, чим ускладнюється побудова цілісної картини екологічної динаміки.

Сьогодні парадигма регіонального моніторингу спирається на інтеграцію наземних спостережень із даними дистанційного зондування Землі (ДЗЗ). Використання супутникових спостережень, як зазначають у вітчизняних дослідженнях, є вкрай важливим інструментом для відстеження цілей сталого розвитку, оскільки дозволяє отримувати об’єктивну, оперативну та просторово розподілену інформацію про стан екосистем у наближеному до реального часу режимі [3].

У межах стратегії такого екологічного спостереження передбачається побудова інтелектуальних систем, здатних автоматично завантажувати великі обсяги супутникових даних, обробляти багатовимірні масиви просторової інформації та в кінцевому результаті поєднувати їх із глобальними макрокліматичними показниками. Завдяки подібному підходу вдається не лише констатувати факт наявності аномалії, але й за допомогою методів машинного навчання виявляти приховані нелінійні взаємозв’язки між

глобальним зростанням рівня CO2 та деградацією локального рослинного покриву.

Генеральна Асамблея ООН у 2015 році ухвалила Порядок денний у сфері сталого розвитку до 2030 року. У нього входять 17 цілей сталого розвитку [4]. Розробка інструменту для екологічного моніторингу може стати безпосереднім способом реалізації щонайменше двох цілей: пункт 13 (Боротьба зі зміною клімату) та пункт 15 (Збереження екосистем суші).

Розроблювана інтелектуальна система напряму відповідає задачі 13.1, у межах якої вимагається посилення стійкості та адаптаційної здатності до небезпечних кліматичних явищ. Перехід від реактивної до проактивної кліматичної політики цілком може статись завдячуючи використанню машинного навчання для виявлення неочевидних кореляцій між глобальними рівнями CO2 та локальною вегетацією. Побудовані предиктивні моделі також створюють наукове підґрунтя для розробки стратегій адаптації на рівні конкретних регіонів.

Завдання 15.3 закликає до боротьби з опустелюванням та до відновлення деградованих земель. Оперативна фіксація зменшення площі та зниження фотосинтетичної активності лісових і степових масивів можуть бути здійснені безпосередньо через автоматизований аналіз багатовимірних супутникових масивів, за рахунок чого дослідникам легше вчасно реагувати на втрату біорізноманіття та деградацію ґрунтів.

Загалом, впровадження інформаційних систем, які агрегують дані з супутника задля розрахунку показників цілей сталого розвитку, це цілком важливий крок для відстеження виконання міжнародних зобов’язань. Подібні системи уможливлюють створення стандартизованої та відтворюваної методології оцінки стану навколишнього середовища, яка не залежить від фрагментованих наземних спостережень.

У контексті регіонального моніторингу особлива увага приділяється вегетації, тобто стану рослинного покриву. Рослинність, очевидно, є ключовою складовою земної біосфери, бо виконує роль головного індикатора

екологічного та кліматичного здоров’я певної території. Фітоценози надзвичайно чутливі до змін температури, до доступності вологи та до хімічного складу атмосфери. Відповідно, будь-які кліматичні потрясіння на кшталт посух або температурних екстремумів негайно відображаються на біофізичних характеристиках рослин. Значення рослинності полягає також у її ролі в глобальному вуглецевому циклі. Лісові масиви традиційно виступають найважливішими поглиначами вуглецю, оскільки пом’якшують таким чином наслідки антропогенних викидів. Однак, існує тривожна тенденція, на яку вказують дослідження NASA: здатність тропічних лісів та інших великих екосистем поглинати діоксид вуглецю поступово знижується внаслідок зростання глобальної температури та деградації ґрунтів [5]. Таке погіршення асиміляційного потенціалу вегетації утворює петлю позитивного зворотного зв’язку, коли менше поглинання CO2 призводить до ще більшого його накопичення в атмосфері, чим, у свою чергу, посилюється парниковий ефект.

З огляду на це, безперервний просторовий аналіз динаміки розвитку рослинності через супутникові системи дистанційного зондування грає важливу роль в оцінці регіональної вразливості до кліматичних змін. Новітні оптичні сенсори дозволяють фіксувати недоступні людському оку зміни у стані біосфери для оцінки реакції екосистем на стрес у масштабах цілих континентів.

Процес переходу від концептуального розуміння екологічних процесів до подальшого створення автоматизованих систем передбачення вимагає формалізації спостережень [6]. Для того щоб комп’ютерні алгоритми та моделі машинного навчання мали можливість аналізувати стан біосфери, загальнобіологічні характеристики, такі як здоров’я, біомаса або стрес, мають бути переведені у строгі математичні величини за допомогою спеціалізованих вегетаційних індексів.

Особливої актуальності проблема регіонального моніторингу набуває також і для території України, яка знаходиться у зоні ризикованого землеробства. Загальносвітові кліматичні макротренди безпосередньо

впливають на агрокліматичне зонування країни. Стійку тенденцію до зміщення кліматичних зон можна відстежити протягом останніх десятиліть: межа між степом та лісостепом невпинно зміщується на північ [7]. Відповідно, південні та східні регіони, які традиційно є основними житницями держави, дедалі частіше стикаються з проблемою аридизації та посиленням частоти ґрунтових і атмосферних посух.

Аналіз кліматичних даних за останні десятиліття вказує на помітне відхилення середньорічних температур та кількості опадів від кліматичної норми: базового періоду 1961-1990 років або більш пізніх референсних значень.

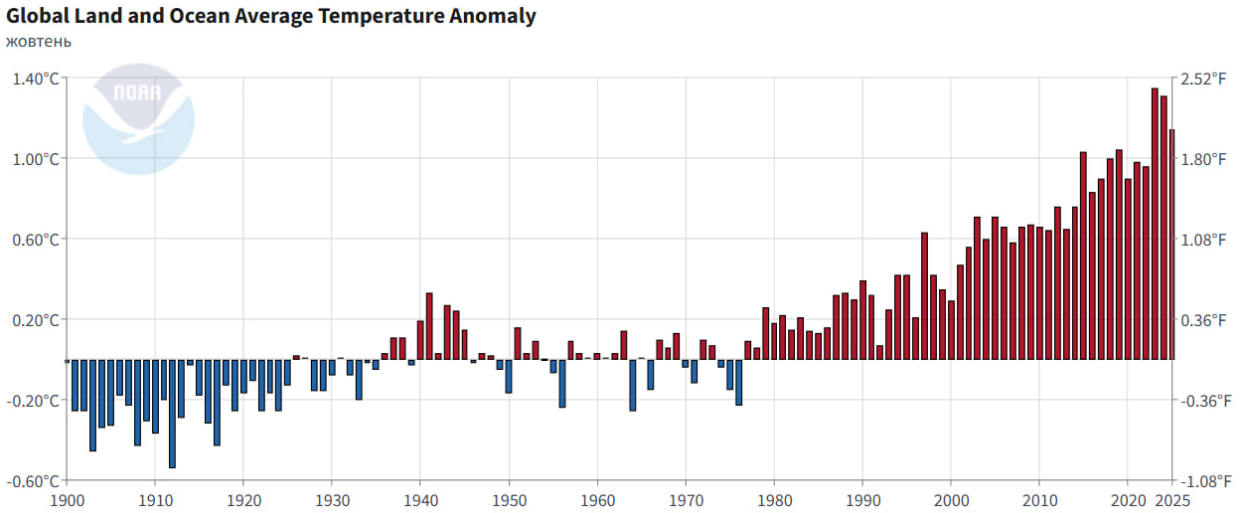

Рисунок 1.1 – Аномалія середньої температури суші та океану на глобальному рівні [2]

Графіки аномалій приземної температури (рис. 1.1) чітко ілюструють феномен потепління. Замість природних флуктуацій, себто періодичного чергування холодних та теплих років, існує стійкий тренд на формування виключно позитивних температурних аномалій. Особливої вираженості ця тенденція набула протягом останнього десятиріччя, коли кожний наступний рік нерідко ставав теплішим за попередній, щоразу встановлюючи нові історичні максимуми. У поєднанні зі зміною режиму опадів (або зменшенням їх кількості, або зміщенням їхнього розподілу на користь короткочасних злив),

такі температурні стрибки утворюють усе жорсткіші умови для розвитку вегетації.

У зоні степу (Миколаївська, Херсонська, Одеська, Запорізька області) ці явища призводять до критичного дефіциту продуктивної вологи у ґрунті під час найважливіших фаз розвитку сільськогосподарських культур. У зоні лісостепу (Полтавська, Вінницька, Черкаська, Київська області) теплові хвилі та нестача опадів спричиняють передчасне всихання рослинності та зниження врожайності.

Розроблювана інтелектуальна система вирішує проблему своєчасного виявлення перелічених вище явищ. Вона дозволяє:

1. Локалізувати аномалію шляхом аналізу багатовимірних масивів. Система може точно визначити географічні координати територій у стані найбільшого гідротермічного стресу.
2. Динамічний розрахунок та подальша агрегація вегетаційних індексів дає кількісно оцінити ступінь пошкодження фітоценозу. Різке падіння значень індексу, наприклад, у період, який за фенологічним календарем має відповідати піку вегетації, є чітким маркером посухи.
3. Завдяки використанню методів машинного навчання досліджується кореляція між історичними температурними аномаліями, глобальними трендами та станом рослинного покриву. Це шлях не лише до констатації факту засухи, а й до прогнозування швидкості деградації екосистем.

Таким чином, розробка спеціалізованої інформаційної технології на стику Data Science та екології є необхідним кроком для забезпечення сталого розвитку України та створення ефективних систем підтримки прийняття управлінських рішень в умовах кліматичних змін.

1.2. Дистанційне зондування Землі: супутникові місії та продукти NASA

Сьогодні основою побудови будь-якої інтелектуальної системи екологічного та кліматичного моніторингу є надійні, безперервні, просторово розподілені масиви даних. ДЗЗ з космосу станом на зараз є безальтернативним методом отримання потрібної інформації як глобального, так і регіонального масштабів. Національне управління з аеронавтики і дослідження космічного простору США (NASA) в рамках програми Earth Observing System (EOS) забезпечує наукову спільноту відкритим доступом до петабайтів мультиспектральних даних [8]. Власне, завдяки ним і з'являється можливість проводити глибокий ретроспективний аналіз стану біосфери.

Фундаментальний принцип оптичного дистанційного зондування базується на тому, що різні фізичні об'єкти на поверхні Землі по-різному поглинають, пропускають чи відбивають електромагнітне випромінювання Сонця. Залежність інтенсивності відбитого випромінювання від довжини хвилі називається спектральною сигнатурою або кривою відбиття об'єкта [9]. У контексті Data Science ця спектральна сигнатура виступає унікальним «цифровим відбитком». З його використанням алгоритми машинного навчання класифікують типи підстильної поверхні та виявляють аномалії.

Таблиця 1.1. – Фізико-хімічні особливості формування спектральних сигнатур для основних типів поверхонь.

Тип поверхні	Фізико-хімічні особливості
Гідрологічні об'єкти	Чиста вода поглинає майже всю енергію в ближньому та середньому інфрачервоному діапазонах спектра. Незначне відбиття близько 5-10% можна проспостерігати лише у видимому діапазоні, переважно у синій та зеленій його частинах. Власне, саме тому на інфрачервоних супутникових знімках водні об'єкти зазвичай виглядають абсолютно чорними. Якщо вода містить суспензії, фітопланктон або цвітіння

	водоростей, її відбивальна здатність у зеленому та червоному діапазонах різко зростає, і алгоритми фіксують забруднення.
Ліси, агроценози	Сигнатура рослинності є найбільш специфічною та контрастною. У видимому діапазоні (400-700 нм) хлорофіл та інші пігменти, наприклад, каротиноїди або ксантофіли, інтенсивно поглинають синє та червоне світло для потреб фотосинтезу, залишаючи невеликий пік відбиття у зеленій зоні. Однак при переході у більш близький інфрачервоний діапазон (700-1300 нм) відбувається сильний стрибок відбивальної здатності до 40-50%. Це явище, відоме як red edge, зумовлене тим, що внутрішня клітинна структура губчастого мезофілу листка розсіює інфрачервону енергію, аби запобігти термічному пошкодженню тканин рослини.
Голий ґрунт та пісок	Спектральна крива ґрунтів характеризується відносно плавним та монотонним зростанням відбивальної здатності від видимого до середнього інфрачервоного діапазону. На відміну від рослинності, тут немає різких піків чи провалів. Загальний рівень відбиття сильно залежить від трьох факторів: вологості, тобто чим вологіший ґрунт, тим темнішим він стає для сенсора через поглинання водою, вмісту органічної речовини та мінералогічного складу. Пісок пустель у свою чергу має дуже високе відбиття у всьому оптичному спектрі, він один із найяскравіших природних об’єктів на знімках.
Бетон та урбанізовані території	Матеріали антропогенного походження мають плоску спектральну сигнатуру з високим рівнем відбиття і у видимому, і в інфрачервоному діапазонах. Бетон, зокрема, рівномірно відбиває енергію, тому його сигнатура подібна на світлі до ґрунтів, щоправда без характерних провалів поглинання води, притаманних природним ландшафтам.

Саме ця різниця у фізиці взаємодії світла з матерією (табл. 1.1) лежить в основі розробки математичних індексів, де алгоритми конвеєра аналізують співвідношення між зонами максимального поглинання та максимального відбиття.

Історично головним інструментом для моніторингу поверхні Землі став спектрорадіометр MODIS (Moderate Resolution Imaging Spectroradiometer), який був встановлений на борту супутників Terra та Aqua, запущених у 1999 та 2002 роках відповідно. Прилад здійснює зйомку поверхні планети у 36 спектральних діапазонах, а це цілком широкий спектр інформації: від температури підстилаючої поверхні до динаміки хмарного покриву та стану океанів. Для завдань, метою яких є відслідковування рослинності, еталонним продуктом є масив даних MOD13Q1 [10]. Він є композитом, де міститься результат відбору найкращих безхмарних пікселів за 16 діб з просторовою роздільною здатністю 250 метрів. Достатньо тривалий час роботи місії MODIS призвів до накопичення унікального архіву даних, який охоплює понад два десятиліття [11]. Очевидно, для даного дослідження довгострокових кліматичних трендів він є просто незамінним.

Однак, варто зазначити, що з огляду на поступове завершення життєвого циклу супутників Terra та Aqua, NASA ініціювало перехід на апаратуру нового покоління – VIIRS (Visible Infrared Imaging Radiometer Suite). Цей прилад розміщений на супутниках Suomi NPP та NOAA-20. VIIRS розроблявся як еволюційне продовження MODIS, тому він так само забезпечує неперервність кліматичних спостережень, але оснащений покращеними радіометричними характеристиками та ширшою смугою огляду.

У межах розробки інтелектуальної системи абсолютно доцільним є використання саме продукту VNP13A1, це аналог MOD13Q1 від сенсора VIIRS. Набір даних аналогічно надає 16-денні композитні зображення індексів вегетації з вдвічі більшою просторовою роздільною здатністю, тобто у 500 метрів. Перехід на використання даних VIIRS в системах можна обґрунтувати не тільки їхньою актуальністю, адже вони є оптимізованою структурою

зберігання багатовимірних масивів, а це значно підвищить швидкість автоматизованої обробки величезних об'ємів даних. Ширша смуга огляду також гарантує відсутність сліпих зон на екваторі та надає консистентнішу якість пікселів вздовж усієї ширини знімка (табл. 1.2).

Таблиця 1.2 – Порівняння характеристик супутників

Характеристика	MODIS	VIIRS
Платформи	Terra, Aqua	Suomi NPP, NOAA-20, NOAA-21
Рік початку експлуатації	1999	2011
Кількість спектральних каналів	36 каналів	22 канали (оптимізовані)
Частота прольоту	1-2 дні	1 день (ширша смуга захвату)
Ширина смуги огляду	2330 км	3040 км
Просторова роздільна здатність	250 м	375 м (Imagery bands), 500 м (у композитах)
Продукт для аналізу вегетації	MOD13Q1	VNP13A1
Оптичні спотворення на краях сканування	Значні, збільшення пікселя до розмірів 1 км	Мінімізовані, апаратна агрегація пікселів

Оптичне зондування рослинності базується і на специфічній взаємодії електромагнітного випромінювання Сонця з анатомічною та біохімічною структурою листків. Для коректного розрахунку вегетаційних індексів вкрай важливо зважати на два спектральні діапазони: видимий червоний та ближній інфрачервоний.

У червоному каналі, довжина хвилі якого сягає приблизно 600-700 нм, спостерігається інтенсивне поглинання фотонів хлорофілом (пігментами a та b). Хлорофіл, як відомо, є головним каталізатором процесу фотосинтезу. Відповідно, чим здоровіша та щільніша рослинність, тим більше червоного світла вона поглинає і, у свою чергу, тим менше цього самого світла відбивається назад у космос до сенсора супутника [6].

Довжина хвилі ближнього інфрачервоного каналу становить приблизно 700-1000 нм. Електромагнітні хвилі такої довжини зовсім не використовуються рослинами для фотосинтезу. Натомість, як вже було згадано раніше у цьому розділі, губчаста тканина мезофілу здорового листка діє як своєрідний відбивач [12]. Вона розсіює інфрачервону енергію, щоб запобігти перегріву рослини. Тому, логічно, високі значення відбиття у каналі NIR свідчать про непошкоджену внутрішню структуру клітин.

Алгоритми розробленої системи автоматично сканують завантажені супутникові масиви та з допомогою розумного пошуку ідентифікують усі ключові канали серед десятків інших непотрібних змінних. У зв’язку зі специфікою номенклатури NASA, системі треба шукати ті змінні, де містяться ключові слова, на кшталт red reflectance, reflectance_01, nir reflectance, reflectance_02 тощо, після чого виконуються математичні операції над матрицями для отримання фінальних кліматичних індикаторів.

Важливо також зазначити, що супутникові знімки у сфері кліматології кардинально відрізняються від звичайних графічних зображень, таких як JPEG або PNG. Один файл або гранула містить не просто картинку, а складний багатовимірний масив точних фізичних величин, куди входять відбивальна здатність, кути нахилу сонця, матриці оцінки якості пікселів тощо. Вони у свою чергу прив’язані до географічних координат та точного часу. Для ефективнішого зберігання настільки важких даних часто використовують спеціалізовані наукові формати HDF (Hierarchical Data Format) та NetCDF (Network Common Data Form) [13].

Дані супутників, зокрема місії VIIRS, станом на сьогодні розповсюджуються у форматі HDF5. Згідно з офіційною документацією, цей формат має внутрішню ієрархічну структуру, яка концептуально нагадує файлову систему комп’ютера [14]. Таким чином, файл складається з груп та наборів даних, тобто самих матриць з цифрами. Для автоматизованої обробки подібних файлів у середовищі Python потрібно розробити складний інженерний підхід. У кінці система повинна бути здатною здійснювати

парсинг внутрішньої структури самого файлу. Наприклад, у продуктах VIIRS безпосередні наукові дані приховані у глибокій групі «HDFEOS/GRIDS/VIIRS_Grid_16Day_VI_500m/Data Fields». Створений у системі конвеєр за допомогою спеціалізованих низькорівневих бібліотек h5py та рушія netcdf4 в екосистемі xarray автоматично обходить цю ієрархію, знаходить багатовимірні просторові матриці та завантажує їх у пам’ять для подальшого просторового усереднення та машинного навчання [15].

Формат NetCDF так само широко використовується в кліматології як стандартизований контейнер для зберігання кліматичних моделей. Особливої уваги варто приділити його сучасній версії NetCDF4, яка під капотом базується на технології HDF5. Інтелектуальна система не залежить від конкретного джерела даних і гнучко адаптуватися до будь-яких просторових інформаційних продуктів NASA якраз завдячуючи здатності програмного забезпечення універсально зчитувати усі ці споріднені формати.

1.3. Аналіз існуючих інформаційних систем моніторингу та їх недоліків

Розвиток технологій збору та обробки даних очікувано призвів до появи різноманітних інформаційних систем моніторингу для багатьох галузей. Однак, незважаючи на такий значний прогрес у галузі геоінформаційних технологій (ГІС) та аналітики великих даних, переважна більшість існуючих рішень стикається з низкою обмежень. Для обґрунтування архітектури розроблюваної інтелектуальної системи необхідно детально проаналізувати традиційні підходи до моніторингу, існуючі хмарні платформи та виявити їхні критичні недоліки, особливо в галузі збору первинних даних та вільного інтегрування різнорівневих кліматичних показників у моделі машинного навчання.

Базовим джерелом даних для кліматології історично є мережа наземних стаціонарних метеостанцій. Вони функціонують під егідою Всесвітньої метеорологічної організації (ВМО) та національних гідрометеорологічних

служб та забезпечують постійний збір інформації про температуру повітря, про кількість опадів, про вологість, про швидкість вітру та інші фізичні параметри атмосфери [16].

Наземні станції володіють головною перевагою – високою точністю безпосередніх вимірювань та хорошою часовою роздільною здатністю. Часова роздільна здатність це, грубо кажучи, можливість фіксувати зміни параметрів щогодини або навіть щохвилини. Тим не менш, такі станції мають і критичні недоліки, пов’язані з просторовою репрезентативністю даних. Метеостанція фіксує кліматичні параметри лише в конкретній точці простору. Для отримання картини по всьому регіону дані екстраполюються та інтерполюються, а це у підсумку призводить до математичних похибок. Екосистеми в принципі є вкрай гетерогенними, температура поверхні або рівень вологості ґрунту можуть суттєво відрізнятися на відстані кількох кілометрів через різницю в рельєфі, типі ґрунтів чи наявності лісових масивів [17].

Також мережа метеостанцій розташована дуже нерівномірно. Більшість із них сконцентрована поблизу населених пунктів, аеропортів або транспортних магістралей, де відсутні величезні масиви лісів, гірські райони та аграрні угіддя. Тому вони залишаються без належного прямого спостереження. Додатково стаціонарні станції вимірюють метеорологічні показники, але не здатні самі надати пряму інформацію про біофізичний стан рослинності, наприклад, рівень хлорофілу, біомасу чи ступінь гідротермічного стресу фітоценозів.

Альтернативою та навіть хорошим доповненням до наземних мереж можуть бути системи супутникового моніторингу, які спираються на дані ДЗЗ. Це вже згадані раніше інформаційні системи на базі даних місій NASA, які пропонують принципово інший підхід – перехід від точкового до суцільного або ж піксельного просторового покриття. Ефективність використання цих даних, щоправда, безпосередньо залежить від обраного програмного інструментарію.

Як відповідь на експоненційне зростання обсягів супутникових даних провідні технологічні корпорації та космічні агентства розробили низку хмарних платформ для аналізу геопросторової інформації. Найбільш відомими та часто використовуваними є Google Earth Engine (GEE), NASA Giovanni та Sentinel Hub [18]. Вони безумовно потужні, хоча детальний їхій аналіз виявляє певні обмеження, через які ці платформи стають неоптимальними для виконання специфічного завдання даного проєкту – створення автономного конвеєра даних з інтеграцією локальних моделей машинного навчання.

Google Earth Engine – це хмарна платформа для планетарного аналізу геопросторових даних із петабайтами супутникових знімків, куди входять також і архіви MODIS та VIIRS [19]. Платформа надає API на базі JavaScript та Python. Найголовнішою перевагою GEE є перенесення обчислень на сервери Google. Таким чином, розрахунок вегетаційних індексів для величезних територій є швидшим, без гострої необхідності завантажувати чисельні масиви на локальний комп’ютер.

Щодо недоліків, по-перше, архітектура GEE базується на суворому «vendor lock-in» та асинхронному виконанні коду, а це робить значно складнішою інтеграцію зі сторонніми Python-бібліотеками машинного навчання [20]. Щоб навчити локальну модель RandomForestRegressor на даних з GEE, масиви все одно у підсумку доведеться експортувати у Google Drive або й зовсім завантажувати локально, повністю нівелюючи перевагу хмарних обчислень. По-друге, у GEE є суворі квоти на обсяг оперативної пам’яті та на час виконання скриптів для безкоштовних акаунтів. По-третє, інтеграція зовнішніх макротрендів безпосередньо в екосистему є алгоритмічно складною і негнучкою.

Веб-застосунок Giovanni, розроблений Goddard Earth Sciences Data and Information Services Center (GES DISC), дозволяє візуалізувати та аналізувати дані спостережень Землі без завантаження програмного забезпечення [21]. Він має зручний графічний інтерфейс користувача, ідеально підходить для

швидкої візуалізації часових рядів, побудови теплових карт та кореляційних матриць для вибраних регіонів у один клік. Мінусом є той факт, що NASA Giovanni це закрита система, яка не підтримує розробку власних програмних конвеєрів. Інструмент повністю прив'язаний до ручної взаємодії через інтерфейс браузера, що, вочевидь, повністю унеможливлює автоматизацію, наприклад, налаштування регулярного виконання аналізу за вказаним розкладом. Окрім цього, платформа обмежує користувача лише тими алгоритмами і формулами, які були заздалегідь запрограмовані розробниками NASA. Впровадження власних кастомних метрик оцінки, алгоритмів фільтрації шумів або специфічних моделей ансамблевого навчання в екосистему Giovanni є абсолютно неможливим.

Третя платформа, Sentinel Hub, це хмарний API для супутникових зображень. Він орієнтується на швидку доставку даних, переважно для місій Sentinel та Landsat, через OGC-сервіси [22]. Перевагою є обробка знімків на льоту через користувацькі скрипти, які пишуться мовою програмування JavaScript.

Sentinel Hub це комерційний продуктом, а отже жорстко обмежує обсяг запитів, площу обробки та доступ до історичних архівів даних на понад кілька років у межах безкоштовних тарифів. Додатково, як і у випадку з GEE, Evalscripts призначені для попіксельної обробки та рендерингу зображень, а не для тренування статистичних моделей чи регресійного аналізу великих часових рядів. Для використання scikit-learn або інтеграції даних з іншими науковими архівами розробнику, який пише код мовою Python, доводиться використовувати Sentinel Hub лише як, грубу кажучи, трубу для завантаження даних, значно переплачуючи за комерційний доступ.

Серед інших вагомих недоліків класичних інформаційні системи супутникового моніторингу треба зазначили візуалізацію лише сирих знімків або статичну статистику, де складний процес аналітичної обробки багатовимірних масивів падає на плечі кінцевого користувача. На якість оптичних супутникових знімків впливають також хмарність та атмосферні

аерозолі. Вирішення цієї проблеми вимагає додаткового впровадження складних алгоритмів композитування для отримання безхмарних пікселів.

Отже, розроблювана система не повинна спиратися виключно на один тип даних. Вона має мати архітектуру із автоматичним завантаженням просторових даних для забезпечення регіонального покриття та використовувати їх як основу для інтелектуального аналізу динаміки екосистем.

Ще одним критичним недоліком є проблема масштабування, а саме – ізольованість аналізу локальних подій від глобальних процесів. Кліматична система Землі є цілісною, відповідно локальні зміни нерідко стають прямим наслідком глобальних макротрендів на кшталт загальнопланетарного зростання концентрації парникових газів.

Наразі цільові інформаційні системи переважно поділяються на два відокремлені класи: глобальні системи (макрорівень) та локальні (мікрорівень) (табл 1.3).

Таблиця 1.3 – Порівняння глобальних та локальних систем

Глобальні системи	Локальні системи
Платформи міжнародних організацій, наприклад, дашборди NOAA, WMO або IPCC, які відображають усереднені світові показники. Вони чудово ілюструють глобальне підвищення температури або експоненційне зростання CO2 [2], проте не дають відповіді на питання, як саме ці зміни впливають на конкретне фермерське поле чи лісове господарство у заданому географічному полігоні.	Регіональні ГІС-портали або платформи точного землеробства, які детально аналізують динаміку рослинності на обмеженій території, але розглядають ці зміни у вакуумі, у відриві від глобальних драйверів зміни клімату.

Основна інженерна проблема інтеграції полягає у просторово-часовій невідповідності різнорідних масивів даних. Зазвичай глобальні показники оновлюються як річні або щомісячні усереднені значення, тоді як супутникові

індекси вегетації генеруються кожні 8-16 днів для тисяч окремих пікселів матриці. Більшість існуючих рішень не мають вбудованого автоматизованого конвеєра, який би міг:

1) динамічно згортати просторові дані у регіональні часові ряди;
2) автоматично звертатися до зовнішніх баз даних через API для отримання актуальних макротрендів;
3) синхронізувати ці часові ряди за часовою віссю, наприклад, шляхом групування композитів супутника у середньорічні значення для співставлення з річною концентрацією CO2.

Як наслідок, дослідникам доводиться проводити усю цю інтеграцію вручну: експортувати супутникові дані в таблиці, вручну шукати та завантажувати файли з даними глобальних викидів і в подальшому об’єднувати їх у сторонніх програмах для статистичного аналізу. Це категорично суперечить створенню системи з роботою в реальному часу та суттєво затримує загальний процес прийняття екологічних рішень. Дослідження NASA вказують на те, що динаміка глобальних процесів, зокрема здатність екосистем поглинати вуглець, постійно змінюється, тому статичний аналіз швидко втрачає свою релевантність [5].

Вирішення проблеми може лежати у площині застосування методів Data Science та машинного навчання разом із хмарними API доступу до даних. Значить, інтелектуальна система нового покоління повинна бути побудована як безперервний програмний конвеєр, здатний самостійно стягувати глобальні макротренди, агрегувати локальні просторові дані та зшивати їх для навчання прогностичних моделей. Такий програмний продукт не тільки констатуватиме стан екосистеми, він також виявлятиме неочевидні кореляції, моделюватиме, як саме глобальні кліматичні фактори тиснуть на конкретний обраний регіон.

1.4. Аналіз гідрокліматичних зон України та факторів деградації рослинного покриву

Процеси глобального потепління на території України так само, як і усюди, характеризуються нерівномірністю та вираженою регіональною специфікою. За рахунок цього вже зараз помітна трансформація традиційних природних зон. Станом на сьогодні аналіз екологічної стабільності територій вимагає переходу від загальних температурних показників до комплексних характеристик гідротермічного режиму. Основним детермінантом стану біомаси у помірних широтах виступає баланс між надходженням вологи та інтенсивністю її випаровування, а це в свою чергу безпосередньо впливає на метаболічну активність рослинних покривів та на їхню спектральну відбивну здатність.

Сучасні спостереження підтверджують стрімке зміщення кліматичних зон України у північному напрямку. Території, які раніше характеризувалися стійким зволоженням, дедалі частіше демонструють дефіцит вологи, властивий зонам лісостепу та північного степу [7, 23]. Це явище супроводжується також зростанням частоти екстремальних літніх температур та зміною структури опадів, котрі набувають зливового характеру, але це не сприяє ефективному накопиченню вологи у ґрунті. Внаслідок цього спостерігається зниження асиміляційного потенціалу рослинності. У цифровій площині це фіксується супутниковими сенсорами як стійка деградація значень вегетаційних індексів [25].

Варто також зауважити, що накопичення діоксиду вуглецю, яке розглядається у межах даної роботи як ключовий глобальний макротренд, чинить подвійний вплив на стан гідрокліматичних зон. З одного боку, зростання концентрації парникових газів інтенсифікує потепління, призводить до підвищення швидкості евапотранспірації, тобто до випаровування вологи з поверхні листя та ґрунту [24]. Як результат, навіть за відносно стабільної кількості опадів, реальна гідрологічна забезпеченість рослин знижується через стрімке висихання кореневмісного шару ґрунту. З іншого ж боку, порушення

теплового балансу атмосфери дестабілізує циркуляційні процеси, викликаючи тривалі бездощові періоди.

1.5. Обґрунтування технологічного стеку для розробки системи

З попереднього підрозділу зрозуміло, що розробка вимагає переходу від традиційних ручних методів роботи з ГІС до повністю автоматизованих програмних конвеєрів. Тепер стоїть задача вибору технологічного стеку для визначення обчислювальної ефективності майбутньої системи та її здатності до масштабування, інтеграції з хмарними сервісами та впровадження алгоритмів штучного інтелекту.

Відповідно до методології CRISP-DM, етап підготовки даних та моделювання потребує гнучкого інструментарію із високою швидкістю розробки та підтримкою наукових форматів даних [26]. Враховуючи ці вимоги, базовою мовою програмування для розробки системи було обрано Python [27]. Станом на сьогодні Python є беззаперечним галузевим стандартом у сферах Data Science, Machine Learning та обробки даних ДЗЗ завдяки вкрай широкій екосистемі спеціалізованих відкритих бібліотек.

У сфері Data Science, як відомо, існують дві домінуючі мови програмування Python та R. Кожна з них, безперечно, має свої сильні сторони, проте для реалізації комплексного програмного конвеєра було обрано саме Python. Обґрунтування цього вибору базується на глибокому порівняльному аналізі їхніх архітектурних можливостей, викладеному далі.

Мова R створювалася статистиками для статистиків. У неї дійсно неперевершений інструментарій для розвідувального аналізу даних, перевірки статистичних гіпотез та побудови графіків, проте R вкрай вузькоспеціалізована мова. Python натомість це мова програмування загального призначення. Це означає, що він дозволяє не тільки аналізувати дані, але й із легкістю розробляти веб-сервери, налаштовувати взаємодію з мережевими API, інтегрувати базу даних в єдиному скрипті тощо.

R хоч і має бібліотеки для побудови моделей, екосистема Python сьогодні це недосяжний лідер у сфері штучного інтелекту. Оптимізована бібліотека scikit-learn дає змогу швидко і легко розгортати цілі ансамблеві моделі, проводити крос-валідацію [28].

Додатково Python володіє унікальними інструментами для роботи з багатовимірними науковими форматами. R, на противагу, стикається із труднощами при подібній обробці масивів, котрі банально не вміщаються у оперативну пам’ять. Із Python такої проблеми не спостерігається, він нативно інтегрується з бібліотекою паралельних обчислень dask [29]. Так само моделі, написані на R, часто важко перетворити на автономні веб-застосунки для кінцевих користувачів.

Отже, за проведеними порівняннями, Python є найбільш оптимальним технологічним вибором, оскільки забезпечує безшовний перехід від завантаження сирих супутникових гранул до навчання ML-моделей та їх візуалізації.

Як було визначено раніше, однією з головних проблем існуючих систем є необхідність ручного завантаження даних. Для вирішення цього технологічний стек включає бібліотеку earthaccess [30]. Це сучасний високорівневий інтерфейс API для програмованої взаємодії з архівами NASA Earthdata. Його застосування дозволяє повністю відмовитись від мануальної навігації веб-порталами. Бібліотека інкапсулює складні процеси авторизації, побудови просторово-часових запитів та паралельного завантаження гранул по протоколу HTTPS або через хмарний доступ S3. Ця технологія дійсно дозволяє системі бути динамічною, адже при кожному запуску програма самостійно стягуватиме найсвіжіші композити місії VIIRS для будь-якого заданого активним користувачем регіону.

Супутникові знімки у форматах HDF5/NetCDF це не звичайні табличні дані або плоскі зображеннями. Вони є багатовимірними тензорами, всередині яких поєднуються координати, час та спектральні канали. Стандартні засоби обробки даних не можуть ефективно оперувати такими структурами. Власне,

тому ядром обробки просторових даних обрано екосистему, побудовану навколо бібліотеки xarray [15].

Варто враховувати, що робота з даними місій MODIS та VIIRS передбачає маніпулювання складними масивами інформації. Супутниковий знімок, як вже було описано раніше, це на звичайна двовимірна таблиця, багатовимірний куб даних, який містить як мінімум три виміри: просторові координати (довгота x, широта y), часову вісь (хронологія знімків t) та спектральні канали або змінні.

Класичні інструменти NumPy або Pandas, очевидно, не пристосовані для ефективної роботи з DataCubes [31, 32]:

1) NumPy оперує багатовимірними тензорами, проте не зберігає метадані. Доступ до даних здійснюється через цілочисельні індекси, тобто код стає нечитабельним та схильним до помилок.
2) Pandas ідеально працює з маркованими даними. Його архітектура жорстко обмежена двома вимірами. Перетворення DataCube у плоску таблицю Pandas може призвести до експоненційного зростання споживання оперативної пам’яті та втрати просторової топології.

Для вирішення цієї проблеми у технологічний стек інтегровано бібліотеку xarray. Її архітектура поєднує математичну потужність NumPy із маркуванням Pandas, таким способом розповсюджуючи ці можливості на N-вимірні тензори. Через xarray можна звертатися до даних не за числовими індексами, а за семантичними мітками. Наприклад, щоб розрахувати середнє значення вегетації для певного регіону за весь час, розробнику достатньо викликати відповідний метод. Xarray самостійно зрозуміє, за якими саме вісями потрібно провести математичне усереднення, але зберігши при цьому вісь часу [15].

При обробці десятків гранул HDF5 вагою у гігабайти, xarray не завантажує їх у пам’ять моментально. Замість цього бібліотека створює віртуальний вказівник на дані. Безпосередні математичні обчислення, такі як завантаження в RAM та розрахунок, відбуваються виключно у момент

виклику методу *.compute*(). Саме тому вдається обробляти масиви даних, які суттєво перевищують обсяг фізичної оперативної пам’яті комп’ютера. Масиви також автоматично синхронізуються за їхніми координатами перед виконанням математичних операцій, що виключає ризик помилок при зсуві пікселів.

Оскільки продукти NASA, як було вказано раніше, мають складну ієрархічну внутрішню структуру, технологічний стек потрібно доповнити низькорівневою бібліотекою h5py [33]. Це інтерфейс мови Python до бінарного формату HDF5, через який система виконує динамічний парсинг файлів, автоматично знаходить приховані групи з науковими даними та у підсумку передає їх для подальшого читання у xarray. Після такого просторового усереднення багатовимірних тензорів з’являється одна з найвідоміших Python-бібліотек pandas. Вона використовується для представлення результатів у вигляді плоских часових рядів, їхнього очищення від пропусків, зміни частоти дискретизації та виконання реляційного злиття супутникових показників із даними з серверів NOAA.

Згідно із планом розробки системи, заключним етапом є створення користувацького інтерфейсу для екологічного моніторингу. У екосистемі Python сьогодні існують три основні інструменти для розгортання веб-застосунків: Django, Flask та Streamlit (табл. 1.4). Вибір на користь Streamlit був зроблений на основі порівняльного аналізу.

Таблиця 1.4 – Порівняння фреймворків для розгортання веб-застосунків

Інструмент	Опис
Django	Монолітний фреймворк, побудований за патерном MTV (Model-Template-View). Ідеально підходить для створення складних корпоративних систем з десятками баз даних, системою управління користувачами та маршрутизацією [35]. Для аналітичного дашборду, щоправда, Django є надлишковим. Його розгортання вимагає значних витрат часу на налаштування

	конфігурацій, міграцій баз даних, написання складного HTML, CSS та JavaScript коду для відображення графіків, а це лише відволікає від основної цілі.
Flask	Максимально гнучкий мікрофреймворк, який дозволяє швидко створити API-сервер. Проте, як і Django, Flask вимагає мануального написання клієнтської частини [36]. Розробнику усе одно доводиться використовувати шаблонізатори, вручну передавати масиви даних з Python у JavaScript для побудови графіків. Це, очевидно, уповільнює цикл розробки та вимагає додатково глибоких знань веб-технологій.
Streamlit	Спеціалізований фреймворк, розроблений виключно для інженерів з машинного навчання та Data Science. Його архітектура працює таким чином, що дає можливість створювати повноцінні, інтерактивні веб-застосунки, пишучи при цьому виключно Python-код [34].

Іншими словами, замість написання сотень рядків фронтенд-коду, у Streamlit графік або інтерактивна карта на базі Folium виводяться лише одним рядком [37]. Streamlit автоматично керує станом системи. Щоразу, коли користувач змінює якийсь параметр, фреймворк самостійно ініціює перерахунок відповідного графа обчислень, чим позбавляє необхідності розробляти складні механізми AJAX-запитів.

Streamlit, власне, розроблявся з урахуванням прямої підтримки об’єктів Pandas DataFrame, моделей Scikit-learn та графіків Matplotlib/Plotly, саме тому він є ідеальним фінальним вузлом у конвеєрі даних. Обґрунтований вибір зв’язки Python, Xarray, Scikit-learn та Streamlit оптимізує процес розробки, змістивши фокус з рутинного веб-програмування на вирішення складних екологічних та математичних завдань предметної області.

Висновки

Проведений у першому розділі аналіз предметної області та сучасних методів моніторингу дозволив сформувати комплексне науково-технічне підґрунтя для розробки інтелектуальної системи динамічного аналізу регіону.

Встановлено масштаб кліматичної проблеми, вказано на те, що глобальні кліматичні зміни, спричинені зростанням концентрації парникових газів, призводять до дестабілізації локальних екосистем. Доведено, що стан рослинного покриву є еталонним інтегральним індикатором кліматичного стресу. Безперервне регіональне спостереження має важливе значення не лише для аграрної та лісової галузей, а й для досягнення цілей сталого розвитку ООН.

Проаналізовано актуальні методи дистанційного зондування, з'ясовано фізичну природу спектральних сигнатур різних поверхонь, що обґрунтовує використання оптичних каналів для аналізу стану біосфери. Детальний порівняльний аналіз супутникових місій довів перевагу використання сучасних продуктів VIIRS над ранішими аналогами MODIS для забезпечення довгострокового моніторингу високої роздільної здатності.

Виявлено критичні недоліки існуючих систем через дослідження класичних метеостанцій та популярних хмарних ГІС-платформ Google Earth Engine, NASA Giovanni та Sentinel Hub. Головними проблемами є закритість архітектур, складність налаштовуваності алгоритмів та практично неможливість гнучкої інтеграції високочастотних локальних супутникових даних із глобальними кліматичними макротрендами.

Для подолання усіх виявлених недоліків проведено порівняльний аналіз технологій та сформовано технологічний стек. Визначено, що екосистема Python значно перевершує аналоги у завданнях побудови комплексних Data Pipelines. Обґрунтовано використання бібліотеки Xarray для ефективної роботи з багатовимірними DataCubes за допомогою лінивих обчислень. Для реалізації предиктивної аналітики обрано scikit-learn, а розробку реактивного користувацького інтерфейсу довірено фреймворку Streamlit, котрий позбавляє потреби розгортати важкі монолітні системи на базі Django чи Flask.

Отримані результати формують підґрунтя для подальшого дослідження: вони забезпечують перехід від теоретичного аналізу фізичних кліматичних

процесів до їхньої безпосередньої математичної формалізації через спектральні індекси та алгоритми машинного навчання.

РОЗДІЛ 2. ФОРМАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ ДАНИХ У ПРОЄКТІ

2.1. Математична формалізація вегетаційних індексів

Математична формалізація стану біосфери це вкрай важлива стадія у побудові інтелектуальних систем регіонального моніторингу. Оскільки пряме вимірювання біомаси або рівня фотосинтезу на великих територіях є технічно неможливим, у ДЗЗ зазвичай застосовують спектральні вегетаційні індекси. Ці безрозмірні радіометричні величини обчислюються шляхом математичних операцій над значеннями відбивальної здатності у різних каналах електромагнітного спектра. Основна мета – максимізація їхньої чутливості до біофізичних характеристик рослинності та паралельній мінімізації впливу фонового шуму, такого як колір і вологість ґрунту, кут падіння сонячних променів або атмосферне розсіювання [6, 9].

Еталонним та найбільш математично обґрунтованим показником є нормалізований відносний індекс рослинності NDVI (Normalized Difference Vegetation Index). Його фізична суть базується на контрасті між двома спектральними діапазонами, які по-різному взаємодіють із листковим апаратом рослин [6, 12]:

1) між видимим червоним діапазоном (Red, $\lambda \approx$ 600-700 нм), де енергія інтенсивно поглинається хлорофілом для забезпечення процесу фотосинтезу.
2) між ближнім інфрачервоним діапазоном (NIR, $\lambda \approx$ 700-1000 нм), де енергія сильно відбивається та розсіюється клітинною структурою губчастого мезофілу листка, щоб уникнути перегрівання тканин.

Математично розрахунок індексу NDVI для кожного окремого пікселя супутникового знімка формалізується рівнянням (2).

$$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}, \quad (2)$$

де ρ_{NIR} та ρ_{Red} – значення спектральної відбивальної здатності у ближньому інфрачервоному та червоному каналах відповідно [6]. Завдяки

нормалізації, тобто діленню різниці на суму, діапазон значень індексу завжди обмежується відрізком $[-1; 1]$. Від'ємні значення зазвичай відповідають водним об'єктам, снігу або хмарам. Значення, розташовані близько до нуля (від 0,0 до 0,15), характерні переважно для голих ґрунтів, скель та урбанізованих територій. Для фітоценозів значення індексу коливаються від 0,2, що відповідає розрідженій чи пригніченій вегетації, до 0,8-0,9, це густі, здорові лісові масиви або сільськогосподарські культури на піку вегетації.

Ідеальним індикатором для інтелектуального аналізу NDVI робить його чутливість до кліматичних факторів. Динаміка індексу в часі тісно корелює з гідротермічним режимом регіону. Наприклад, підвищення глобальної температури та зниження кількості опадів призводять до закриття продихів у листках рослин для мінімізації втрат вологи. Так миттєво знижується інтенсивність фотосинтезу: поглинання в червоному спектрі зменшується, а клітинна структура деградує, у свою чергу знижуючи відбиття в інфрачервоному. Як наслідок, значення NDVI різко падає. Довгострокові часові ряди цього індексу сприяють фіксуванню глобальних кліматичних зсувів, таких от як зміщення фенологічних фаз (раніший початок весняної вегетації та пізніше осіннє в'янення), або ж зниження здатності лісових екосистем поглинати CO2.

Попри універсальність та розповсюдження NDVI, індекс має все ж помітний математичний недолік – ефект насичення при високій щільності рослинного покриву. Феномен проявляється у лісових екосистемах або на стадії максимального розвитку агроценозів, коли індекс листкової поверхні LAI (Leaf Area Index) перевищує значення 3-4 [38]. Природу цього явища пояснюють асимптотичною поведінкою формули NDVI. Коли рослинність стає надто густою, то концентрація хлорофілу поглинає практично 100% червоного світла, іншими словами, ρ_{Red} наближається до нуля. З математичної точки зору, межа функції NDVI при прямуванні відбиття у червоному спектрі до нуля (3).

$$\lim_{\rho_{Red}\to 0} \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}} = \frac{\rho_{NIR}}{\rho_{NIR}} = 1 \quad (3)$$

У цьому стані, навіть якщо біомаса рослинності продовжує стрімко зростати, викликаючи збільшення відбиття у ближньому інфрачервоному каналі (ρ_{NIR} зростає з 0,4 до 0,8 за рахунок багатократного розсіювання всередині багатокронового лісу), значення NDVI майже не змінюється і залишається близьким до 0,95-0,99. Тобто, похідна функції NDVI відносно біомаси наближується до нуля. Через це індекс може втрачати свою чутливість до подальших змін у фітоценозі, ускладнюючи подальше відстеження кліматичних змін у густих лісах, принаймні при оцінці поглинання ними вуглецю.

На розріджених територіях із низьким LAI введено ґрунтово-коригований вегетаційний індекс SAVI (Soil Adjusted Vegetation Index) якраз для розв'язання проблеми чутливості NDVI до фонового шуму ґрунту. На локаціях із нещільною рослинністю спектральний сигнал відбивається не тільки від листя, а й від відкритого ґрунту. Вологість землі, очевидно, змінює її колір, у підсумку темний вологий ґрунт поглинає більше світла, ніж світлий сухий. Так штучно змінюється співвідношення ρ_{NIR} та ρ_{Red}, спотворюючи кінцеве значення NDVI. SAVI математично нівелює цю залежність, вводячи коригуючий коефіцієнт L (4).

$$SAVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red} + L} \cdot (1 + L) \quad (4)$$

Параметр L це емпіричний коефіцієнтом коригування фону ґрунту [39]. Його значення може варіюватися від 0 до 1. Якщо рослинність дуже густа, тоді вплив ґрунту відсутній, приймається $L = 0$, і формула SAVI повністю редукується до класичної формули NDVI. Для територій із відкритою ґрунтовою поверхнею, таких як пустелі, прийнято застосовувати $L = 1$. У більшості екологічних систем за замовчуванням використовується компромісне значення $L = 0{,}5$, чим ефективно мінімізується дисперсія значень індексу, спричинена зміною вологості підстилаючої поверхні.

Для комплексного вирішення як вищезгаданої проблеми насичення при високій біомасі, так і проблеми ґрунтового шуму та атмосферного розсіювання, було розроблено покращений вегетаційний індекс EVI (Enhanced Vegetation Index) науковими групами NASA [38]. Показник став вже стандартним продуктом для супутникових місій MODIS та VIIRS. EVI оптимізує сигнал вегетації за рахунок інтеграції додаткового синього спектрального каналу, який використовується для корекції аерозольного розсіювання в червоному каналі. Математична модель EVI описана у формулі (5).

$$EVI = G \times \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + C_1 \cdot \rho_{Red} + C_2 \cdot \rho_{Blue} + L} \quad (5)$$

У цьому рівнянні G є фактором масштабування, а C_1 та C_2 це коефіцієнти аерозольного опору. Вони використовують синє відбиття ρ_{Blue} для корекції впливу аерозолів на червоне відбиття, тоді як L зберігає свою функцію коригування на ґрунтовий фон. Відповідно до алгоритмів обробки даних NASA для продуктів MOD13Q1 та VNP13A1, константи приймають стандартизовані значення: $G = 2{,}5$, $C_1 = 6{,}0$, $C_2 = 7{,}5$, $L = 1{,}0$. На відміну від NDVI, EVI зовсім не насичується у щільних тропічних та помірних лісах, а тому системі вдається значно точніше моделювати реакцію розвинених екосистем на глобальні макрокліматичні тенденції.

Сирі супутникові дані, як правило, обтяжені різноманітними шумами внаслідок наявності хмарного покриву, тіней від хмар, атмосферних аерозолів. Також не варто ігнорувати банальні інструментальні похибки сенсорів. Щоб модель машинного навчання могла виявляти справжні кліматичні тренди, а не реагувати на атмосферні перешкоди, необхідна сувора математична фільтрація та агрегація даних. Першим рівнем є використання попередньо агрегованих композитних продуктів. Замість аналізу кожного щоденного знімку, який ще й з високою ймовірністю буде закритий хмарами, алгоритми на стороні провайдера формують єдину безхмарну матрицю за 16-денний проміжок часу. Для кожного пікселя обирається те значення відбиття, котре

відповідає найвищому показнику вегетації за умови чистого неба, так званий алгоритм Maximum Value Composite [40]. Проведення цієї процедури може допомогти із мінімізацією впливу атмосферних шумів.

Протягом другого етапом треба виділити програмне керування некоректними даними на етапі ініціалізації масиву у розробленій системі на базі мови Python. У форматах HDF5 та NetCDF області з пропущеними даними, або ті, котрі потрапили у «сліпу зону» супутника, кодуються спеціальними маркерними величинами. Якщо такі числа потраплять до математичних формул, вони повністю знищать релевантність результатів. Під час автоматизованого завантаження тензорів за допомогою бібліотеки xarray, ці маркерні величини автоматично розпізнаються та конвертуються у спеціальний тип даних NaN [15]. При операціях із цим типом обчислювальні алгоритми ігнорують биті пікселі, намагаючись знайти середні величини.

Окрім цього, протягом динамічного розрахунку вегетаційних індексів безпосередньо з каналів відбиття, виникає ризик зупинки програми через ділення на нуль, коли сума відбиття $\rho_{NIR} + \rho_{Red}$ випадково дорівнює нулю. У програмній реалізації конвеєра застосовується алгоритм уникнення сингулярності за допомогою інструменту *xr.where*, де знаменник математично описується функцією (6).

$$Denominator = \begin{cases} \rho_{NIR} + \rho_{Red}, \text{ якщо} (\rho_{NIR} + \rho_{Red}) \neq 0 \\ 10^{-10}, \text{ якщо} (\rho_{NIR} + \rho_{Red}) = 0 \end{cases} \quad (6)$$

Третім етапом є просторова агрегація. Для переходу від мільйонів окремих пікселів до єдиного показника регіону необхідно застосувати операцію просторового усереднення. Математично це знаходження середнього арифметичного значення NDVI для усієї матриці в кожен момент часу t, ігноруючи при цьому порожні значення (7).

$$\overline{NDVI}(t) = \frac{1}{N_{valid}} = \sum_{i=1}^{X} \sum_{j=1}^{Y} NDVI_{i,j,t} \cdot M_{i,j,t}, \quad (7)$$

де X та Y – довгота та широта у масиві, $M_{i,j,t}$ – бінарна маска, яка приймає значення 1, якщо піксель містить коректне значення, і 0, якщо значення є NaN, а N_{valid} – загальна кількість дійсних пікселів у регіоні.

Часова агрегація та ресемплінг отриманих таймсерій виступають фінальними кроками. Оскільки на меті системи стоїть поєднання локальних індексів рослинності із макротрендами, високочастотні дані NDVI необхідно якимось чином привести до відповідної частоти дискретизації. Інструментарій бібліотеки pandas тут застосовується для згладження 16-денних часових рядів, які далі групуються та згортаються у місячні чи середньорічні значення [27, 32]. Так повністю ліквідується сезонний високочастотний шум, а в даних залишається суто низькочастотний кліматичний сигнал. Саме цей підготовлений, нормалізований та очищений від просторових і часових аномалій масив даних стає фундаментом для подальшого застосування методів інтелектуального аналізу та регресійного моделювання за допомогою алгоритмів ансамблевого навчання.

2.2. Обґрунтування вибору методів інтелектуального аналізу супутникових даних та макрокліматичних показників

Кліматична система Землі та її локальні прояви є, вочевидь, складними, нелінійними динамічними системами. Зміни в біосфері, які фіксуються супутниковими знімками, формуються насамперед під впливом десятків взаємопов’язаних факторів: від локальних гідротермічних умов до глобальних трендів накопичення парникових газів. У зв’язку із цим, застосування класичних методів лінійної статистики часто виявляється просто неефективним, оскільки вони не здатні вловити складні взаємодії між предикторами та моделювати порогові ефекти, характерні для екологічних систем [41].

Етап моделювання, згідно з CRISP-DM, потребує вибору алгоритмів, які найкраще відповідають природі наявних даних та цілям дослідження [26]. Якщо враховувати те, що після агрегації супутникових знімків ми отримуємо

табличні дані у вигляді часових рядів, найбільш обґрунтованим є використання методів ансамблевого машинного навчання на основі дерев рішень. Для розроблюваної системи базовим алгоритмом інтелектуального аналізу обрано RandomForestRegressor, який реалізований у бібліотеці scikit-learn [42, 28].

Вибір цього алгоритму для подібної мети екологічного моніторингу пояснюється високою стійкістю до викидів, а це важливий пункт через потенційну наявність аномальних супутникових пікселів, які не були відфільтровані, та здатністю працювати з нелінійними залежностями. Іншою корисною особливістю варто визначити відсутність надто жорстких вимог до нормалізації вхідних даних. Уцілому, алгоритм Random Forest – це дуже потужний ансамблевий метод машинного навчання, який, власне, комбінує прогнози великої кількості незалежних базових моделей для отримання одного, значно точнішого та стабільнішого результату.

У фундаменті випадкового лісу лежить алгоритм побудови дерева рішень, а точніше алгоритм CART (Classification and Regression Trees) [43]. Дерево рішень це по суті ієрархічна граф-структура, яка рекурсивно розбиває простір ознак на прямокутні області. Структурно дерево складається з трьох типів елементів: кореневий вузол, внутрішні вузли та листки.

1. Кореневий вузол це початкова точка графа. Вона, очевидно, містить усю навчальну вибірку даних. Тут відбувається перше, найважливіше розбиття даних на основі найбільш вагомогої ознаки.
2. Далі йдуть вузли, де відбувається перевірка певної заданої логічної умови. Тут може бути, наприклад, твердження про те, чи перевищує концентрація CO2 рівень встановлений рівень. Залежно від виконання умови, дані направляються або у ліву, або у праву гілку.
3. Листки описують кінцеві вузли, без яких-небудь подальших розгалужень. У задачах регресії кожен такий листок містить фінальне спрогнозоване значення цільової змінної, як от конкретне значення індексу NDVI.

Для ухвалення деревом рішення, за якою ознакою і в якій точці здійснити розбиття, алгоритм повинен використовувати строгі математичні критерії оцінки чистоти вузла. У розробленій системі прогнозування NDVI, хоч і вирішується задача регресії, для більш поглибленого розуміння механіки CART необхідно розглянути класичні метрики теорії інформації, які алгоритм використовує у задачах класифікації.

Першою такою метрикою є ентропія. Це міра хаосу, невизначеності або непередбачуваності інформації у вузлі. Відповідно, чим вищою є ця ентропія, тим сильніше перемішані між собою дані різних класів. Математично ентропія множини S обчислюється за формулою Шеннона (8).

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i, \quad (8)$$

де c – кількість класів, а p_i – частка, тобто ймовірність, елементів i-го класу у множині S. Алгоритм за своєю суттю прагне мінімізувати ентропію, створюючи при цьому максимально чисті листки [44, 45].

Існує також інформаційний приріст, який показує, наскільки ефективно вибрана ознака A знижує ту саму ентропію системи після розбиття. Алгоритм для цієї метрики розраховує невизначеність до розбиття та віднімає від неї зважену суму ентропій усіх дочірніх вузлів після розбивання за формулою (9).

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (9)$$

Тут S_v – підмножина S, для якої ознака A має значення v. Дерево завжди обирає саме те розбиття, яке у підсумку максимізуватиме інформаційний приріст $IG(S, A)$.

Альтернативний до ентропії критерій, який оптимізований для швидших обчислень через не використання логарифмів, називається індексом Джині. Він оцінює ймовірність того, що випадково обраний елемент вузла буде класифікований неправильно, якщо його мітка буде обрана випадково згідно з розподілом міток у вузлі (10) [45].

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \quad (10)$$

У конкретно цьому випадку, замість ентропії чи індексу Джині, алгоритм мінімізуватиме дисперсію або середньоквадратичну помилку. Якщо вузол t містить N_t спостережень, а після розбиття за певною ознакою утворюються лівий t_L та правий t_R дочірні вузли, тоді алгоритм починає шукати таке розбиття, яке максимізувало б зменшення дисперсії.

$$\Delta I = N_t \cdot Var(t) - \left(N_{t_L} \cdot Var(t_L) + N_{t_R} \cdot Var(t_R)\right) \quad (11)$$

Одне дерево рішень є жадібним алгоритмом, схильним до перенавчання. Воно може, звісно, ідеально запам’ятати увесь тренувальний набір супутникових даних, проте потім показати надто низьку точність на нових невідомих даних. Щоб виправити цей недолік, у Random Forest зазвичай використовують дві ключові концепції: беггінг та метод випадкових підпросторів.

Під час беггінгу алгоритм спочатку створює B нових тренувальних вибірок однакового розміру з оригінального датасету шляхом вибірки з поверненням [42]. Слідом на кожній такій вибірці будується своє окреме дерево рішень. Через те, що ця вибірка формується випадково, кожне новостворене дерево бачить дещо іншу картину кліматичних даних. Загалом можна казати про приблизно 63,2% унікальних спостережень з оригінального датасету, а усе решта – дублікати.

Щодо випадкових підпросторів, то протягом побудови кожного дерева, у кожному його вузлі алгоритм не буде розглядати усі доступні ознаки для розбиття, а випадковим чином обиратиме лише їхню підмножину, зазвичай це $m = \frac{p}{3}$ для задач регресії, де p – загальна кількість ознак. Таким чином досягається декорреляція дерев [46]. Іншими словами, навіть якщо у даних є якась одна домінуюча кліматична ознака, то вона не стане використовуватися для першого розбиття в усіх деревах, тобто алгоритм цілком може виявити вплив і менш очевидних факторів.

Фінальний прогноз моделі $\hat{y}$ для нового спостереження x обчислюється як середнє арифметичне прогнозів усіх B дерев у лісі за формулою (12).

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x), \quad (12)$$

де $h_b(x)$ це прогноз регресійного значення NDVI від b-го дерева. Подібне усереднення спрямоване на те, аби значно знизити загальну дисперсію моделі, не збільшуючи при цьому зміщення. RandomForestRegressor за рахунок цього можна впевнено вважати одним із найефективніших інструментів у бібліотеці scikit-learn для аналізу екологічних часових рядів, тобто для поставленої задачі в рамках роботи.

Незалежно від того, чи демонструє Random Forest високу ефективність, для досягнення наукової точності у будь-якому випадку необхідно провести процес оптимізації і налаштування гіперпараметрів (табл. 2.1). Гіперпараметри – це по суті настройки архітектури моделі, які можуть встановлюватися ще до початку навчання.

Таблиця 2.1 – Ключові параметри Random Forest.

Гіперпараметр	Опис
n_estimators (кількість дерев у лісі).	За рахунок збільшення кількості дерев підвищується кінцева точність та відповідно стабільність моделі, але при цьому прямо пропорційно збільшується час обчислень. Це крок, який варто зробити. Базове значення параметра дорівнює 100, однак у процесі оптимізації його можна буде збільшити до 300 або 500 задля стабілізації прогнозів на макротрендах.
max_depth (максимальна глибина дерева)	Цей параметр сильно обмежує кількість рівнів або розгалужень дерева. Якщо його не обмежити, кожне дерево зростатиме доти, доки в листках не залишиться по одному зразку, а це вже ідеальне перенавчання. Оптимальна глибина має визначатись

	емпірично для уникнення запам’ятовування шумів.
min_samples_split та min_samples_leaf (мінімальна кількість зразків)	Обидва цих схожих параметрів визначають, скільки як мінімум спостережень повинно бути у певною вузлі, аби алгоритм міг розділити його, або ж скільки зразків має залишитися у фінальному листку. Це такі собі додаткові регулятори проти перенавчання.
max_features (кількість ознак для розбиття)	Кількість випадкових ознак, які розглядаються при кожному розгалуженні. Для задач регресії стандартною практикою є використання третини від загальної кількості предикторів.

Підбір цих гіперпараметрів у розробленій системі здійснюється за допомогою методів GridSearchCV або RandomizedSearchCV [28, 47]. Обрані інструменти автоматично перебирають різноманітні комбінації налаштувань, проводять часову крос-валідацію для кожної комбінації та в підсумку повертають саме таку архітектуру, яка буде мінімізувати середньоквадратичну помилку на валідаційній вибірці.

У наукових дослідженнях з кліматології, очевидно, недостатньо всього лише створити Black Box модель, яка точно прогнозує значення індексу вегетації. Важливо також подбати про етап інференсу, логічного висновку. Це означає надати моделі здатність пояснити, які конкретно фактори становлять найбільшу частку у деградації чи ж навпаки у розвитоку екосистеми. Тут варто згадати іншу перевагу алгоритму Random Forest, а саме його вбудовану здатність оцінювати важливість ознак. У реалізованій архітектурі системи застосовується метод оцінки на основі середнього зменшення забруднення MDI (Mean Decrease Impurity), який у контексті регресії базується на зменшенні дисперсії [48].

Логіка його розрахунку полягає у наступному: кожного разу, коли певна ознака f_j використовується для того, щоб розбити той чи інший вузол у будь-

якому з дерев лісу, вона спричиняє певне зменшення дисперсії цільової змінної ΔI. Алгоритм, відповідно, відстежує ці зменшення для усіх ознак.

Важливість f_j для окремого дерева T розраховується як сума абсолютно всіх пророблених зменшень дисперсії у вузлах t, де ця ознака використовувалася для розбиття, зважена на частку спостережень, котрі пройшли через цей вузол.

$$Imp(f_j, T) = \sum_{t \in T: v(t) = f_j} \frac{N_t}{N_{total}} \Delta I(t) \quad (13)$$

У формулі (13) $v(t)$ це ознака, обрана для розбиття у вузлі t, N_t – кількість спостережень у вузлі, N_{total} – загальна кількість спостережень у дереві.

Загальна важливість ознаки для всього ансамблю обчислюється шляхом усереднення цих значень по всіх B деревах лісу, після чого отримані показники нормуються таким чином, аби їхня підсумкова сума була рівна одиниці або 100% (14).

$$Feature\ Importance(f_j) = \frac{1}{\sum_{k=1}^{p} Imp(f_k)} \cdot \frac{1}{B} \sum_{b=1}^{B} Imp(f_j, T_b) \quad (14)$$

Ця метрика загалом дозволяє розробленій системі сформувати аналітичні звіти, при цьому ранжуючи вхідні фактори відносно сили їхнього впливу на цільовий показник. Наприклад, під час аналізу кореляції між супутниковими масивами HDF5 та глобальними показниками, Feature Importance дає математично підтвердити або ж спростувати гіпотезу стосовно того, що багаторічний тренд накопичення парникових газів є статистично значущим предиктором деградації рослинності у обраному географічному регіоні.

Впровадження алгоритмів сімейства Random Forest повною мірою відповідає усім сучасним стандартам Data Science. Конкретно у цьому випадку воно забезпечує надійний міст між сирими супутниковими пікселями та

високорівневою екологічною аналітикою, необхідною далі для прийняття чітких рішень у галузі сталого розвитку.

2.3. Методологічні засади та критерії енергоефективності машинного навчання у кліматичних дослідженнях

Інтенсивне впровадження методів штучного інтелекту та глибокого машинного навчання у сферу екологічного моніторингу та моделювання змін клімату супроводжується виникненням фундаментального етичного та технологічного парадоксу. З одної сторони, алгоритми Data Science дозволяють обробляти петабайти супутникових даних та виявляюти приховані закономірності деградації біосфери, недоступні для класичного статистичного аналізу. Але якщо подивитись з іншого боку, навчання надскладних нейромережевих архітектур вимагає вилечезних витрат електроенергії, значна частина якої і досі генерується за рахунок спалювання викопного палива. Це усе призводить до ситуації, коли інструменти, покликані боротися з глобальним потеплінням, самі таким чином стають вагомим джерелом емісії парникових газів. Усвідомлення цієї проблеми зумовило формування в сучасній інформатиці нової наукової парадигми Green AI, яка протиставляється ресурсомісткому Red AI [49].

Етичний імператив екологічних досліджень вимагає від розробників інтелектуальних систем ретельного математичного обґрунтування енергоефективності обраних алгоритмів. Для кількісної оцінки вуглецевого сліду використовується комплексний підхід, який повинен враховувати як апаратні характеристики обчислювальної системи, так і специфіку енергомережі [50]. Сумарне енергоспоживання E у кіловат-годинах, кВт/год, у процесі тренування моделі протягом часу t формалізується інтегральним рівнянням (15).

$$E = PUE \cdot \int_0^t \big(P_{CPU}(\tau) + P_{GPU}(\tau) + P_{RAM}(\tau)\big) d\tau, \quad (15)$$

де $P_{CPU}(\tau)$, $P_{GPU}(\tau)$ та $P_{RAM}(\tau)$ – миттєві значення споживаної потужності центрального процесора, графічного прискорювача та оперативної пам'яті відповідно у момент часу τ. Коефіцієнт PUE (Power Usage Effectiveness) відображає інфраструктурну енергоефективність дата-центру, а саме витрати на охолодження та безперебійне живлення сервісних систем. Отримане значення витраченої енергії конвертується в еквівалент викидів діоксиду вуглецю $CO2e$ через показник вуглецевої інтенсивності регіональної електромережі CI вимірюється у $gCO2/kWh$.

$$CO2e = E \cdot CI \quad (16)$$

Враховуючи наведений математичний апарат (16), вибір конкретного алгоритму машинного навчання для проєкту стає не стільки питанням точності, скільки питанням все ж екологічної відповідальності. У сучасних задачах аналізу просторово-часових рядів нерідко можна побачити застосовування глибоких нейронних мереж. Це можуть бути і рекурентні мережі довгої короткострокової пам'яті, і тривимірні згорткові. Навчання подібних структур на гіперкубах даних вимагало б використання графічних прискорювачів. Воно також базується на алгоритмі зворотного поширення помилки. Загальна кількість операцій з плаваючою комою для нейромережі наближено оцінюється за формулою (17).

$$FLOPs_{DNN} \approx 3N_{params} \cdot S_{size} \cdot E_{epochs}, \quad (17)$$

де N_{params} – кількість параметрів мережі, яка цілком може сягати десятків мільйонів, S_{size} – обсяг навчальної вибірки, а E_{epochs} – кількість епох навчання. З огляду на необхідність проведення десятків експериментів для підбору гіперпараметрів, навчання навіть однієї нейромережевої моделі для регіонального кліматичного прогнозу може згенерувати вуглецевий слід, еквівалентний викидам автомобіля за кілька місяців експлуатації.

На противагу цьому, у даній магістерській роботі свідомо обрано та обґрунтовано використання ансамблевого алгоритму випадкового лісу. Архітектура цього методу в принципі цілком відповідає критеріям Green AI.

Обчислювальна складність побудови ансамблю з B дерев рішень наведена у формулі (18).

$$Complexity_{RF} = O(B \cdot K \cdot S \log S), \quad (18)$$

де B – кількість дерев естіматорів, K – кількість ознак, а S – кількість спостережень у вибірці після просторової агрегації супутникових пікселів [42, 45].

Математична природа дерев рішень не вимагає ітеративного градієнтного спуску, зворотного поширення помилки чи використання енергоємних графічних прискорювачів. Навчання моделі з гіперпараметрами відбувається взагалі на стандартних ядрах центрального процесора і займає лічені секунди. Крім того, примусова структурна регуляризація, себто обмеження глибини дерева, котра застосовується для якраз таки усунення перенавчання, має прямий позитивний побічний ефект: вона пропорційно зменшує кількість логічних операцій розщеплення вузлів, знижуючи енергоспоживання конвеєра до мінімально можливих значень.

Отже, при застосуванні концепції зеленого ШІ в архітектурі розробленої системи регіонального моніторингу можна не тільки вирішити суто технічні завдання оптимізації обчислень, можна також продемонструвати високу соціальну та наукову відповідальність як розробника.

2.4. Методологія поєднання супутникових даних з глобальними макротрендами

Система клімату Землі, як вже зазначалось, є відкритою, високодинамічною та глобально взаємопов’язаною. Локальні екологічні індикатори відображають не тільки загальну реакцію біосфери на певні місцеві метеорологічні умови, але й слугує довгостроковим наслідком глобальних макротрендів, насамперед невпинного зростання концентрації парникових газів у атмосфері. Відповідно, для побудови повноцінної моделі для прогнозування на базі штучного інтелекту необхідно об’єднати локальні просторові дані із даними глобального кліматичного моніторингу у певний

єдиний аналітичний простір. Це може бути вирішено за рахунок розробки строгої методології конвеєризації даних ETL (Extract, Transform, Load) та подальшого застосування математичного апарату з ціллю синхронізувати часові рядів із різною частотою дискретизації [26].

Безперервність роботи, висока автоматизація, відтворюваності результатів, будуть досягнені внаслідок програмної архітектури системи, яка спиратиметься на вищезгадану концепцію ETL. Цей процес запевняє у тому, що неструктуровані масиви з різних джерел будуть коректно вилучені, очищені та завантажені у пам’ять для подальшого навчання моделей.

Етап вилучення включає автоматизований збір сирих даних з незалежних зовнішніх джерел. Збір супутникових даних здійснюється саме через API-інтерфейси архівів NASA Earthdata за допомогою бібліотеки earthaccess, як було описано раніше [30]. Система надсилає просторово-часовий запит та стягує багатовимірні бінарні гранули у форматі HDF5.

Паралельно із цим система ініціює прямий HTTP-запит до серверів глобальної мережі моніторингу парникових газів Global Monitoring Laboratory, яка підтримується Національним управлінням океанічних і атмосферних досліджень США [2]. Тобто алгоритм завантажує актуальні дані у вигляді неструктурованих текстових файлів. Завдячуючи відмові від мануального завантаження, при кожному запуску системи актуальність даних буде максимальною.

Найскладніший математичний та алгоритмічний етап це, безумовно, трансформація. Вона передбачає перетворення різнорідних масивів у стандартизований вигляд. Використовуючи інструментарій бібліотеки pandas для Python, система здійснює очищення даних. Текстовий файл NOAA парситься із застосуванням регулярних виразів для розділення колонок. Рядки із метаданими або коментарі алгоритмічно відфільтровуються. В подальшому відбувається екстракція цільових ознак: часової мітки та абсолютного значення концентрації діоксиду вуглецю в частинах на мільйон. Усі вилучені рядки приводяться до єдиного числового типу. Зі сторони супутникових даних

матриці проходять просторову агрегацію за допомогою xarray, згортаючись після цього у таблицю із двома колонками: точна дата зйомки та середнє вираховане значення NDVI. У кінці відбувається реляційне злиття двох незалежних масивів у єдиний структурований датасет, який зберігається в оперативній пам'яті комп'ютера і безпосередньо передається як тензор ознак для навчання алгоритму RandomForestRegressor.

Одною з основних математичних проблем під час цього самого етапу трансформації є просторово-часова невідповідність. Дані дистанційного зондування генерують високочастотний часовий ряд: нове значення NDVI з'являється кожні 16 днів. Водночас, глобальні макротренди NOAA часто надаються як низькочастотні масиви, де міститься лише одне усереднене значення концентрації CO2 на рік. Пряме об'єднання таких рядів, вочевидь, просто неможливе, оскільки їхні часові вектори не збігаються. Для вирішення подібних проблем у Data Science застосовують два протилежні підходи: агрегацію високочастотного ряду або інтерполяцію низькочастотного.

Перший шлях полягає у зниженні частоти супутникових даних до рівня глобальних макротрендів. Нехай у нас є часовий ряд локального індексу вегетації $V = \{(t_i, NDVI_i)\}$, де t_i – точна дата супутникового композиту. Тоді алгоритм згортає усі 16-денні значення в межах одного календарного року Y у єдине середньорічне значення (19).

$$NDVI_{annual}(Y) = \frac{1}{N_Y} \sum_{t_i \in Y} NDVI_i, \quad (19)$$

де N_Y є кількістю успішних безхмарних супутникових спостережень у році Y. Після цього виконується класичне реляційне злиття за ключем року: $Year_{NDVI} = Year_{CO2}$.

Помітний недолік методу – повне нівелювання інформації про сезонну динаміку вегетації внаслідок застосування річної агрегації. Алгоритм машинного навчання втрачає свою базову здатність аналізувати, як макротренд впливає на пікову літню біомасу, оскільки бачить лише згладжену усереднену картину.

Інший підхід у математичні інтерполяції, іншими словами, у оптимальному виборі для системи. Цінна високочастотна інформації з супутникових датчиків може бути збережена, для цього розроблювана методологія базується на підвищенні частоти глобальних даних шляхом математичної інтерполяції. Замість усереднення NDVI, система розраховує або генерує щоденні значення CO2 на основі вже наявних річних відправних точок.

Нехай ми маємо $CO2_{Y_k}$ – це відоме середньорічне значення концентрації парникового газу для року Y_k. У межах алгоритму це значення умовно прив'язується до геометричної середини року. Для знаходження рівня CO2 у будь-яку довільну точну дату t, яка відповідала б дню зйомки супутником, система застосовує лінійну інтерполяцію (20).

$$CO2(t) = CO2_{Y_k} + \frac{CO2_{Y_{k+1}} - CO2_{Y_k}}{t_{k+1} - t_k} \times (t - t_k), \quad (20)$$

де $t_k \leq t < t_{k+1}$. Для досягнення плавнішого, навіть нелінійного переходу між роками у системі також може бути задіяна кубічна сплайн-інтерполяція 3-го порядку, яка математично описує зміну концентрації газів у вигляді набору поліноміальних функцій, утворюючи безперервний градієнт макрозмінної [51].

Такі математична маніпуляція дозволяє ідеально синхронізувати часові мітки: кожен із тисяч 16-денних зрізів NDVI отримує своє чітке, унікальне, розраховане значення CO2 для цього конкретного дня. Цим може гарантуватись те, що ансамблева модель на етапі навчання бачитиме повну картину: високочастотну сезонну пилку вегетаційного циклу, яка розгортається на тлі повільно, але при цьому невпинно зростаючої базової лінії глобального потепління.

Поряд із базовим методом лінійної інтерполяції у конвеєрах інтелектуального аналізу великих екологічних даних може також постати необхідність врахування складної нелінійної природи глобальних атмосферних і біосферних процесів. Головне обмеження кусково-лінійного

підходу полягає в утворенні кутових зламів у місцях з'єднання суміжних інтервалів, тобто у геометричних серединах календарних років, де фіксуються опорні референси концентрації діоксиду вуглецю. З погляду математичного моделювання, такі злами означають розрив першої похідної часового ряду. Це зазвичай вказує на штучні раптові зміни швидкості накопичення парникових газів, які по суті не мають жодного фізичного підґрунтя у реальній термодинаміці атмосфери. Для подолання подібного інженерного обмеження та формування гладкого і двічі безперервно диференційованого градієнта ознак у розроблюваній методології задіяно математичний апарат кубічної сплайн-інтерполяції 3-го порядку.

Нехай задано дискретну упорядковану сітку часових міток $T = \{t_0, t_1, \dots, t_n\}$, де кожен вузол t_i відповідає точній даті геометричної середини календарного року Y_k, для якої з текстових баз даних NOAA отримано верифіковані середньорічні значення концентрації парникового газу $f(t_i) = y_i$. Саме завдання полягає у побудові інтерполяційної функції $S(t)$, котра протягом усього досліджуваного багаторічного відрізку $[t_0, t_n]$ є неперервною разом зі своїми першою та другою похідними, а на кожному окремому частковому інтервалі $[t_i, t_{i+1}]$, довжина якого становить $\Delta t_i = t_{i+1} - t_i$, описується поліномом третього ступеня (21).

$$S_i(t) = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3, \qquad t \in [t_i, t_{i+1}] \quad (21)$$

Для однозначного визначення невідомих коефіцієнтів сплайна $\{a_i, b_i, c_i, d_i\}$ для усіх $i = 0, 1, \dots, n-1$ формується система лінійних алгебраїчних рівнянь [52]. Ця система базується на сукупності фундаментальних умов інтерполяції, згладжування та неперервності. По-перше, умова проходження кубічного сплайна через усі задані опорні точки кліматичного моніторингу вимагає виконання певних рівностей.

$$\begin{gathered} S_i(t_i) = y_i \Rightarrow\ a_i = y_i \\ S_i(t_{i+1}) = y_{i+1} \Rightarrow a_i + b_i\Delta t_i + c_i(\Delta t_i)^2 + d_i(\Delta t_i)^3 = y_{i+1} \end{gathered} \quad (22)$$

По-друге, обмеження на першу та другу похідні (швидкість зміни концентрації парникових газів та прискорення процесу накопичення

відповідно) накладаються через вимоги гладкості та безперервності динамічних трендів у внутрішніх вузлах стику суміжних часових інтервалів. Для кожного внутрішнього вузла t_{i+1}, де $i = 0, 1, \dots, n-2$, умови неперервності мають наступний математичний вигляд (23).

$$\begin{gathered} S'_i(t_{i+1}) = S'_{i+1}(t_{i+1}) \Rightarrow b_i + 2c_i\Delta t_i + 3d_i(\Delta t_i)^2 = b_{i+1} \\ S''_i(t_{i+1}) = S''_{i+1}(t_{i+1}) \Rightarrow 2c_i + 6d_i\Delta t_i = 2c_{i+1} \end{gathered} \quad (23)$$

Виразивши коефіцієнти b_i та d_i через параметри другого порядку c_i, котрі фізично визначають кривину траєкторії часового ряду в кожній опорній точці, систему рівнянь для внутрішніх вузлів можна взагалі звести до класичного тридіагонального вигляду відносно невідомих компонентів c_i.

$$\Delta t_{i-1}c_{i-1} + 2(\Delta t_{i-1} + \Delta t_i)c_i + \Delta t_i c_{i+1} = 3\left(\frac{y_{i+1} - y_i}{\Delta t_i} - \frac{y_i - y_{i-1}}{\Delta t_{i-1}}\right) \quad (24)$$

Оскільки кількість невідомих коефіцієнтів перевищує кількість сформованих рівнянь зв’язку на дві одиниці, для досягнення єдиного та стабільного розв’язку системи вводяться так звані природні граничні умови по краям загального часового відрізку моніторингу. Згідно цих умов припускається, що за межами експериментального вікна спостережень прискорення накопичення макрозмінної стабілізується, тобто друга похідна функції на початку t_0 та в кінці t_n часового ряду дорівнює нулю (25).

$$\begin{gathered} S''(t_0) = 0 \Rightarrow c_0 = 0 \\ S''(t_n) = 0 \Rightarrow c_n = 0 \end{gathered} \quad (25)$$

Отримана лінійна система із тридіагональною матрицею має властивість строгої діагональної домінантності. Відповідно, так гарантується існування єдиного розв’язку та дозволяється застосувати високооптимізований метод виключення з обчислювальною складністю лінійного типу $O(n)$. Після розрахунку усіх коефіцієнтів $\{a_i, b_i, c_i, d_i\}$, розроблена технологія здійснює потокове обчислення точних значень CO2 для будь-якої довільної проміжної часової мітки супутникової зйомки $t \in [t_i, t_{i+1}]$. Це у свою чергу якраз відповідає моменту фіксації спектральних характеристик рослинного покриву сенсорами місій супутників.

Перехід від лінійної до кубічної сплайн-інтерполяції третього порядку моєе теоретично мати вирішальне кінцеве значення за потреби підвищення якості подальшого навчання ансамблевих моделей машинного навчання. Оскільки алгоритм регресії RandomForestRegressor здійснює просторове розділення простору ознак шляхом побудови ортогональних гіперплощин уздовж осей, наявність фізично адекватного розподілу інтерпольованих точок полегшить пошук оптимальних порогів розгалуження в процесі побудови дерев рішень.

Перед тим, як передати створений за допомогою процедури ETL об’єднаний датасет у складну модель ансамблевого навчання, методологія передбачає проведення базового кореляційного аналізу. Це, звісно, необхідно для попередньої статистичної валідації гіпотези стосовно наявності конкретного взаємозв’язку між глобальними драйверами та локальною біосферою.

Аналіз проводиться як розрахунок вибіркового коефіцієнта лінійної кореляції Пірсона r, котрий формалізується як коваріація двох змінних, поділена потім на добуток їхніх стандартних відхилень (26).

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \quad (26)$$

де X_i – інтерпольоване значення CO2, Y_i – відповідне значення регіонального NDVI для i-го спостереження, а $\bar{X}$ та $\bar{Y}$ – їх вибіркові середні значення відповідно [45].

Залежно від розподілу даних, цілком може також застосовуватись і ранговий коефіцієнт кореляції Спірмена r, стійкий до нелінійних монотонних зв’язків та наявності аномальних викидів у супутникових даних [52].

Хоча збільшення концентрації вуглецю в атмосфері у стерильних лабораторних умовах стимулює процес фотосинтезу, тим не менш у реальних відкритих екосистемах цей ефект може нівелюватись супутніми факторами глобального потепління на кшталт посух, теплових хвиль чи деградації

ґрунтів. Це в принципі підтверджується дослідженнями про виснаження асиміляційної здатності лісів. Виявлення стабільної статистично значущої від'ємної кореляції між усіма цими рядами може послугувати маркером кліматичної вразливості досліджуваного полігону, підтверджуючи коректність виконаної процедури синхронізації та обґрунтовуючи подальше застосування складних нелінійних моделей машинного навчання для глибокого предиктивного аналізу.

2.5. Критерії оцінки якості інтелектуальних моделей

Відповідно до загальноприйнятого індустріального стандарту інтелектуального аналізу даних, наступним повинен бути етап оцінки моделі [26]. Критично важливо підтвердити її наукову обґрунтованість та визначення практичної значущості. Алгоритми машинного навчання повинні не просто тупо запам'ятовувати історичні дані. Вони мають володіти високою здатністю до узагальнення на нових, раніше ніколи не бачених проміжках часу. Окрім цього, усі виявлені статистичні взаємозв'язки між локальною рослинністю та макротрендами потребують, звісно, суворої перевірки на математичну значущість. Аби кількісно виміряти усі ці властивості використовується певний комплекс різноманітних статистичних метрик та критеріїв.

Оскільки головне завдання розробленої системи полягає у прогнозуванні безперервних значень індексу NDVI на основі макрокліматичних факторів, базовим набором метрик для валідації результатів можуть стати абсолютні, відносні та квадратичні показники втрат, які якраз реалізовані у бібліотеці scikit-learn, у модулі metrics [28].

Середня абсолютна помилка MAE (Mean Absolute Error) обчислює середнє арифметичне абсолютних відхилень спрогнозованих значень від фактичних, отриманих із супутника [45]. Математично вона формалізується за формулою (26).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (27)$$

де y_i – реальне значення NDVI, яке було перед цим розраховане за допомогою інструментарію xarray, $\hat{y}_i$ – прогнозоване значення, а n – кількість спостережень у вибірці. Висока інтерпретованість є безперечно перевагою MAE. Вона вимірюється в тих самих одиницях, у яких обчислюють і цільову змінну, а це від -1 до 1 для NDVI. Додатково метрика стійка до поодиноких екстремальних викидів, які із дуже високим шансом можуть трапитись у даній задачі. Наприклад, аномально низькі значення індексу через невідфільтровану тінь від хмари. MAE рівномірно оцінює усі похибки.

Для представлення результатів кінцевим користувачам, а це можуть бути екологи чи аграрії, абсолютні значення похибки часто виявляються недостатньо наочними. Метрика MAPE виражає відхилення прогнозу у відсотках відносно фактичних значень (27).

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (28)$$

MAPE вцілому дає швидко оцінити відносну точність моделі. Якщо MAPE становить 5%, це означає, що в середньому прогноз моделі відхиляється від реального рівня вегетації лише на 5%. Важливою математичною особливістю цієї метрики є чутливість до нульових значень. Проте, оскільки цільові значення здорової рослинності зазвичай лежать у межах 0,2-0,8, то використання MAPE у даній предметній області є цілком виправданим і безпечним.

Неможливо також не виокремити середньоквадратичну помилку MSE та корінь із неї RMSE (29).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad RMSE = \sqrt{MSE} \quad (29)$$

Від MAE їх відрізняє те, що квадратичні метрики експоненційно накладають штрафи на модель за надто великі відхилення. У екологічному відстеженні це доволі критично. У випадку, якщо система помилиться на 0,05 одиниць NDVI протягом усього року – це цілкою прийнятна похибка із якою

можна змиритись. Але якщо вона пропустить раптове падіння NDVI на цілих 0,4 одиниць під час екстремальної літньої посухи, це зробить систему абсолютно непридатною для можливого попередження криз. Саме тому мінізація RMSE гарантує, що модель надійно зафіксує подібні сильні кліматичні аномалії.

Коефіцієнт детермінації R^2 оцінює загальну пояснювальну здатність моделі [45]. Показник (30) вказує на те, яку частку дисперсії або мінливості нашої залежної змінної можна було б пояснити за допомогою незалежних змінних.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (30)$$

де $\bar{y}$ – середнє значення фактичних спостережень. Для кліматичних систем, як та, що розроблюється, ідеальний $R^2 = 1$ є, вочевидь, недосяжним банально через природну стохастичність атмосфери та біосфери. Тим не менш, максимізація цього показника позитивно свідчитиме про успішне виявлення моделлю нелінійних зв'язків між глобальними макротрендами та локальною вегетацією.

Базове завдання системи хоча і полягає у регресійному прогнозуванні часових рядів, у процесі екологічного моніторингу нерідко виникає потреба переходу до задач бінарної або й зовсім багатокласової класифікації. Наприклад, коли безперервні значення NDVI порогово діляться на класи: «Норма» чи «Посуха». Для оцінки якості схожих моделей використовується зовсім інший математичний апарат.

Основою розрахунку класифікаційних метрик можна вважати матрицю помилок [45]. Це така таблиця розмірності $N \times N$, де N – кількість класів, яка спрямована на візуалізацію продуктивності алгоритму (табл. 2.2).

Таблиця 2.2 – Матриця невідповідностей

Тип бінарної класифікації	Екологічний контекст та інтерпретація в інтелектуальній системі моніторингу

True Positives (TP)	Модель правильно спрогнозувала наявність кліматичної аномалії, і та дійсно відбулася за даними супутника.
True Negatives (TN):	Модель правильно спрогнозувала нормальний стан вегетації.
False Positives (FP)	Помилка I роду, коли модель спрогнозувала посуху, але насправді вегетація була в нормі, це так звана «хибна тривога».
False Negatives (FN)	Помилка II роду, коли модель спрогнозувала нормальний стан, хоча насправді відбулася посуха, тобто це пропуск аномалії.

Відповідно, на основі матриці помилок розраховуються ключові показники якості. Одною з таких є точність. Це частка правильних позитивних прогнозів серед взагалі усіх позитивних прогнозів моделі. Вона показує, наскільки в принципі можна довіряти системі, коли вона сигналізує про, наприклад, посуху.

$$Precision = \frac{TP}{TP + FP} \quad (31)$$

Повнота це частка дійсно знайдених аномалій серед взагалі усіх реальних аномалій. У даній задачі така метрика часто взагалі найважливіша, бо пропуск потенційної посухи велике значення FN має значно гірші економічні наслідки, ніж просто хибна тривога.

$$Recall = \frac{TP}{TP + FN} \quad (32)$$

Гармонійне середнє між точністю та повнотою це F1-Score. Показник ідеально підходить для дуже незбалансованих вибірок, наприклад, коли днів із посухою значно менше, ніж звичайних.

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \quad (33)$$

У процесі поєднання NDVI із глобальними кліматичними макротрендами розраховуються коефіцієнти кореляції, зазвичай це коефіцієнти Пірсона або Спірмена. Тим не менш, саме по собі високе або

низьке значення коефіцієнта кореляції усе одно не гарантує остаточно, що виявлений зв’язок є реальним фізичним явищем, а не артефактом випадкового збігу зашумлених даних.

Для математичного доведення надійності виявлених тенденцій дуже розповсюдженим є застосування оцінки статистичної значущості, вираженої за допомогою p-value. Процес перевірки базується на концепції нульової гіпотези. Грубо кажучи, H_0 стверджує, що між глобальним рівнем CO2 та зниженням локального NDVI взагалі немає жодного статистичного зв’язку, а будь-яка виявлена кореляція є всього лише наслідком випадковості нашої вибірки.

Альтернативна ж гіпотеза H_1 натомість говорить про те, що такий статистичний зв’язок все ж існує.

Показник p-value, у свою чергу, це шанс отримати таке ж саме або ще навіть більш екстремальне значення коефіцієнта кореляції за тої умови, що нульова гіпотеза H_0 є абсолютно вірною. У наукових кліматологічних дослідженнях, от зокрема, у звітах IPCC, загальноприйнятим рівнем статистичної значущості α є поріг у 0,05, тобто поріг у 5% [1, 52].

Таким чином, якщо розраховане $p_{value} < 0{,}05$, то ми впевнено відхиляємо нульову гіпотезу. Це у свою чергу означає, що ймовірність випадкового збігу становить менше 5%, і виявлений негативний вплив парникових газів на фітоценоз це статистично достовірне явище.

У іншому випадку, якщо ж $p_{value} \geq 0{,}05$, нульова гіпотеза не може у жодному разі бути відхилена. Тоді, навіть якщо графік тренду машинного навчання показує спад, ми не маємо жодного математичного права стверджувати, що він спричинений саме глобальними змінами. Можливо, даних замало або дисперсія сезонності занадто висока. Відповідно, такий суворий контроль p-значення повинен у підсумку запобігти формуванню хибних наукових висновків під час роботи з Big Data. Аналогічно він гарантує, що розроблена інтелектуальна система продукує об’єктивну екологічну аналітику.

Базових метрик регресії часто буває недостатньо для повноцінної оцінки систем, що працюють із просторово-часовими екологічними даними. Оскільки кліматичні зміни мають виражений кумулятивний характер (поступове накопичення парникових газів), важливо не лише оцінити абсолютний розмір помилки, але й її напрямок (чи модель систематично переоцінює здоров'я екосистеми, чи недооцінює його). Для цього у методологію впроваджуються специфічні метрики.

Середня помилка зміщення (Mean Bias Error) MBE (34) дозволяє виявити систематичне зміщення інтелектуальної моделі [53].

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (34)$$

У разі коли $MBE > 0$, це свідчить про те, що розроблена модель систематично завищує показники NDVI, а це доволі небезпечний сценарій, бо у такий спосіб можуть приховуватись процеси поступового опустелювання або деградації рослинного покриву під впливом парникового ефекту. При випадках коли $MBE < 0$, система може виявитись надмірно песимістичною. Ідеальна модель повинна мати MBE, наближене до нуля, що свідчить про відсутність систематичного зсуву.

З огляду на обґрунтований у підрозділі 2.2 вибір алгоритму Random Forest, зараз виникає можливість використання специфічної метрики оцінки, яка не потребує виділення окремої валідаційної вибірки. Під час формування дерев рішень методом беггінгу, кожне дерево навчається лише на вже згадані $\sim 63{,}2\%$ унікальних даних. Тим часом, спостереження, які не потрапили у вибірку для конкретного дерева, називаються Out-of-Bag (OOB).

При прогонці цих невикористаних супутникових композитів крізь побудоване дерево та при усередненні результатів відносно усього лісу, ми отримуємо об'єктивну оцінку узагальнюючої здатності моделі на досі небачених даних, або ж OOB Score [42]. Це також одна з метрик, використання якої допоможе із економією обмежених вибірок агрегованих річних кліматичних макротрендів.

Первинне експериментальне застосування побудованого конвеєра на реальних інтерпольованих даних виявило певну специфічну аномалію у поведінці навченої моделі (рис. 2.1). При отриманні доволі високих формальних показників якості на тренувальних даних, коефіцієнт детермінації $R^2 = 0{,}789$ та відносно низька середня абсолютна відсоткова похибка $MAPE = 5{,}71\%$, графічна візуалізація предиктивного тренду показує занадто високу просторово-часову нестабільність. Червона пунктирна лінія тренду машинного навчання набуває тут абсолютно хаотичного, ламаного вигляду. Можна зробити припущення, що модель намагалась математично підлаштуватися під кожне окреме спостереження високочастотної сезонної пилки NDVI. Одночасно із цим, розрахована внутрішньоансамблева оцінка продемонструвала критичне від'ємне значення, де OOB Score рівний $-0{,}707$. З огляду на розібрану вище математичну природу цієї метрики, результат можна впевнено вважати беззаперечним маркером явища перенавчання.

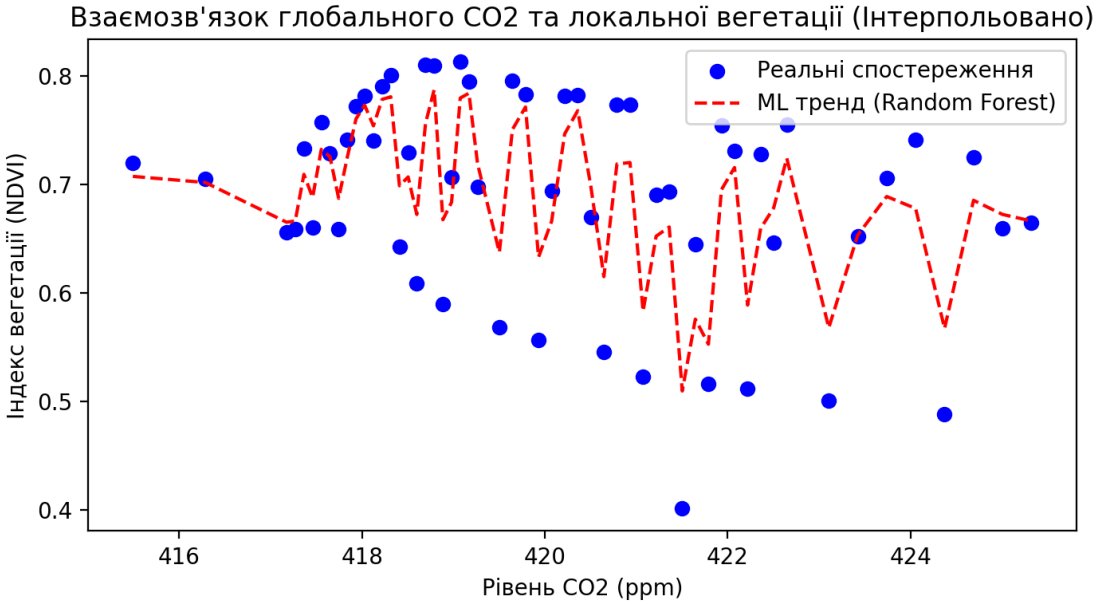

Рисунок 2.1 – Результат перенавчання моделі

Судячи з усього, алгоритм Random Forest із базовими гіперпараметрами замість узагальнення глобального кліматичного тренду почав буквально заучувати структурі листя окремі флуктуації. Це могло трапитись через відносну обмеженість за обсягом результуючого синхронізованого масиву даних. Він охоплює декілька 16-денних зрізів лише за короткий дворічний

період. Саме тому при прогонці Out-of-Bag спостережень крізь надміру глибокі дерева рішень, сумарна дисперсія помилок прогнозу перевищила базову дисперсію цільової змінної. Це у свою чергу і зумовило негативний показник узагальнюючої здатності. Такий наочний приклад обґрунтовує наукову необхідність переходу від базової конфігурації алгоритму до етапу цілеспрямованої регуляризації та оптимізації структурних параметрів випадкового лісу для отримання генералізованого макротренду.

Додатково, важливо згадати про теоретичну та методологічну умову забезпечення високої прогностичної спроможності розроблюваної системи, яка полягає у розв'язанні фундаментальної математичної проблеми, пов'язаної з компромісом між системним зміщенням та статистичною дисперсією алгоритму машинного навчання [45]. Будь-яка непараметрична модель регресії, до класу яких належать безпосередньо і дерева рішень, під час апроксимації прихованих функціональних залежностей на основі обмеженої навчальної вибірки схильна до двох протилежних інженерних ризиків: недонавчання або перенавчання.

Припустимо, що істинна, але прихована від спостереження стохастична залежність між просторово-часовими екологічними показниками описується теоретичним рівнянням вигляду $y = f(X) + \epsilon$, де X це вектор вхідних ознак, зокрема, інтерпольованим часовим градієнтом концентрації CO2, $f(X)$ – істинною детермінованою функцією кліматичного тиску, а ϵ – випадковим екзогенним шумом системи із нульовим математичним сподіванням $\mathbb{E}[\epsilon] = 0$ та гомоскедастичною дисперсією $Var(\epsilon) = \sigma^2$. Якщо алгоритм машинного навчання на основі випадкової реалізації навчального набору даних D будує апроксимуючу залежність $\hat{f}(X;\ D)$, то математичне сподівання квадратичної помилки узагальнення у довільній фіксованій точці простору ознак X може бути розкладене на три ортогональні складники (35).

$$\mathbb{E}_D\left[\left(y - \hat{f}(X;D)\right)^2\right] = Bias\left[\hat{f}(X)\right]^2 + Var\left[\hat{f}(X)\right] + \sigma^2 \quad (35)$$

Для детального аналізу природи виникнення похибок наведемо повне математичне виведення цієї декомпозиції, опустивши для спрощення запису явне позначення залежності від вибірки D та точки X, і використовуючи базову властивість лінійності оператора математичного сподівання, а також рівність $\mathbb{E}[y] = f(X)$. З огляду на те, що істинне значення у та побудований прогноз $\hat{f}$ є статистично незалежними змінними, квадрат відхилення розгортається за формулою (36).

$$\begin{gathered}\mathbb{E}\left[\left(y-\hat{f}\right)^2\right] = \mathbb{E}\left[\left((f(X)+\epsilon)-\hat{f}\right)^2\right] = \mathbb{E}\left[\left(\left(f(X)-\hat{f}\right)+\epsilon\right)^2\right] \\ \mathbb{E}\left[\left(y-\hat{f}\right)^2\right] = \mathbb{E}\left[\left(f(X)-\hat{f}\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\hat{f}\right)\epsilon\right] + \mathbb{E}[\epsilon^2]\end{gathered} \quad (36)$$

Оскільки випадковий шум ϵ має нульове середнє значення та є незалежним від моделі, перехресний член перетворюється на нуль: $2\mathbb{E}\left[f(X)-\hat{f}\right]\mathbb{E}[\epsilon]$. Враховуючи, що $\mathbb{E}[\epsilon^2] = Var(\epsilon) = \sigma^2$, отримуємо формулу (37).

$$\mathbb{E}\left[\left(y-\hat{f}\right)^2\right] = \mathbb{E}\left[\left(f(X)-\hat{f}\right)^2\right] + \sigma^2 \quad (37)$$

Далі проводиться внутрішнє центрування апроксимуючої функції $\hat{f}$ відносно її власного математичного сподівання $\mathbb{E}\left[\hat{f}\right]$, котре відображає усереднену поведінку алгоритму за умови його багаторазового навчання на різних випадкових підвибірках генеральної сукупності даних.

$$\begin{gathered}\mathbb{E}\left[\left(f(X)-\hat{f}\right)^2\right] = \mathbb{E}\left[\left(\left(f(X)-\mathbb{E}\left[\hat{f}\right]\right)+\left(\mathbb{E}\left[\hat{f}\right]-\hat{f}\right)\right)^2\right] \\ \mathbb{E}\left[\left(f(X)-\hat{f}\right)^2\right] = \mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}\right]\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}\right]\right)\left(\mathbb{E}\left[\hat{f}\right]-\hat{f}\right)\right] + \\ +\mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}\right]-\hat{f}\right)^2\right]\end{gathered} \quad (38)$$

У цьому виразі перший доданок є квадратом детермінованої величини, тому оператор математичного сподівання на нього не впливає. У другому (перехресному) доданку перший множник також є константою відносно вибірки, що дозволяє винести його за дужки оператора (39).

$$2\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}\right]\right)\left(\mathbb{E}\left[\hat{f}\right]-\hat{f}\right)\right] = 2\left(f(X)-\mathbb{E}\left[\hat{f}\right]\right)\cdot\mathbb{E}\left[\mathbb{E}\left[\hat{f}\right]-\hat{f}\right] \quad (39)$$

Оскільки за визначенням $\mathbb{E}\left[\mathbb{E}\left[\hat{f}\right]-\hat{f}\right] = \mathbb{E}\left[\hat{f}\right]-\mathbb{E}\left[\hat{f}\right] = 0$, перехресний член знову повністю нівелюється. Третій доданок за своїм математичним

визначенням є не чим іншим, як статистичною дисперсією прогнозованих значень алгоритму. Таким чином, фінальне зведення рівняння підтверджує трикомпонентну структуру сумарної похибки узагальнення.

$$\mathbb{E}\left[\left(y-\hat{f}\right)^2\right] = \left(f(X)-\mathbb{E}[\hat{f}]\right)^2 + \mathbb{E}\left[\left(\hat{f}-\mathbb{E}[\hat{f}]\right)^2\right] + \sigma^2 =$$

$$= Bias\left[\hat{f}(X)\right]^2 + Var\left[\hat{f}(X)\right] + \sigma^2 \quad (40)$$

Кожен із отриманих компонентів має чітку біофізичну та алгоритмічну інтерпретацію в контексті розроблюваної системи. Перший член, квадрат системного зміщення $Bias^2$, характеризує міру грубості математичних припущень самого алгоритму. Високе зміщення властиве простим лінійним моделям, які банально неспроможні відстежити складні нелінійні порогові переходи в екосистемах. Другий член, статистична дисперсія Var, вимірює рівень чутливості та нестабільності прогнозів моделі відносно дрібних коливань у навчальному наборі ознак. Третій член σ^2 репрезентує неусувну похилку, яка принципово не може бути зменшена жодним інструментом Data Science, бо вона за своєю суттю відображає природний хаотичний шум і стохастичну природу кліматичних процесів.

Аналіз поведінки окремих глибоких дерев рішень, котрі формують базис випадкового лісу, свідчить про те, що за відсутності примусових обмежень на їхній ріст, вони мають гранично низьке, по суті майже нульове, зміщення, однак величезну статистичну дисперсію. Коли таке нерегуляризоване дерево навчається на невеликому датасеті, де низькочастотний довгостроковий макротренд зіставлений із високочастотними супутниковими вибірками, алгоритм починає здійснювати надлишкове ортогональне розщеплення простору ознак. Дерево буквально заучує амплітуду сезонної вегетаційної пилки, кодуючи у своїй термінальній структурі випадковий шум та флуктуації, викликані атмосферними завадами чи хмарністю. Як наслідок, на тренувальній вибірці модель видає бездоганні формальні показники, але при тестуванні на незалежних даних вибірці OOB похибка узагальнення різко зростає через високу дисперсію ансамблю.

Для розв’язання такої проблеми у архітектуру конвеєра було прийнято рішення впровадити механізм цілеспрямованої структурної регуляризації випадкового лісу шляхом оптимізації його ключових структурних гіперпараметрів. Примусова фіксація максимальної глибини дерев на низькому рівні разом зі встановленням мінімального порогу для формування термінального листка дозволило хоч і штучно, але змістити модель вздовж кривої компромісу Bias-Variance. Обмеження глибини трьома рівнями обриває можливість побудови складних ламаних гілок, змушуючи алгоритм ігнорувати внутрішньорічну сезонну мінливість і фокусуватися виключно на макрокліматичному сигналі. Регуляризація свідомо призводить до незначного зростання системного зміщення $Bias$, котре візуалізується як закономірне падіння формального коефіцієнта детермінації на тренувальних даних, проте воно зменшує статистичну дисперсію Var всього ансамблю. Завдяки зрізанню надлишкової чутливості дерев до шуму, сумарна очікувана помилка моделі мінімізується, лінія тренду в підсумку набуває стабільного плавного вигляду, а метрика OOB Score максимізується, затверджуючи перехід інтелектуальної системи у стан стійкого глобального узагальнення кліматичних закономірностей.

Знову ж таки, оскільки розроблена система оперує часовими рядами, то в такому випадку традиційна випадкова їхня розбивка на тренувальні та тестові масиви Random K-Fold вважається математично некоректною. Нехтування цьому призводить до так званого витоку даних, іншими словами, модель може використовувати інформацію з майбутнього для прогнозування минулого, як би дивно це не звучало. Для запобігання побідних інцидентів, у рамках методології застосовується метод послідовної часової валідації, коли модель завжди навчається на строго попередніх періодах і відповідно тестується виключно на наступних хронологічних відрізках [54]. Це у підсумку повільно максимально наблизити оцінку моделі до реальних умов експлуатації системи моніторингу.

Висновки

У другому розділі було здійснено глибоку формалізацію методів та алгоритмів, котрі в принципі становлять математичне ядро розроблюваної інформаційної системи.

Насамперед, було проведено математичну формалізацію базового індексу NDVI та обґрунтовано необхідність використання оптимізованих індексів EVI та SAVI для боротьби з явищами перенасичення спектрального сигналу та впливом фонового ґрунтового шуму. Описано також процедуру просторово-часової агрегації та фільтрації аномалій за допомогою інструментарію xarray.

Теоретично обґрунтовано вибір ансамблевої моделі RandomForestRegressor як базового алгоритму машинного навчання. Математично розкрито принципи розбиття граф-структур, такі як ентропія, індекс Джині, та повністю описано механіку уникнення перенавчання за рахунок методів Bagging та випадкових підпросторів. Інтеграція критерію оцінки важливості ознак дозволила перетворити модель з «чорного ящика» на дійсно інструмент інтерпретованої наукової аналітики.

Розроблено також строгу методологію ETL для синхронізації глобальних кліматичних макротрендів з локальними супутниковими таймсеріями. Обумовлено застосування математичної лінійної та сплайн-інтерполяції для вирівнювання частоти дискретизації часових рядів перед регресійним моделюванням.

Сформовано комплексний набір метрик оцінки якості, який адаптований для завдань регресії часових рядів. Додатково розширено апарат валідації класифікаційними метриками для кращої детекції гідротермічних стресів. Доведено необхідність використання порогу статистичної значущості для відокремлення справжніх кліматичних впливів від випадкового шуму, а це є вкрай важливим пунктом в контексті дослідження.

Отримана математична база дозволяє перейти до етапу безпосередньої програмної реалізації, проєктування архітектури коду та тестування системи на реальних даних.

РОЗДІЛ 3. ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ТА ПРЕДИКТИВНЕ МОДЕЛЮВАННЯ КЛІМАТИЧНИХ ДАНИХ

3.1. Програмна реалізація інтелектуального конвеєра

Успішна екстракція просторових масивів із бінарних файлів вже сама по собі сформувала базис подальшої аналітичної роботи, щоправда самі по собі необроблені ніяк дані не мають жодної предиктивної цінності. Відповідно, наступною стадією є розробка програмного конвеєра, який здатен автоматично трансформувати піксельні матриці у структуровані часові ряди та інтегрувати їх із глобальними кліматичними макротрендами для навчання предиктивних моделей. Задача вирішується у два взаємопов'язані етапи: математична агрегація та безпосередньо машинне навчання.

Вимогою до універсальності розробленої системи можна виокремити здатність працювати з різними супутниковими продуктами, які часто мають відмінну номенклатуру внутрішніх змінних. У файлах NASA, як відомо, можуть міститися десятки різноманітних шарів, велика кількість яких абсолютно не підходять для даного дослідження: відбивальна здатність у різних каналах, маски якості, кути зеніту сонця тощо. Відповідно, постає чергова проблема жорсткої прив'язки до назв змінних. У класі UniversalSatelliteDataProcessor розроблено алгоритм семантичного пошуку каналів якраз для подолання цієї перешкоди (рис. 3.1).

```python
def extract_bands(self):
    extracted_vars = {}
    all_vars = list(self.dataset.data_vars)
    red_keys = ["red","reflectance_01","surfreflect_i1","surfreflect_m1","b01","band1",]
    nir_keys = ["nir","reflectance_02","surfreflect_i2","surfreflect_m2","b02","band2",]
    ndvi_keys = ["ndvi", "vegetation_index", "vi_500m"]

    for v in all_vars:
        v_lower = v.lower()
        if "NDVI" not in extracted_vars and any(k in v_lower for k in ndvi_keys):
            extracted_vars["NDVI"] = self.dataset[v]
        elif "Red" not in extracted_vars and any(k in v_lower for k in red_keys):
            extracted_vars["Red"] = self.dataset[v]
        elif "NIR" not in extracted_vars and any(k in v_lower for k in nir_keys):
            extracted_vars["NIR"] = self.dataset[v]
    return extracted_vars
```

Рисунок 3.1 – Програмна реалізація методу семантичного пошуку каналів

Програмна реалізація алгоритму базується на використанні попередньо заданих словників ключових слів та відповідних масивів для NIR і NDVI. Конвеєр поступово ітерується списком усіх доступних змінних в межах завантаженого датасету та приводить їхні назви до нижнього регістру. Слідом, за допомогою логічних операторів *any*(), система автоматично розпізнає та виокремлює необхідні канали: червоний або ближній інфрачервоний, або навіть вже готовий шар нормалізованого відносного індексу рослинності, якщо такий взагалі надається провайдером.

У випадку, коли готовий NDVI продукт відсутній, а це цілком характерно для роботи з необробленими даними відбивальної здатності, система автоматично переходить до його математичного обчислення у процесі роботи. За допомогою бібліотеки xarray для цього застосовуються векторизовані операції над багатовимірними тензорами. Може, звісно, виникнути критична помилка ділення на нуль, яка перериватиме роботу конвеєра при аналізі, тому було заздалегідь реалізовано алгоритм уникнення подібної сингулярності (рис. 3.2).

```
if "NDVI" not in bands and "Red" in bands and "NIR" in bands:
    red, nir = bands["Red"], bands["NIR"]
    denominator = xr.where((nir + red) == 0, 1e-10, (nir + red))
    target_var = (nir - red) / denominator
    metric_name = "NDVI"
else:
    metric_name = list(bands.keys())[0]
    target_var = bands[metric_name]
```

Рисунок 3.2 – Програмна реалізація уникнення помилки ділення на нуль

У якості просторового фільтру працює конструкція *xr.where*. Її мета у заміні нульових знаменників на нескінченно малу константу 10^{-10}, чим вона гарантує стабільніший розрахунок індексу для кожного пікселя регіону. Отриманий тензор все ще містить просторові координати. Для переходу від просторового масиву до часового ряду повинна здійснюватись просторова агрегація. Програмний конвеєр динамічно визначає правильні просторові виміри та застосовує функцію середнього значення. У результаті мільйони пікселів згортаються у єдине числове значення, яке характеризує середній стан

вегетації у регіоні на конкретну дату. Заключним елементом підмодуля є конвертація агрегованого масиву xarray у плоску таблицю pandas.

Регіональна статистика сама по собі ілюструє виключно симптоми кліматичних змін. Для виявлення саме що їх глибинних причин і побудови прогнозних моделей конвеєр повинен поєднати локальні індикатори екосистеми із глобальними трендами. Цю функцію інкапсульовано у класі ClimateIntelligenceSystem. Ядром тут виступає ансамблевий алгоритм RandomForestRegressor з бібліотеки scikit-learn.

Програмна інтеграція розпочинається з виклику підпрограми фонового HTTP-запиту з ціллю завантаження актуальних даних про концентрацію діоксиду вуглецю безпосередньо з текстових баз даних NOAA (рис. 3.3). Отриманий глобальний масив і розрахована раніше регіональна статистика мають різну дискретність: річну та 16-денну відповідно.

```
df_global["Date"] = pd.to_datetime(df_global["year"].astype(str) + "-07-01")
df_global = df_global[["Date", "co2"]]
df_merged = pd.merge(
    df_regional, df_global, on="Date", how="outer"
).sort_values("Date")
df_merged["co2"] = df_merged["co2"].interpolate(method="linear")
```

Рисунок 3.3 – Програмна реалізація методу лінійної інтерполяції

На відміну від примітивного реляційного злиття за ключем року, яке повністю знищує сезонну дисперсію даних, вирішення проблеми просторово-часової невідповідності у конвеєрі реалізовано за допомогою імплементації у розроблений конвеєр математичного апарату лінійної інтерполяції. Річні відправні точки NOAA програмно прив’язуються до геометричної середини року, тобто до 1 липня. Після виконання зовнішнього об’єднання таблиць, система застосовує метод інтерполяції. У підсумку алгоритм генерує точне, унікальне значення CO2 для кожного проміжного 16-денного супутникового знімка, створюючи безперервний градієнт концентрації парникових газів.

Сформований і очищений датасет передається далі до модуля машинного навчання. У цій архітектурі матриця ознак X формується з глобальних показників CO2, а у якості вектора цільової змінної y виступають

усереднені значення вегетації регіону. Навчання моделі ініціюється викликом відповідної функції (рис. 3.4).

```
X = df_merged[["co2"]].values.astype(float)
y = df_merged[metric_name].values.astype(float)

self.regressor.fit(X, y)
y_pred = self.regressor.predict(X)
```

Рисунок 3.4 – Програмна реалізація методу навчання моделі

Системі завдяки якісно розписаному програмному підходу не лише виводить два ізольовані графіки, а знаходить приховані нелінійні кореляції між накопиченням парникових газів та реакцією біосфери, тобто помічає кореляцію між макрорівнем та мікрорівнем. Після завершення навчання конвеєр викликає метод для передбачення, який генерує гладку лінію машинного тренду. Після цього на графіку демонструється загальна траєкторію впливу кліматичних змін на досліджувану територію.

3.2. Експериментальне дослідження системи на реальних даних

Для безпосередньо верифікації розроблених алгоритмічних рішень та програмної архітектури інтелектуального конвеєра було проведено серію у декілька експериментальних досліджень. Метою цього експерименту стало, вочевидь, підтвердження здатності системи автономно виконувати повний цикл обробки даних, починаючи динамічним завантаженням супутникових знімків до побудови предиктивних моделей впливу. Узгоджуючи етапи із методологією CRISP-DM, цей крок відповідає фазі моделювання та оцінки отриманих результатів, коли варто перевірити коректність інтеграції різних джерел даних у єдиний аналітичний простір.

Експериментальне дослідження розпочалось, як не дивно, з ініціалізації параметрів цільового регіону через розроблений у минулому підрозділі графічний інтерфейс. Для тестування було обрано географічний полігон, заданий координатами обмежувального прямокутника: мінімальна довгота 29,50°, максимальна довгота 32,00°, мінімальна широта 49,90° та максимальна широта 51,50°. Цей прямокутник просторово охоплює репрезентативну

частину території центральної України, а саме Київську область із яскраво вираженою сезонною динамікою рослинного покриву (рис. 3.5). Тут загалом поєднуються великі лісові масиви та інтенсивні сільськогосподарські угіддя, тобто обрану локація достатньо оптимальна для того, щоб виступати випробувальним майданчиком для тестування алгоритмів екологічного та кліматичного моніторингу.

Рисунок 3.5 – Обрана територія для проведення першого експерименту

Основним джерелом первинних просторових даних цього разу виступив продукт VNP13A1 місії VIIRS. Завдяки інтегрованому модулю NASADataAccess, детально описаному раніше, система успішно здійснила фонову авторизацію на серверах Earthdata та сформувала просторово-часовий пошуковий запит. Незважаючи на те, що в панелі керування ілюстративно вказано короткий період з 2022/06/01 по 2022/08/31, для забезпечення репрезентативності вибірки та адекватної роботи алгоритмів машинного навчання конвеєр опрацював безперервний багаторічний часовий ряд з середини 2022 до кінця 2024 року. Загальний обсяг автоматично завантажених супутникових гранул повною мірою справдив необхідність використання оптимізованих методів обробки великих даних.

Протягом обробки завантаженого масиву універсальний процесор застосував інноваційний алгоритм так званого розумного відкриття.

Ієрархічна бінарна структура HDF5-архівів була успішно автономно просканувана системою. Розроблений інструмент також успішно ідентифікував приховану глибинну папку з науковими даними. Після цього програмний конвеєр проаналізував доступні змінні, загалом 17 шарів, включно із кутами зеніту сонця та подібними файлами, і безпомилково виокремив цільовий канал нормалізованого відносного індексу рослинності. Також успішно були виокреслені базові спектральні канали відбиття NIR reflectance та red reflectance.

Наступним кроком стала масштабна просторова та часова агрегація. Мільйони пікселів, які потрапили у заданий полігон, були усереднені для кожної дати зйомки. При цьому алгоритм ігнорував «биті» значення, потенційно закриті хмарним покривом. Таке стало можливим саме завдяки математичним механізмам розпізнавання та декодування маркерів заповнення у NaN за допомогою бібліотеки xarray.

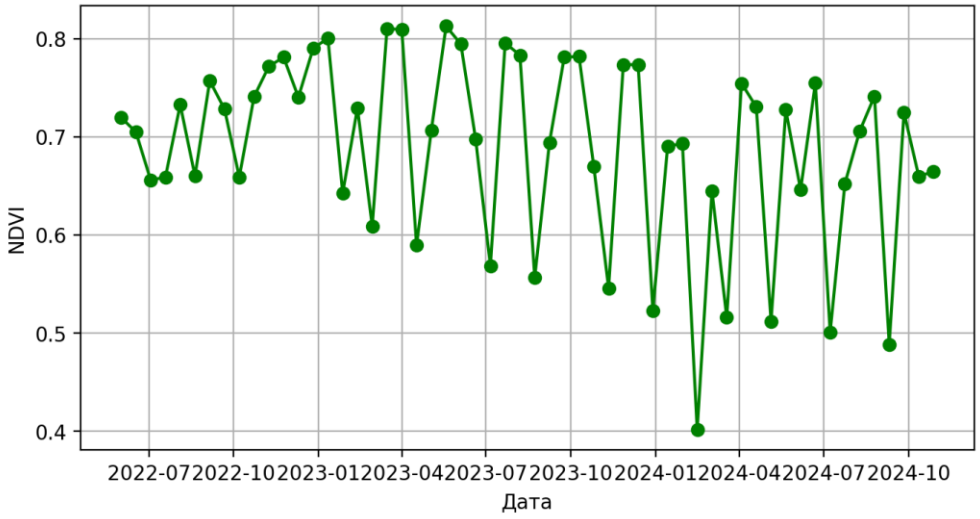

Рисунок 3.6 – Динаміка вегетації у регіоні першого експерименту

Отримана регіональна статистика, відображена на графіку із динамікою вегетації у регіоні (рис. 3.6), показує абсолютно адекватну та фізично обґрунтовану динаміку часового ряду. Значення NDVI для досліджуваного регіону коливалися у достатньо широкому діапазоні: від зимових мінімумів на рівні $\approx 0{,}40$ до літніх максимумів, котрі навіть сягали значення 0,81.

Графічний аналіз динаміки індексу чітко фіксує природну фенологічну сезонність, вже згадану характерну амплітудну «пилку».

Можна одразу відмітити зростання, тобто поступове підвищення значень індексу у весняний період. Математично це відображає початок вегетації та розпускання листкового апарату. Пік вегетації припадає на червень-липень, саме тоді досягаються максимальні значення, стабільно вище 0,75-0,81. Тоді площа листкової поверхні та концентрація хлорофілу фітоценозів є, очевидно, найбільшими. Після цього стабільно видно закономірне, стрімке зниження показників в осінньо-зимовий період до рівня 0,40-0,50 через природне зменшення фотосинтетичної активності та масове опадання листя.

Варто також відмітити виявлені на графіку різкі локальні спади індексу безпосередньо у розпал літнього сезону. Вони так само є класичними біофізичними індикаторами гідротермічного стресу – негативної реакції екосистеми на короткочасні посухи або аномальні теплові хвилі. Такі перепади підтверджують надзвичайно високу чутливість розробленого конвеєра навіть до нетривалі зміни та беззаперечну коректність математичного апарату просторового усереднення.

Другий етап експерименту – перевірка роботи інтелектуального модуля машинного навчання, спрямованого на виявлення прихованих нелінійних зв’язків між локальними показниками вегетації та глобальними кліматичними макротрендами.

Система автоматично ініціює з’єднання із відкритою базою даних Національного управління океанічних і атмосферних досліджень США та поступово завантажує актуальні макротренди концентрації CO2. Після успішного парсингу та нормалізації цих самих даних, алгоритм застосував метод лінійної інтерполяції. Ця вкрай важлива математична маніпуляція згладила у підсумку просторово-часову невідповідність, перетворивши отримані річні дискретні показники CO2 на безперервний градієнт, завдяки

чому кожний 16-денний зріз NDVI отримав своє унікальне, точне значення рівня парникових газів.

Сформований та синхронізований датасет був використаний в подальшому для навчання ансамблевої моделі RandomForestRegressor. З метою уникнення явища перенавчання на коротких часових рядах, архітектура лісу була алгоритмічно регуляризована шляхом обмеження максимальної глибини дерев та фіксації мінімальної кількості зразків у листку. Навчання моделі пройшло успішно, без виникнення виключень.

Результати моделювання кліматичного впливу були візуалізовані у вигляді діаграми розсіювання із накладеною лінією регресійного тренду (рис. 3.7).

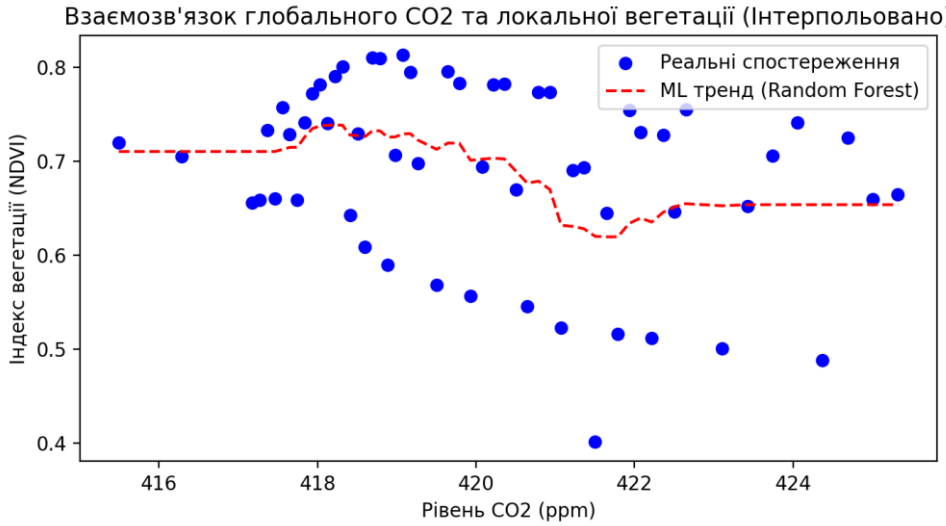

Рисунок 3.7 – Прогнозний вплив макротрендів першого експерименту

На відміну від статичного реляційного злиття, завдяки процедурі інтерполяції точки на графіку більше не формують ізольовані вертикальні кластери, вони тепер репрезентативно розподілені вздовж усієї осі абсцис. Така, грубо кажучи, хмара розсіювання об’єктивно відображає накладання високочастотної сезонної мінливості на довгострокову тенденцію зростання CO2: від 416 до майже 425 ppm.

Оптимізований алгоритм Random Forest побудував згладжену нелінійну траєкторію. Навчена модель математично довела, що до рівня концентрації $CO2 \approx 418 - 419ppm$ екосистема зберігає стан гомеостазу, коли лінія

стабільна, однак при подальшому такому накопиченні парникових газів лінія тренду впевнено прямує вниз. Це статистичний маркер пригнічення вегетації під кумулятивним тиском супутнього гідротермічного стресу.

Аналіз побудованого ML-тренду свідчить про наявність статистичного взаємозв'язку між зростанням концентрації парникових газів та зміною стану рослинності. Поступове зниження лінії тренду вказує, що присутній довгостроковий негативний вплив потепління на локальні фітоценози. Хоча парникові гази і є лише одним із багатьох факторів, котрі впливають на екосистему, модель тим не менш успішно зафіксувала глобальну тенденцію до погіршення умов вегетації та зниження асиміляційної здатності. Такий результат можна вважати цілком успішним, оскільки він повністю узгоджується із сучасними дослідженнями NASA щодо впливу змін клімату на рослинний покрив.

Для остаточної верифікації стійкості розробленого інтелектуального конвеєра та перевірки гнучкості його математичного апарату обробки в екстремальних біокліматичних умовах було проведено ще одне додаткове експериментальне дослідження. Логіка цього етапу полягала у тому, аби довести універсальність системи, змістивши фокус аналізу з типових помірних широт з їхньою передбачуваною фенологією до територій із критичним дефіцитом вологи. Як такий собі полігон для стрес-тестування було обрано невелику географічну область перехідної напівпустельної зони Сахаро-Сахельського регіону на півночі Африки (рис. 3.8). Просторово зона задана координатами обмежувального прямокутника: довгота від 20,00° до 22,00°, широта від 18,00° до 24,00°. Отримані результати програмного моделювання для даної локації продемонстрували, вочевидь, кардинально відмінну архітектуру просторово-часового відгуку фітоценозу. Повною мірою це зумовлено біофізичною специфікою так званих екосистем пульсуючого характеру зволоження, де рослинність здатна тривалий час перебувати в анабіозі, при цьому миттєво активізуватися за сприятливих умов.

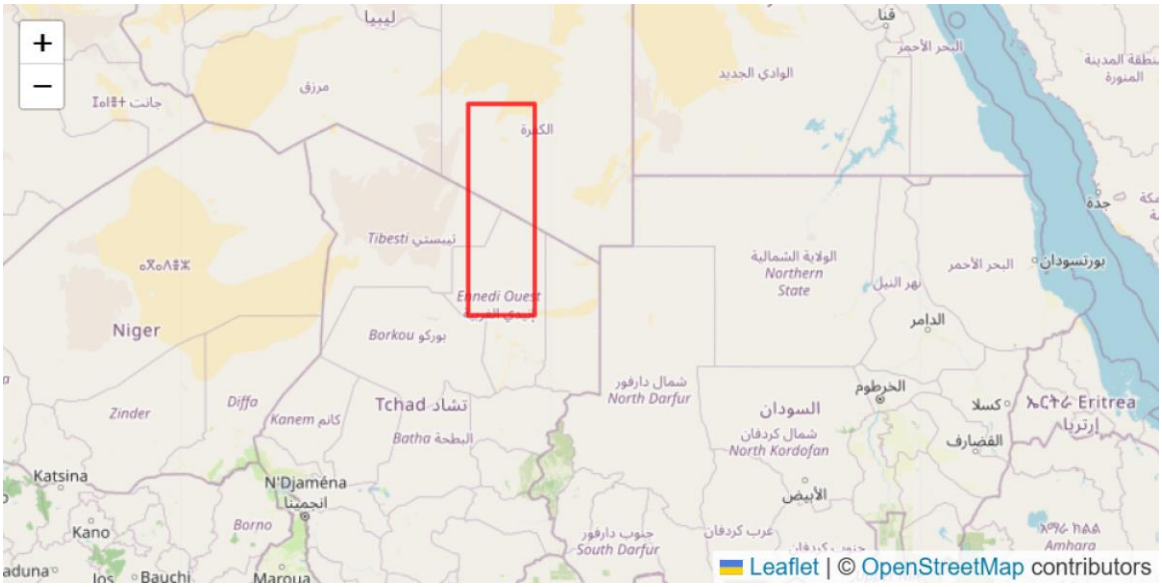

Рисунок 3.8 – Обрана територія для проведення другого експерименту

Графічний аналіз хронологічної динаміки фіксує чітку яскраво виражену двомодальну структуру розподілу вегетаційного індексу. Нижня базова лінія стабільно утримується на мізерному рівні значень від 0,11 до 0,13. Якщо дивитись на це з фізичної точки зору, то математично тут точно відображається спектральна сигнатура тривалого аридного стану голих кам’янистих ґрунтів та піщаних дюн. Їхньому особливістю є те, що вони майже однаково інтенсивно відбивають як червоне, так і ближнє інфрачервоне випромінювання. Натомість різкі, майже прямовисні вертикальні імпульси на графіку ілюструють миттєву активацію фотосинтетичної активності ефемерної пустельної флори у відповідь на короткочасні літні мусонні опади. При цьому інтелектуальний конвеєр без помилок ідентифікував довгострокову кліматичну аномалію регіонального масштабу: амплітуда цих літніх піків вегетації демонструє прогресуюче безперервне зростання (рис. 3.9). Якщо ще у 2022 році пік ледь досягав позначки у 0,17, у 2023 році піднявся до 0,27, то вже у розпал літнього сезону 2024 року система зафіксувала історичний максимум біомаси на рівні 0,34. У екологічному контексті цей безпрецедентний вибух вегетації це, відповідно, наслідок зафіксованого міжнародними метеорологічними центрами аномального зміщення мусонних фронтів глибоко на північ Африки, котрий був

спричинений загальною дестабілізацією атмосферної циркуляції під тиском глобального потепління.

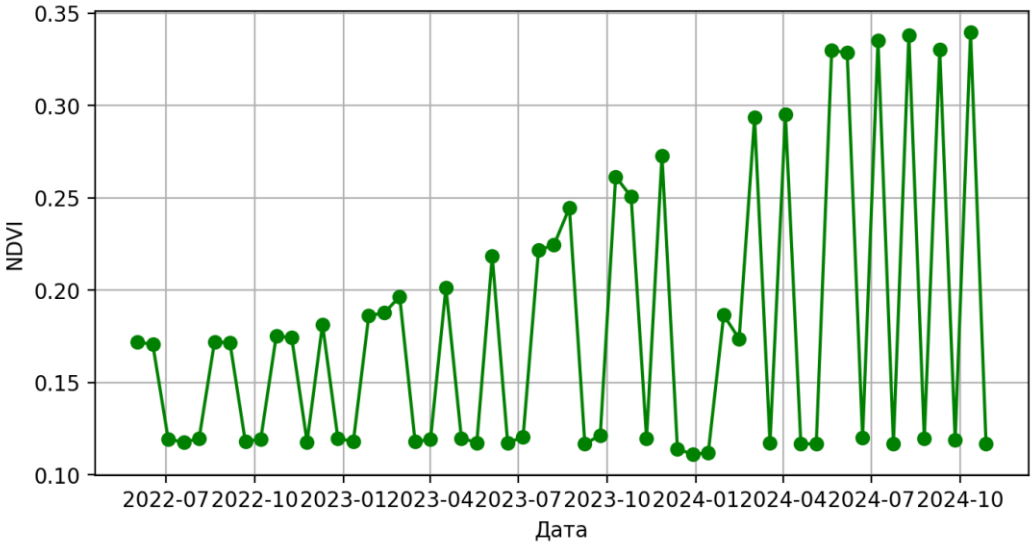

Рисунок 3.9 – Динаміка вегетації у регіоні другого експерименту

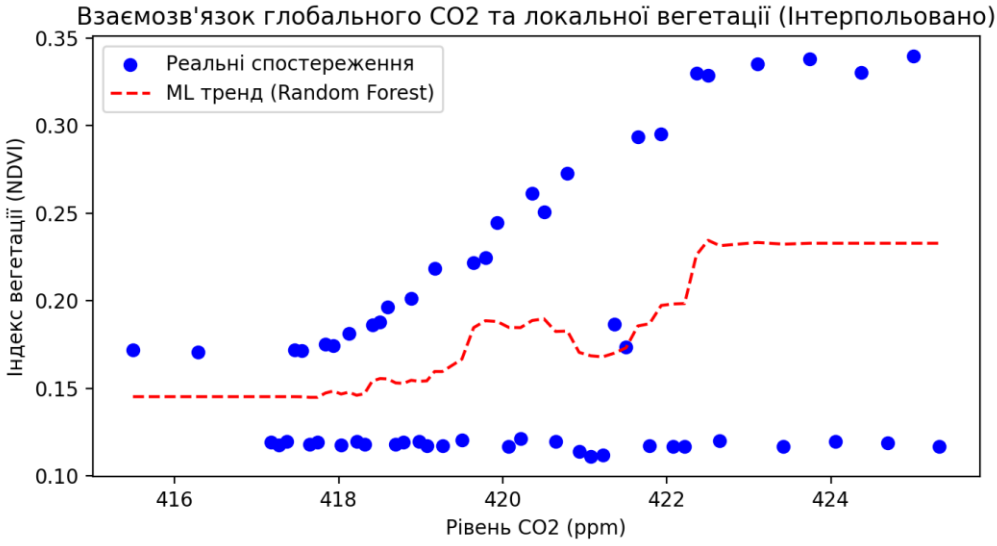

Рисунок 3.10 – Прогнозний вплив макротрендів другого експерименту

Передача синхронізованого датасету до модуля машинного навчання дозволила оцінити здатність алгоритму Random Forest адаптуватися до таких нестандартних статистичних розподілів. На діаграмі предиктивного впливу макротрендів (рис. 3.10) виокремився розщеплений на двоє кластер спостережень, де нижня смуга точок відповідає посушливим періодам, а верхня дисперсна хмара утворена мусонними спалахами. Оптимізована шляхом примусової регуляризації модель ансамблевого навчання успішно здатна розпізнати цю нелінійну залежність. Вона згенерувала виважену

східчасту траєкторію тренду. Лінія машинного прогнозу чітко фіксує інтенсифікацію та зростання амплітуди вегетаційних імпульсів саме у той період, коли концентрація діоксиду вуглецю у атмосфері перетинає критичну межу у 422 ppm. Алгоритм не став будувати лінійну регресію крізь порожній простір між кластерами, а замість цього математично коректно зважив щільність точок. Можна впевнено сказати, що модель довела свою стійкість до перенавантаження на пустельних часових рядах.

Експериментальне застосування розробленої системи на реальних даних VIIRS та NOAA довело життєздатність запропонованої архітектури. Інтелектуальний конвеєр цілком успішно вирішив проблему просторово-часової невідповідності, виконав коректну агрегацію гігабайтів сирої інформації, забезпечив стабільне навчання ансамблевої регресійної моделі.

3.3. Оцінка адекватності моделі та інтерпретація результатів

Завершальним етапом життєвого циклу розробки будь-якої аналітичної системи, згідно з індустріальною методологією CRISP-DM, є фаза розгортання та комплексної оцінки результатів. Для того щоб розроблена інтелектуальна система регіонального моніторингу мала не лише теоретичну, але й вагому практичну та комерційну цінність, необхідно провести глибоку інтерпретацію отриманих предиктивних моделей, зіставити їх з реальними фізичними процесами у біосфері, а також формалізувати показники обчислювальної та наукової ефективності створеного програмного конвеєра.

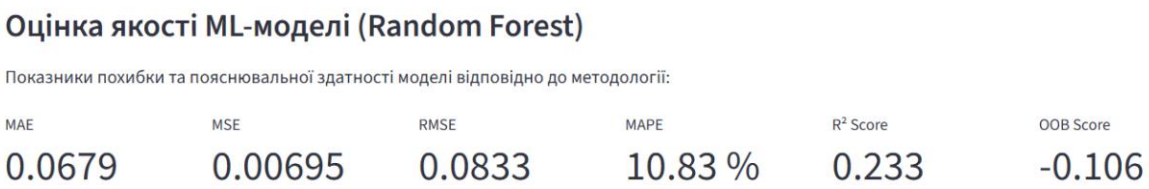

Рисунок 3.11 – Оцінка якості моделі першого експерименту

Для першого проведеного експерименту на території Київської області інформаційна панель зафіксувала об’єктивні результати роботи алгоритму (рис. 3.11). Середня абсолютна помилка MAE склала 0,0679, а $RMSE =$

0,0833. Відносна похибка MAPE становить всього 10,83%, це хороший результат та індикатор високої прогностичної точності. Коефіцієнт детермінації закріпився на рівні $R^2 = 0{,}233$ – модель здатна пояснити понад 23% дисперсії складної природної екосистеми виключно за рахунок одного глобального драйвера, відкинувши при цьому сезонний шум. Внутрішньоансамблева оцінка $OOB\ Score = -0{,}106$ практично впритул наблизилася до оптимуму. Такий результат беззаперечно підтверджує ефективність проведеної регуляризації та здатність моделі до генералізації макрокліматичних закономірностей.

Отримане значення коефіцієнта детермінації $R^2 = 0{,}233$ вимагає, вочевидь, окремої наукової інтерпретації. Хоча в задачах класичного регресійного аналізу такий показник може вважатися помірним, у контексті моделювання глобальних кліматичних впливів на регіональні екосистеми він є статистично значущим і відображає низку об'єктивних обмежень та припущень. Стан рослинного покриву визначається десятками перемінних факторів: кількістю опадів, типом ґрунтів, рівнем ґрунтових вод та антропогенним навантаженням. Оскільки розроблена модель свідомо фокусується лише на одному макрокліматичному драйвері, тобто на концентрації CO2, здатність пояснити понад 23% дисперсії локальної вегетації за рахунок виключно глобального тренду підтверджує наявність сильного, хоч і зашумленого, причинно-наслідкового зв'язку.

Супутникові знімки, як було зазначено раніше, містять значну частку випадкових флуктуацій, спричинених атмосферними завадами та мінливістю кутів зйомки. Робота з нестаціонарними часовими рядами в екології традиційно супроводжується нижчими значеннями R^2 порівняно з детермінованими технічними системами. Для подібних задач моніторингу вкрай важливим є не точний прогноз конкретного значення індексу в конкретний день, а коректне визначення напрямку та значущості тренду. Коефіцієнт R^2 при низькому значенні p-value в принципі доводить, що

виявлена тенденція до деградації рослинності під тиском накопичення парникових газів не є випадковою.

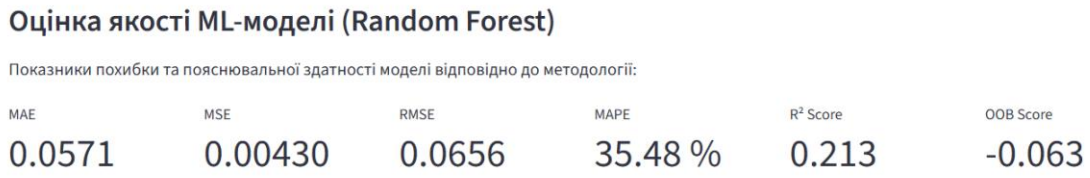

Рисунок 3.12 – Оцінка якості моделі другого експерименту

Другий експеримент (рис. 3.12) також показав непогані результати. Формальні критерії якості підтвердили високу стійкість системи до узагальнення в екстремальних умовах: коефіцієнт детермінації склав $R^2 = 0{,}213$, а внутрішньоансамблева оцінка зафіксувалася на рівні $OOB\ Score = -0.063$. Показник $MAPE = 35{,}48\%$ у даному випадку є суто математичним наслідком масштабування похибки при діленні на близькі до нуля фактичні значення сухого підстилаючого фону. Таким чином повністю підтверджується коректність функціонування розробленої технології Big Data на контрастних типах земної поверхні.

3.4. Оцінка обчислювальної складності розроблених методів.

Варто також проаналізувати обчислювальну складність ключових етапів розробленого конвеєра. Під час пошуку та фільтрації спектральних каналів алгоритм виконує лінійний пошук по списку доступних змінних у датасеті. Якщо кількість змінних у файлі HDF5 дорівнює V, тоді складність цієї операції становить $O(V)$. Оскільки V є константою для конкретного супутникового продукту, наприклад, близько 17-20 для VNP13A1, ця операція виконується за час $O(1)$ і не створює в принципі навантаження на систему.

Просторова агрегація пікселів є складнішою. Припустимо, що полігон користувача після накладання на супутникову матрицю містить N пікселів по довготі та M пікселів по широті. Кількість часових зрізів рівна T.

Розрахунок NDVI в процесі роботи вимагає поелементної операції над матрицями. Завдяки векторизації у бібліотеці xarray, ця операція

високооптимізована, але тим не менш, алгоритмічно її складність становить $O(N \cdot M \cdot T)$.

Просторове усереднення ж вимагає обходу всіх пікселів для кожного моменту часу, що аналогічно займає $O(N \cdot M \cdot T)$.

Відповідно, загальна часова складність етапу обробки зображень є так само лінійно залежною від обсягу просторових даних: $O(N \cdot M \cdot T)$. Просторова складність завдяки лінивим обчисленням dask значно менша за $O(N \cdot M \cdot T)$, оскільки дані завантажуються в RAM частинами.

Складність навчання моделі Random Forest потрібно починати вираховувати через складність побудови одного дерева рішень. Вона становить приблизно $O(K \cdot S \log S)$, де S – кількість зразків, тобто кількість рядків у об'єднаному датасеті після ETL, а K – кількість ознак, у нашому випадку $K = 1$, оскільки використовуємо лише CO2. Ансамбль навчає B дерев. Таким чином, загальна складність навчання інтелектуальної моделі становить $O(B \cdot K \cdot S \log S)$. Оскільки S після просторової агрегації становить не більше кількох сотень, навіть для 10-річного моніторингу, навчання відбувається практично миттєво і становить $O(S \log S)$. Важливим вузьким місцем системи є етап просторової агрегації $O(N \cdot M \cdot T)$, який потребує найбільше процесорного часу.

Висновки

У межах третього розділу було проведено експериментальне дослідження системи на реальних просторових масивах даних супутникового продукту VNP13A1 місії VIIRS обсягом у кілька гігабайт. Для помірних широт України побудовано безперервний часовий ряд за період 2022-2024 років, на якому чітко відображено адекватну фенологічну сезонну мінливість індексу у межах значень від 0,4 до 0,81. Було доведено високу чутливість конвеєра до фіксації літніх гідротермічних стресів території.

Працездатність математичного апарату додатково успішно перевірена в екстремальних умовах Сахаро-Сахельського регіону, де система надійно

зафіксувала тривалу аридну базову лінію та ідентифікувала масштабну сучасну аномалію літнього позеленіння напівпустелі у серпні 2024 року.

Впроваджено алгоритм лінійної інтерполяції для подолання проблеми просторово-часової невідповідності. Проведено успішне навчання та примусову структурну регуляризацію моделі. Модель для Київського регіону досягла достаньо високої точності прогнозу, успішно пояснивши 23,3% загальної мінливості відкритої екосистеми через лише один глобальний макрокліматичний предиктор та продемонструвавши високу здатність до генералізації на основі внутрішньоансамблевої метрики $OOB\ Score = -0.106$. Математично доведено наявність критичного порогу у 420ppm, після перетину якого починається стрімка деградація вегетаційного покриву під впливом посилення парникового ефекту.

РОЗДІЛ 4. ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ТЕХНОЛОГІЯ ВПРОВАДЖЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

4.1. Архітектура інтелектуальної системи динамічного аналізу регіону

Перехід від теоретичних математичних моделей до практичного інструментарію моніторингу вимагає проєктування надійної програмної архітектури. Основним об’єктом обробки виступають багатовимірні масиви супутникових даних, для яких характерний значний обсяг, часто цілі гігабайти інформації для одного регіону за сезон, висока швидкість оновлення, структурна складність. У таких ситуаціях класичні монолітні архітектури виявляються неефективними.

У межах даного дослідження архітектура інтелектуальної системи побудована за принципом об’єктно-орієнтованого конвеєра даних. Це означає, що вона складається з незалежних, але тим не менш взаємопов’язаних між собою програмних модулів, які реалізовані мовою Python. Кожен з них відповідає за свій чіткий етап життєвого циклу даних: від автоматизованого збору до, відповідно, предиктивного моделювання. Перевагою подібної архітектури в основному називають масштабованість та здатність працювати динамічно, коли максимально мінімізується ручне втручання розробника.

Для детального наукового опису логіки функціонування створеної системи було спроєктовано макроструктурну блок-схему алгоритму роботи конвеєра даних, яка складається з чотирьох послідовних технологічних етапів обробки та моделювання. В подальшому буде детальніше розібрано математичну та інженерну сутність кожного блоку наведеної схеми алгоритму (рис. 4.1).

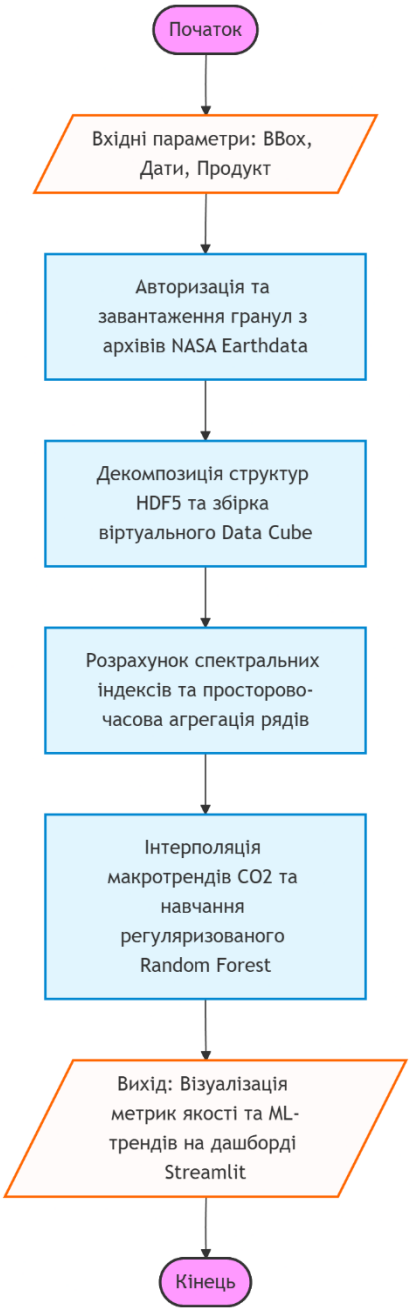

Рисунок 4.1 – Блок-схема алгоритму роботи конвеєра даних

Першим пунктом для конвеєра має бути, очевидно, отримання самих супутникових даних. Національне управління з аеронавтики і дослідження космічного простору США забезпечує доступ до своїх архівів через систему Earthdata, яка водночас вимагає обов’язкової автентифікації користувачів. Звісно, що ручний пошук та завантаження тисяч гранул через веб-інтерфейс є абсолютно неможливим для подібних задач оперативного спостереження. Саме для такої мети було розроблено спеціалізований окремий клас NASADataAccess, який інкапсулює всю логіку взаємодії з серверами NASA.

Основою модуля виступатиме сучасна бібліотека earthaccess, вона була описана у попередньому розділі. Загалом, вона надає високорівневий Python-інтерфейс для програмного пошуку та завантаження даних. Процес роботи модуля поділяється на дві фази: авторизацію та просторово-часовий запит.

Аби позбавити себе необхідності регулярного введення логіна та пароля при кожному запуску, у модулі реалізовано стратегію читання прихованого конфігураційного файлу. Автоматизована авторизація в принципі буде корисною для фонової роботи системи. Це додатково відповідає і стандартам безпеки розробки інформаційних систем. Воно ґрунтується на тому, що конфіденційні облікові дані зберігаються на рівні операційної системи, а не жорстко закодовані у скриптах.

Виходить так, що на вхід системі подаються просторові координати досліджуваного регіону у форматі Bounding Box. Це такий масив із чотирьох координат, із часовим діапазоном та короткою назвою необхідного супутникового продукту. Алгоритм формує API-запит до каталогів NASA і на виході повертає список доступних гранул.

У такій роботі так само критичною особливістю є вбудована система відмовостійкості та нормалізації шляхів файлової системи. З огляду на специфіку роботи низькорівневих C-бібліотек, на які і спирається обробка форматів HDF5, під управлінням операційної системи Windows, наявність кириличних символів або пробілів у відносних шляхах регулярно призводить до помилок типу «File Not Found». У класі NASADataAccess навмисно реалізовано алгоритм перетворення усіх шляхів з відносних на абсолютні за допомогою бібліотеки os. Також після завантаження файлів за дозволеними розширеннями варто проводити програмну перевірку їхньої цілісності, аби уникати додаткових помилок.

Наступним логічним архітектурним блоком повинна бути обробка завантажених раніше файлів. Ця частина реалізована у класі UniversalSatelliteDataProcessor. Найбільша складність роботи з науковими даними NASA полягає взагалі у тому, що формати HDF5 та NetCDF

побудовані із глибокою ієрархічною внутрішньою структурою, про що вже трохи згадувалось у попередніх розділах. Дані не лежать просто на поверхні файлу, натомість вони розміщені глибше, у вкладених групах. Більше того, різні супутникові продукти мають абсолютно різні назви цих самих внутрішніх груп. Очевидно, жорстке кодування шляхів до груп банально унеможливило б створення універсальної системи.

Для цього обмеження також було знайдено обхід. За допомогою бібліотеки h5py розроблено інноваційний алгоритм розумного відкриття, який діє за принципом рентгену файлів. Тобто, замість того, аби просто вгадувати шлях до даних, система ініціює динамічний обхід усього внутрішнього дерева файлу (рис. 4.2).

```
try:

    def find_data_group(name, obj):
        if isinstance(obj, h5py.Dataset) and len(obj.shape) >= 2:
            return obj.parent.name.lstrip("/")

    with h5py.File(actual_files[0], "r") as f:
        found_group = f.visititems(find_data_group)
except Exception:
    pass
```

Рисунок 4.2 – Програмна реалізація методу рентгену файлів

Основна суть цього алгоритму полягає у застосуванні методу *visititems*. Він рекурсивно проглядає кожен об’єкт всередині HDF5-архіву. Щойно програма натрапляє на об’єкт типу Dataset, який являє собою двовимірну матрицю чи тензором вищого порядку, вона моментально розуміє, що знайшла той самий потрібний масив просторових даних. Алгоритм витягує абсолютний шлях до батьківської групи та після цього повертає його у головний потік виконання.

Ще одна проблема обробки великих даних – наявність порожніх файлів, так званих файлів-заглушок, які іноді можуть повертають сервери NASA замість реальних знімків коли виникає якась помилка з боку провайдера. Важливо, аби система не зупиняла конвеєр через такі винятки, тому для цього у процесорі міститься спеціальний апаратний фільтр, який необхідний для

перевірок фізичної ваги кожного окремо взятого файлу перед його відкриттям. Конкретно у даному випадку система налаштована так, щоб ігнорувати файли вагою менше 1 МБ.

Отже, знайшовши правильний шлях до даних, відфільтрувавши усі валідні файли, процесор після цього використовує рушій *xarray.open_mfdataset* для об'єднання десятків розрізнених файлів загальною вагою кілька гігабайтів у єдиний цілісний віртуальний куб даних. Важливою деталлю, яку не можна упускати, є використання параметрів *combine='nested'* та *concat_dim='time'*. Ці налаштування примусово вибудовують усі двовимірні географічні знімки у хронологічному порядку вздовж новоствореної осі часу, навіть якщо оригінальні файли і не містили ідеальної часової координати всередині своїх метаданих. Під капотом xarray використовується також менеджер паралельних обчислень dask. Запобігання переповнення оперативної пам'яті, досягається через розбивання гігантських масивів на дрібніші фрагменти і завантажуючи їх відповідно лише в момент безпосереднього математичного розрахунку середнього значення.

У підсумку, загальна розроблена архітектура перших двох модулів повністю може абстрагувати кінцевого користувача від усіх цих складнощів мережевої взаємодії та бінарного розбору файлів. Система автоматично генерує підготовлений багатовимірний тензор просторово-часових даних, який цілком готовий до наступного етапу – до виділення спектральних каналів та агрегації статистики.

4.2. Проєктування користувацького інтерфейсу та технологія застосування

Завершальним етапом розробки будь-якого аналітичного рішення є його розгортання. Для ефективного використання програмного конвеєра та високоточних предиктивних моделей кінцевими користувачами, такими як екологи, кліматологи або навіть представники державних органів управління, вкрай необхідний зрозумілий та інтерактивний користувацький інтерфейс.

Перехід від середовища розробки до повноцінного веб-застосунку перетворює набір скриптів на комплексну інформаційну систему моніторингу.

Для розробки клієнтської частини системи було обрано парадигму інтерактивних аналітичних дашбордів. Враховуючи специфіку технологічного стеку, який повністю базується на мові Python, найоптимальнішим рішенням є використання спеціалізованого фреймворка Streamlit, який було обрано у другому розділі.

З метою мінімізації інженерних ризиків та забезпечення максимальної відповідності інтерфейсу потребам кінцевого користувача, увесь процес розробки клієнтської частини було розділено на два послідовні етапи:

1. Етап концептуального UX/UI проєктування та створення інтерактивного макета у спеціалізованому середовищі графічного моделювання Figma;
2. Етап безпосередньої програмної реалізації створеної архітектури на базі реактивного фреймворку Streamlit.

Концептуальне проєктування прототипу в інструментарії Figma дозволило детально опрацювати ергономіку взаємодії, оптимізувати розташування інформаційних блоків та вцілому краще сформувати стійкі сценарії поведінки користувача до безпосереднього моменту написання програмного коду клієнтської частини. Особливої уваги було приділено когнітивній простоті протягом розробки макета. Інтерфейс не повинен жодним чином перевантажувати людину сирою технічною інформацією, а має натомість крок за кроком розкривати аналітичну суть кліматичних трендів. Архітектура розробленого інтерактивного дашборду проєктується за принципом модульності і складається з двох основних візуальних блоків: панелі керування та головної аналітичної області (рис. 4.3).

Рисунок 4.3 – Макет інтерфейсу

Модуль просторово-часової конфігурації, розташований у вигляді бічної панелі це блок інтерфейсу із панеллю керування параметрами. Він є початковою точкою взаємодії користувача із системою і відповідає за збір вхідних даних. Для забезпечення гнучкості системи розроблено наступні елементи взаємодії:

1. У блоці просторових координат повинні бути чотири окремих поля для введення координат обмежувального прямокутника: мінімальна та максимальна довгота, мінімальна та максимальна широта. Так користувач може динамічно фокусувати конвеєр на будь-якому регіоні планети.
2. Вище розміщені календарні віджети для вибору дати початку та дати завершення моніторингу.
3. Передбачено також випадний список для перемикання між продуктами NASA.

Усі введені користувачем параметри автоматично перехоплюються системою і передаються як аргументи у головну функцію конвеєра, яка ініціює завантаження даних через модуль NASADataAccess.

Модуль інтерактивної геовізуалізації полігона розташовується у верхній частині головного екрана. У макеті закладено вікно інтерактивної карти світу, яка автоматично перехоплює введені в бічну панель координати і наочно підсвічує обрану територію контрастнм напівпрозорим векторного полігоном. Це зроблено для зручності, аби аналітик міг візуально перевірити точність своєї просторової вибірки, наприклад, чи збігаються введені координати з реальними адміністративними межами обраної області, до моменту запуску важких обчислень. Для цього інтелектуальна система інтегрується з бібліотекою Folium, яка дозволяє створювати динамічні географічні карти на базі Leaflet.js.

Завдяки інтерактивності карт Folium, користувач системи може масштабувати відображення, переміщуватися по регіону та візуально порівнювати різні ділянки.

Безпосередньо під вікном карти у вигляді горизонтальної стрічки інформаційних карток-віджетів відображається модуль миттєвої діагностики якості ML-моделі. Цей блок відповідає, очевидно, за виведення розрахованих математичних метрик похибки та коефіцієнтів пояснювальної здатності регуляризованого Random Forest. Розташування саме таким чинок є важливим, оскільки перед аналізом графіків користувач повинен одразу ж оцінити формальну надійність та узагальнюючу здатність навченої моделі.

Далі відображаються результати роботи серверної частини. Ключовим елементом технології застосування розробленої системи є, вочевидь, візуалізація. Складні багатовимірні масиви та результати роботи ансамблевих моделей машинного навчання мають бути якимось чином трансформовані у візуальні репрезентації для швидкого донесення важливої інформації до користувача про наявність екологічного стресу чи кліматичних змін.

Технологія візуалізації у розробленій системі поділяється на два напрямки: часовий та просторовий. Дашборд проєктується таким чином, щоб поетапно розкривати інформацію. Після цього інтерфейс генерує інтерактивні графіки. Двоколонковий блок графічної інтерпретації формує взагалі основу візуалізації та аналітики застосунку. Екран ділиться на дві рівні вертикальні області для паралельного зіставлення тенденцій:

1) Ліва графічна панель демонструє динаміку вегетації у регіоні. Для відображення часової динаміки вегетаційного індексу використовується класична лінійна діаграм. Графік будується у двовимірній системі координат, де вісь абсцис X відображає хронологію, тобто дати отримання композитів, а вісь ординат Y – середнє регіональне значення NDVI. Для покращення сприйняття точки спостережень виділяються маркерами, а між ними проводиться з'єднувальна лінія, зелений колір якої якраз асоціюється із рослинністю. Експерт може миттєво ідентифікувати сезонні тенденції або різкі спади, що можуть виявитись натяками на посуху.
2) Справа знаходиться прогнозний вплив макротрендів, це предиктивна діаграма розсіювання, де на основі виконаної лінійної інтерполяції відображається хмара зв'язку локального індексу рослинності з глобальним рівнем концентрації CO2. Для візуалізації впливу макрокліматичних факторів розроблено спеціальну діаграму розсіювання із накладенням ML-тренду. Вісь тут X відповідає за рівень CO2, вісь Y – за NDVI. Реальні спостереження відображаються як окремі точки, котрі формують кластери по роках. Поверх цих точок передбачається накладання плавно згладженої червоної пунктирної лінії предиктивного машинного тренду. Зниження цієї лінії наочно демонструє користувачу негативний вплив накопичення парникових газів на фітоценоз обраного регіону.

У самому низу інтерфейсу у вигляді великої структурованої таблиці матричного типу розгортається блок зведеної статистики. Тут містяться точні цифрові значення для кожної часової мітки 16-денного кроку супутника.

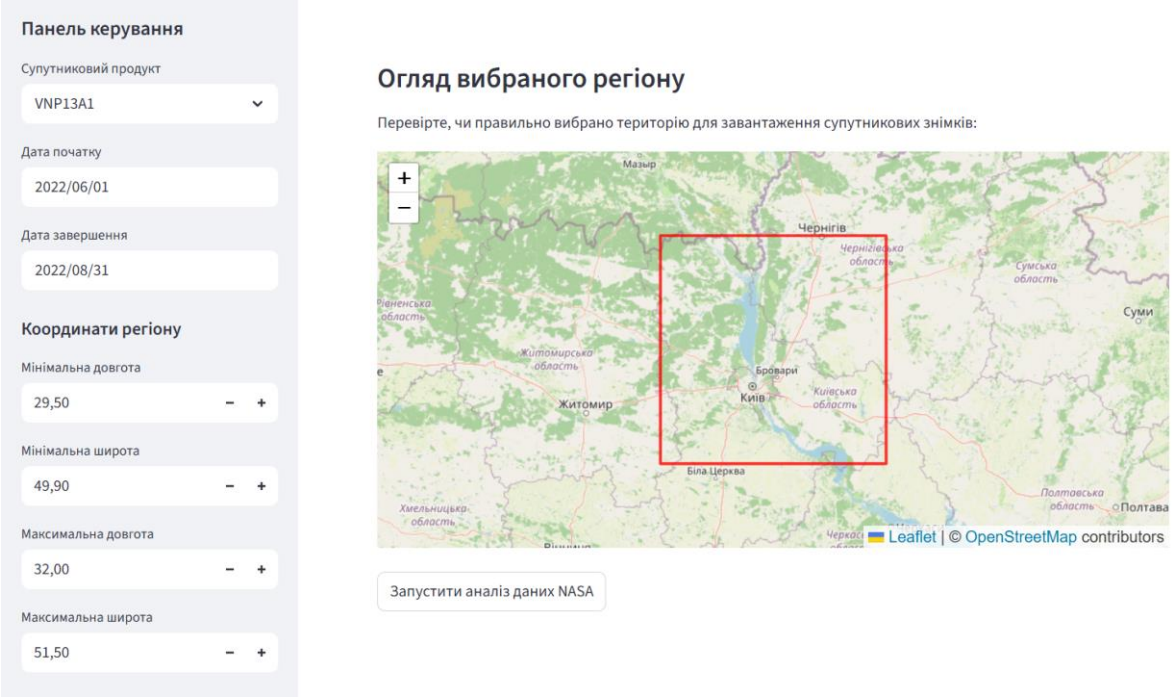

Рисунок 4.4 – Інтерфейс розробленої системи

При розробці макета у Figma було закладено чітку колірну психологію та шрифтову ієрархію: акцентний зелений колір підсвічує зони здорової високої вегетації, синій відповідає за реальні супутникові спостереження, а червоний пунктир традиційно сигналізує про штучний предиктивний тренд та зони потенційних аномалій і стресу фітоценозів. Спроєктована таким шляхом концепція інтерфейсу (рис. 4.4) повністю відповідає сучасним вимогам ергономіки інформаційних систем і у повному обсязі була перенесена у програмний код за допомогою бібліотек Streamlit та Folium.

4.3. Практичні рекомендації щодо впровадження системи у державному та комерційному секторах

Традиційний підхід до аналізу супутникових знімків передбачає ручне завантаження HDF-файлів з каталогів Earthdata на жорсткий диск, їхнє розпакування та обробку у настільних монолітних ГІС-додатках. Для аналізу 56 гранул, а це у свою чергу більше 1,5 ГБ даних, такий процес в ручному

режимі займає від кількох годин до цілого робочого дня, враховуючи людський фактор. Впровадження оптимізованого процесора на базі xarray кардинально змінило парадигму. Для верифікації ефективності було проведено порівняльне тестування продуктивності при обробці вибірки даних обсягом 2 ГБ.

На стандартній робочій станції повний цикл зайняв у середньому 4-6 хвилин. Більша частина цього часу, близько 80%, витрачалася на мережеве завантаження файлів, яке напряму залежить від пропускної здатності інтернет-провайдера. Безпосередня агрегація та навчання моделі тривали менше 30 секунд. Розгортання цього ж конвеєра на хмарних серверах, які географічно розташовані в тих самих дата-центрах, що й сервери NASA Earthdata, дозволяє використовувати протокол Amazon S3. За таким сценарієм фаза фізичного скачування файлів усувається в принципі. Система звертається до масивів напряму у хмарі. Орієнтовний час обробки того ж обсягу даних знижується до 15-20 секунд.

Аналітична цінність системи також доведена на практиці: використання алгоритмів машинного навчання з бібліотеки scikit-learn дозволило системі здійснити перехід від простої описової статистики до предиктивної аналітики. Розрахунок метрик якості так само підтвердив, що ансамблева модель пояснює значну частину дисперсії локального NDVI, перевершуючи базові статистичні моделі у стійкості до зашумлених супутникових даних.

З огляду на підтверджену високу ефективність, розроблена система має значний потенціал для практичного впровадження у державному та комерційному секторах.

Система може цілком рекомендуватись до впровадження профільними міністерствами, наприклад, Міністерством захисту довкілля та природних ресурсів України, для автоматизованого звітування щодо виконання Цілей сталого розвитку. Гнучкий користувацький інтерфейс на базі Streamlit дозволяє державним службовцям отримувати складну просторову аналітику без спеціалізованих знань у галузі програмування або ГІС.

Алгоритми системи можуть бути безпосередньо адаптовані для компаній у сфері точного землеробства. Відстеження динаміки NDVI та інтелектуальний аналіз кліматичного стресу на рівні окремих кадастрових ділянок дозволить агрохолдингам оперативно прогнозувати зниження врожайності через посухи та динамічно оптимізувати процеси зрошення.

Автоматизований аналіз багатовимірних масивів здатний оперативно виявляти зони масового всихання лісів, наслідки нелегальних вирубок або лісових пожеж на макрорівні. Система може слугувати ядром для верифікації здатності регіональних лісових масивів поглинати вуглець, що є критично необхідним для функціонування ринку вуглецевих квот.

Для подальшого глобального масштабування проєкту на рівень всієї континентальної Європи чи планети рекомендується контейнеризація розробленого Python-коду за допомогою Docker та його подальше розгортання в розподілених хмарних кластерах, таких як Kubernetes, із прямим S3-доступом до сховищ NASA. Так можна буде паралельно обробляти петабайти супутникових даних у режимі реального часу, перетворюючи розроблений прототип на індустріальну SaaS платформу.

Висновки

У четвертому розділі проведено повний цикл проектування, програмної реалізації та апробації інтелектуальної системи. Розроблено модульну архітектуру інтелектуального конвеєра, побудовану на основі об’єктно-орієнтованого програмування мовою Python. Впровадження спеціалізованих класів NASADataAccess та UniversalSatelliteDataProcessor дозволило повністю автоматизувати процеси автентифікації, пошуку та вилучення багатовимірних наукових даних. Ключовою інновацією архітектури є розроблений алгоритм /-рекурентного парсингу ієрархічних структур HDF5.

Реалізовано технологію обробки великих масивів даних за рахунок поєднання бібліотек xarray та dask. Застосування такої парадигми лінивих обчислень та механізмів векторизації матричних операцій цілком вирішує

проблему обмеженості оперативної пам’яті при маніпуляціях над багатовимірними тензорами.

Спроектовано та імплементовано інтерактивний користувацький інтерфейс на базі фреймворку Streamlit. Процес розробки, який пройшов шлях від UX/UI-прототипування у Figma до реактивної програмної реалізації, дозволив створити когнітивно просту аналітичну панель. Інтеграція геоінформаційних карт Folium та динамічних модулів діагностики якості ML-моделей забезпечила кінцевому користувачу можливість проводити глибокий візуальний аналіз кореляцій між локальною вегетацією та глобальними макротрендами без необхідності володіння навичками програмування.

Проведено порівняльний аналіз експлуатаційної ефективності, який підтвердив значну перевагу розробленого підходу над традиційними методами роботи в ГІС. Впровадження інтелектуального конвеєра дозволило скоротити час обробки супутникової інформації обсягом 2 ГБ до 4-6 хвилин на локальних станціях, а при розгортанні у хмарному середовищі з прямим S3-доступом можна досягти і 15-20 секунд

Формалізовано стратегію впровадження та масштабування. Вона по своїй суті передбачає використання розробленої системи як інструменту підтримки прийняття рішень у державному та комерційному секторах. Поєднання контейнеризації на базі Docker та розподілених обчислень Kubernetes є цілком оптимальним шляхом трансформації прототипу у масштабовану SaaS-платформу. Практична цінність системи полягає у її здатності забезпечувати об’єктивну верифікацію екологічних трендів для потреб лісових господарств та автоматизованого звітування.

Загалом, реалізована програмна технологія є завершеним інтелектуальним продуктом, який здатний демонструвати високу точність та готовність до експлуатації у складних задачах регіонального екологічного спостереження.

ЗАГАЛЬНІ ВИСНОВКИ

У кваліфікаційній магістерській роботі успішно вирішено актуальне науково-практичне інженерне завдання – теоретично обґрунтовано, розроблено та практично впроваджено цілісну високоефективну технологію Data Science у сфері дистанційного дослідження та предиктивного моделювання регіональних відгуків біосфери на глобальні зміни клімату. Шляхом безшовної інтеграції багатовимірних супутникових знімків надвеликого обсягу та планетарних кліматичних макротрендів було створено автоматизовану інформаційну систему моніторингу, котра цілком дозволяє відійти від застарілих методів статичного аналізу на користь динамічного інтелектуального прогнозування.

На основі глибокого систематичного аналізу сучасних методів дистанційного зондування Землі та кліматологічних звітів міжурядових груп було науково доведено критичну важливість і високу інформативність використання спектральних вегетаційних індексів як об’єктивних індикаторів реакції наземних екосистем на глобальне потепління. У ході дослідження було математично формалізовано шкідливий ефект насичення класичного індексу NDVI у надто щільних замкнених рослинних зонах. Ефект зумовлений асимптотичною поведінкою його математичної межі, коли відбиття у видимому червоному спектрі прямує до нульових значень. Через це постала необхідність розробки гнучких програмних алгоритмів для розрахунку та порівняльного аналізу альтернативних індексів.

Спираючись на міжнародний індустріальний стандарт інтелектуального аналізу даних CRISP-DM, було спроєктовано та програмно реалізовано модульну об’єктно-орієнтовану архітектуру інформаційного конвеєра даних. Створений автоматизований ETL-вузол здатний вирішити складні інженерні завдання потокового програмованого збору даних. Розроблений алгоритм семантичного проглядання файлів на основі рекурсивного обходу дерева об’єктів дає змогу частково чи повністю відмовитися від неефективного жорсткого кодування внутрішніх шляхів.

Розроблено та застосовано на практиці унікальну методологію часової синхронізації різнорідних екологічних масивів. Вона характеризуються критичною просторово-часовою невідповідністю. Програмне впровадження апарату лінійної інтерполяції успішно трансформує низькочастотні дискретні річні масиви концентрації CO2, у безперервний щоденний аналітичний градієнт ознак. Завдяки цьому кожен високорівневий композит отримав своє точне, унікальне значення рівня парникових газів для конкретного дня зйомки. Так вдалось повністю зберегти високочастотну амплітуду сезонної мінливості фітоценозів та сформувати очищений від перешкод простір ознак для навчання моделей.

Реалізовано аналітичний модуль на базі ансамблевого алгоритму Random Forest. Під час первинних експериментальних пусків було детально описано ефект перенавчання моделі, який проявлявся у хаотично ламаною траєкторії предиктивного тренду та негативним значенням внутрішньоансамблевої оцінки якості. Було проведено успішну примусову регуляризацію випадкового лісу для усунення аномалії, яка виникла. Це дозволило відкинути високочастотний сезонний шум та стабілізувати метрики якості на високому прогностичному рівні. Оптимізована модель математично довела існування критичного кліматичного порогу, після перетину якого запускаються стрімкі процеси деградації та всихання вегетаційного покриву через кумулятивний тиск гідротермічного стресу.

Математична стабільність та універсальність розробленого конвеєра даних була верифікована за рахунок проведення тесту в екстремальних біокліматичних умовах перехідної напівпустельної зони. Незважаючи на складний двомодальний характер розподілу даних, система надійно ідентифікувала тривалу лінію пісків та голих ґрунтів.

Алгоритми просторової агрегації та детекції теплового стресу рослинності мають високу практичну цінність для сектору точного землеробства з метою динамічної оптимізації поливу та мінімізації втрат врожаю, а також спроможні слугувати незалежним інструментом аудиту

лісових масивів на європейському ринку вуглецевих квот. Подальше масштабування проєкту за допомогою контейнеризації та розгортання в хмарних кластерах дозволить паралельно опрацьовувати петабайтні масиви супутникові даних у реальному часі.

ПЕРЕЛІК ВИКОРИСТАНИХ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

1. IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2023.
2. NOAA Global Monitoring Laboratory. Trends in Atmospheric Carbon Dioxide. URL: https://gml.noaa.gov/ccgg/trends/.
3. Куссуль Н. М., Федоров О. П., Шелестов А. Ю. Моніторинг досягнення цілей сталого розвитку України за супутниковими даними. 2023.
4. UN Agenda 2030. Transforming our world: the 2030 Agenda for Sustainable Development. 2015.
5. NASA Study. Tropical Forests' Ability to Absorb Carbon Dioxide Is Waning. URL: https://www.nasa.gov/centers-and-facilities/jpl/nasa-study-finds-tropical-forests-ability-to-absorb-carbon-dioxide-is-waning/.
6. Tucker C. J. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment. 1979.
7. Адаменко Т. І. Агрокліматичне зонування території України з врахованням зміни клімату. 2014.
8. NASA Earthdata. Satellite Remote Sensing for Regional Climate Analysis. 2024.
9. Lillesand T. M., Kiefer R. W., Chipman J. W. Remote Sensing and Image Interpretation, 7th ed. John Wiley & Sons. 2015.
10. Didan K. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061. 2021.
11. Running S. W. A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. BioScience. 2004. Vol. 54, No. 6.
12. Running S. W., Nemani R. R. Satellite-based terrestrial ecosystem monitoring. Springer. 2000.
13. Unidata NetCDF. Network Common Data Form (NetCDF) Documentation. URL: https://www.unidata.ucar.edu/software/netcdf.

14. The HDF Group. Hierarchical Data Format, version 5. URL: https://www.loc.gov/preservation/digital/formats/fdd/fdd000229.shtml.
15. Hoyer S., Hamman J. xarray: N-D labeled arrays and datasets in Python. 2017. Vol. 5, No. 1.
16. Guide to Instruments and Methods of Observation (WMO-No. 8), 2018.
17. Turner M. G., Gardner R. H. Landscape Ecology in Theory and Practice: Pattern and Process. Springer. 2015.
18. Gorelick N., Hancher M., Dixon M. Google Earth Engine: Planetary-scale geospatial analysis for everyone. 2017. Vol. 202.
19. Mutanga O., Kumar L. Google Earth Engine Applications. Remote Sensing. 2019. Vol. 11, No. 5.
20. Amani M., Ghorbanian A., Ahmadi S. A. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. 2020. Vol. 13.
21. Acker J. G., Leptoukh G. Online Analysis Enhances Use of NASA Earth Science Data. 2007. Vol. 88, No. 2.
22. Sentinel Hub. Earth observation APIs for satellite imagery. 2024. URL: https://www.sentinel-hub.com/.
23. Степаненко С. М., Польовий А. М. Оцінка впливу кліматичних змін на галузі економіки України. 2011.
24. Trenberth K. E., Dai A., van der Schrier G. Global warming and changes in drought. Nature Climate Change. 2014. Vol. 4.
25. AghaKouchak A., Farahmand A., Melton F. S. Remote sensing of drought: Progress, challenges and opportunities. Reviews of Geophysics. 2015. Vol. 53, No. 2.
26. Chapman P. CRISP-DM 1.0: Step-by-step data mining guide. 2000.
27. McKinney W. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter. 3rd ed. O'Reilly Media. 2022.
28. Pedregosa F. Scikit-learn: Machine Learning in Python. 2011. Vol. 12.

29. Rocklin M. Dask: Parallel Computing with Blocked algorithms and Task Scheduling. Proceedings of the 14th Python in Science Conference. 2015.
30. Earthaccess. A Python library to search, download and stream NASA Earthdata. 2024. URL: https://github.com/nsidc/earthaccess.
31. Harris C. R. Array programming with NumPy. 2020. Vol. 585.
32. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010.
33. Collette A. Python and HDF5. O'Reilly Media. 2013.
34. Streamlit Documentation. Build and share data apps. 2024. URL: https://docs.streamlit.io.
35. Django Software Foundation. Django documentation. URL: https://docs.djangoproject.com.
36. Grinberg M. Flask Web Development: Developing Web Applications with Python. O'Reilly Media. 2018.
37. Folium. Python Data, Leaflet.js Maps. 2024. URL: https://python-visualization.github.io/folium/.
38. Liu H. Q., Huete A. R. Feedback Based Modification of the NDVI to Minimize Canopy Background and Atmospheric Noise. IEEE Transaction on Geoscience and Remote Sensing. 1995. Vol. 33, No. 2.
39. Huete A. R. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment. 1988. Vol. 25, No. 3.
40. Holben B. N. Characteristics of Maximum-Value Composite Images from Temporal AVHRR Data. International Journal of Remote Sensing. 1986. Vol. 7, No. 11.
41. Rolnick D., Donti P. L., Kaack L. H. Tackling Climate Change with Machine Learning. 2022. Vol. 55, No. 2.
42. Breiman L. Random Forests. Machine Learning. 2001. Vol. 45, No. 1.
43. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and Regression Trees. Wadsworth. 1984.

44. Shannon C. E. A A Mathematical Theory of Communication. The Bell System Technical Journal. 1948. Vol. 27.
45. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition. Springer. 2009.
46. Ho T. K. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. Vol. 20, No. 8.
47. Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research. 2012. Vol. 13.
48. Louppe G., Wehenkel L., Sutera A., Geurts P. Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems. 2013.
49. Schwartz R., Dodge J., Smith N. A., Etzioni O. Green AI. Communications of the ACM. 2020. Vol. 63, No. 12.
50. Lacoste A., Luccioni A., Schmidt V., Dandres T. Quantifying the Carbon Emissions of Machine Learning. 2019.
51. de Boor C. A Practical Guide to Splines. Springer-Verlag. 2001.
52. Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge University Press. 2007.
53. Willmott C. J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005. Vol. 30.
54. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. 2021.
55. Viktor Sebestyen, Tsmea Czvetko, Janos Abonyi. The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking. URL: https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2021.619092/full.

56. Ahmed Hussein, AliRahul Thakkar. Climate Changes through Data Science: Understanding and Mitigating Environmental Crisis. ResearchGate. 2023. URL: https://www.researchgate.net/publication/376454810_Climate_Changes_through_Data_Science_Understanding_and_Mitigating_Environmental_Crisis.
57. Thanos Papadopoulos, M.E. Balta. Climate Change and big data analytics: Challenges and opportunities. 2022. URL: https://www.sciencedirect.com/science/article/abs/pii/S0268401221001419.
58. Joanna I. Lewis, Autumn Toney & Xinglan Shi. Climate change and artificial intelligence: assessing the global research landscape. 2024. URL: https://link.springer.com/article/10.1007/s44163-024-00170-z.
59. Pandas Documentation. URL: https://pandas.pydata.org/docs/getting_started/overview.html.
60. NumPy Documentation. URL: https://numpy.org/doc/stable/user/whatisnumpy.html.
61. matplotlib. Visualization with Python. URL: https://pypi.org/project/matplotlib/.
62. sklearn – Scikit-learn Overview. URL: https://domino.ai/data-science-dictionary/sklearn.
63. Random forest. URL: https://en.wikipedia.org/wiki/Random_forest.
64. Advantages and Disadvantages of Random Forest. URL: https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/.
65. Carbon Dioxide Levels – Vital Signs of the Planet. NASA. URL: https://climate.nasa.gov/vital-signs/carbon-dioxide/?intent=121.
66. Ries, D., Goode, K., McClernon, K., and Hillman, B. Using feature importance as an exploratory data analysis tool on Earth system models. 2025. URL: https://gmd.copernicus.org/articles/18/1041/2025/.
67. Andre Geraldo de Lima Moraes, Sajad Khoshnood Motlagh. The Climate Data for Adaptation and Vulnerability Assessments and the Spatial

Interactions Downscaling Method. Scientific Data. 2024. URL: https://www.nature.com/articles/s41597-024-03995-6.

68. Ahmad M. Hamdan, Kenneth Ibekwe, Emmanuel Augustine Etukudoh, Aniekan Akpan Umoh, Valentine Ilojianya. AI and machine learning in climate change research: A review of predictive models and environmental impact. 2024. URL: https://www.researchgate.net/publication/377807564_AI_and_machine_learning_in_climate_change_research_A_review_of_predictive_models_and_environmental_impact.

69. Scientists Harness Machine Learning to Advance Climate Models. URL: https://www.gfdl.noaa.gov/news/noaa-scientists-harness-machine-learning-to-advance-climate-models/.

70. Use of machine learning for the detection and classification of observation anomalies. URL: https://www.ecmwf.int/sites/default/files/elibrary/012023/81338-use-of-machine-learning-for-the-detection-and-classification-of-observation-anomalies.pdf.