ШИФР: SWIFTT BTI

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ІНЖЕНЕРІЇ ОЗНАК ДЛЯ ЗАДАЧІ СЕМАНТИЧНОЇ СЕГМЕНТАЦІЇ СТАНУ ЛІСІВ ЗА СУПУТНИКОВИМИ ДАНИМИ

Зміст

Анотація

Актуальність теми цієї роботи полягає у розробленні методів штучного інтелекту для автоматизації побудови та пошуку інформативних ознак, що дозволяють якісно вирішувати задачі моніторингу навколишнього середовища на основі супутникових даних. Подібні завдання ставляться в проектах програми Horizon Europe, наприклад SWIFTT (https://swiftt.eu/). Такі методи дозволять в майбутньому оперативно виявляти та вирішувати проблеми, такі як екологічні зміни та загрози для природи, що є критичними для збереження біорізноманіття та здоров'я екосистем, зокрема, виявлення захворювання лісу.

Метою роботи є підвищення ефективності супутникового моніторингу стану навколишнього середовища на основі інтелектуального аналізу супутникових даних.

Завдання:

- Розробка методу інженерії ознак з використанням оптимізаційних алгоритмів штучного інтелекту.
- Побудова та навчання нейронних мереж для інтелектуальної класифікації стану лісу.
- Аналіз ефективності запропонованих інтелектуальних підходів.

Методика дослідження: у роботі використано методи науки про дані, системного аналізу, математичної статистики, методів оптимізації, машинного навчання, комп’ютерного зору, що є актуальними для досліджень в області штучного інтелекту, а також проведено комп’ютерний експеримент.

Загальна характеристика: Розроблено автоматизований підхід на основі генетичних алгоритмів оптимізації та нейронних мереж для інтелектуальної обробки супутникових даних. Запропоновано нову цільову функцію інформативності ознак та класи стійких до шумів вегетаційних індексів. Підхід дозволяє автоматизувати процес інженерії ознак, підвищити незалежність ознак та якість семантичної сегментації в задачах дистанційного зондування.

Проведено аналіз ефективності на реальних даних та продемонстровано можливість узагальнення моделей.

Ключові слова: нейронні мережі, комп'ютерний зір, штучний інтелект, вегетаційні індекси, інженерія ознак, генетичний алгоритм, багатошаровий перцептрон, дистанційне зондування, семантична сегментація, аналіз даних.

Вступ

У сучасному світі все частіше виникають завдання моніторингу навколишнього середовища на основі дистанційних спостережень, зокрема супутникових даних. Такі задачі по своїй суті є задачами семантичної сегментації на основі супутникових знімків і є дуже важливими у сфері штучного інтелекту та комп'ютерного зору.

На сьогоднішній день існує велика кількість робіт, присвячених використанню методів машинного навчання для вирішення цих задач. У них аналізуються сезонні зміни спектральних каналів зображення [1], використовуються методи глибинного навчання [2, 3, 4, 5], створюються нові вегетаційні індекси [6].

Проте в багатьох випадках при розв'язанні поставлених перед ними задач автори або обмежуються лише спектральними каналами, або додатково до них використовують кілька добре відомих вегетаційних індексів. Але водночас не аналізують чи спроможні обрані ними ознаки якісно вирішувати поставлену задачу (чи є вони інформативними) та наскільки добре вони поєднуються між собою. Немаловажливим є й те, що процес вибору набору ознак частіше за все виконується власноруч або ж з використанням знань експертів в предметній області.

Зважаючи на це, метою нашого дослідження було знаходження методів, що допомогли б автоматизувати процес вибору інформативних ознак. Для цього нам було необхідно створити метод чисельної оцінки інформативності як окремих ознак, так і їх сукупностей, а також дослідити та реалізувати можливість використання методів оптимізації для знаходження оптимальних наборів ознак (в межах поставленої задачі семантичної сегментації).

Для того щоб не обмежувати вибір ознак лише в межах відомих та зменшити кількість необхідної людської роботи, перед нами також стояла задача інженерії ознак. Отже, це дослідження грунтується не стільки на використанні

вегетаційних індексів, скільки на класах вегетаційних індексів, серед яких, за допомогою методів штучного інтелекту, а саме - генетичного алгоритму, і відбуватиметься формування набору інформативних ознак.

Постановка задачі

Поставлена перед нами задача полягає в побудові на основі супутникових знімків бінарної сегментації на певній ділянці в межах якогось класу підстилаючої поверхні (листяний/хвойний ліс, море, поле…).

Кожна ділянка A асоційована з умовною сіткою на поверхності Землі розмірності $w \times h$ пікселів, де w - ширина ділянки, h - висота ділянки.

Ділянка A складається з маски цільової сегментації (ground truth) M_S, маски класу підстилаючої поверхні M_F та набором X супутникових знімків. $A = (X, M_S, M_F)$.

M_S є булевою матрицею розмірності $w \times h$. Пікселі зі значенням 1 відповідають цільовому стану сегментації.

M_F є матрицею розмірності $w \times h$. Кожний елемент матриці може мати N_S значення: $M_F(i, j) \in \{0, \ldots, N_S - 1\}$. 0 - клас, в межах якого проводиться сегментація, $1, \ldots, N_S - 1$ - інші класи.

Кожний знімок X_i з набору X є тензором розміності $w \times h \times b$, де b - кількість спектральних каналів. При навчанні моделей, X_1 є знімком за той день, що відповідає даті актуальної маски цільової сегментації M_S.

Опис даних

У рамках проекту Horizon Europe SWIFTT (Satellites for Wilderness Inspection and Forest Threat Tracking) (https://swiftt.eu/), перед нами стоїть задача виявлення зараженого жуками короїдами [7] хвойного лісу.

Для виконання цього завдання в межах проекту SWIFTT було отримано дані у вигляді супутникових зображень Sentinel-2 та масок захворювання, які відповідають ділянкам лісових масивів в регіоні Гранд-Ест, Франція (рис. 1 (а)). Ці дані використовувалися для знаходження наборів ознак та навчання моделей, розроблених в даній роботі, та їх чисельного оцінювання.

(а) (б)

Рис. 1 - Розташування ділянок для яких надані дані. (а) - навчальні ділянки ; чорні точки - ділянки; (б) - тестова ділянка

Для перевірки можливості узагальнення отриманих моделей за допомогою них було проведено виявлення хворого лісу на зображеннях за 2018 рік на північ від Києва, поблизу Чорнобиля (рис. 1 (б)).

У межах поставленого нами завдання як вхідні дані використовуються багатоспектральні зображення Sentinel-2. Як відомо [8], такі зображення містять інформацію про поверхню Землі в різних спектральних діапазонах, що дозволяє використовувати їх для різноманітних цілей, таких як моніторинг земельного покриву, вимірювання рослинності та виявлення змін у навколишньому середовищі.

На рис. 2 показано 13 каналів Sentinel-2, кожен з яких має свою довжину хвилі. Канали поділяються на три групи: видимі канали (B1, B2, B3, B4), ближні інфрачервоні канали (B5, B6, B7, B8, B8A, B9), короткохвильові інфрачервоні канали (B10, B11, B12). Просторова роздільна здатність каналів Sentinel-2 становить 10 м, 20 м або 60 м, що залежить від конкретного діапазону. За браком каналів B9, B10 у наданих даних, у роботі вони не використовуються.

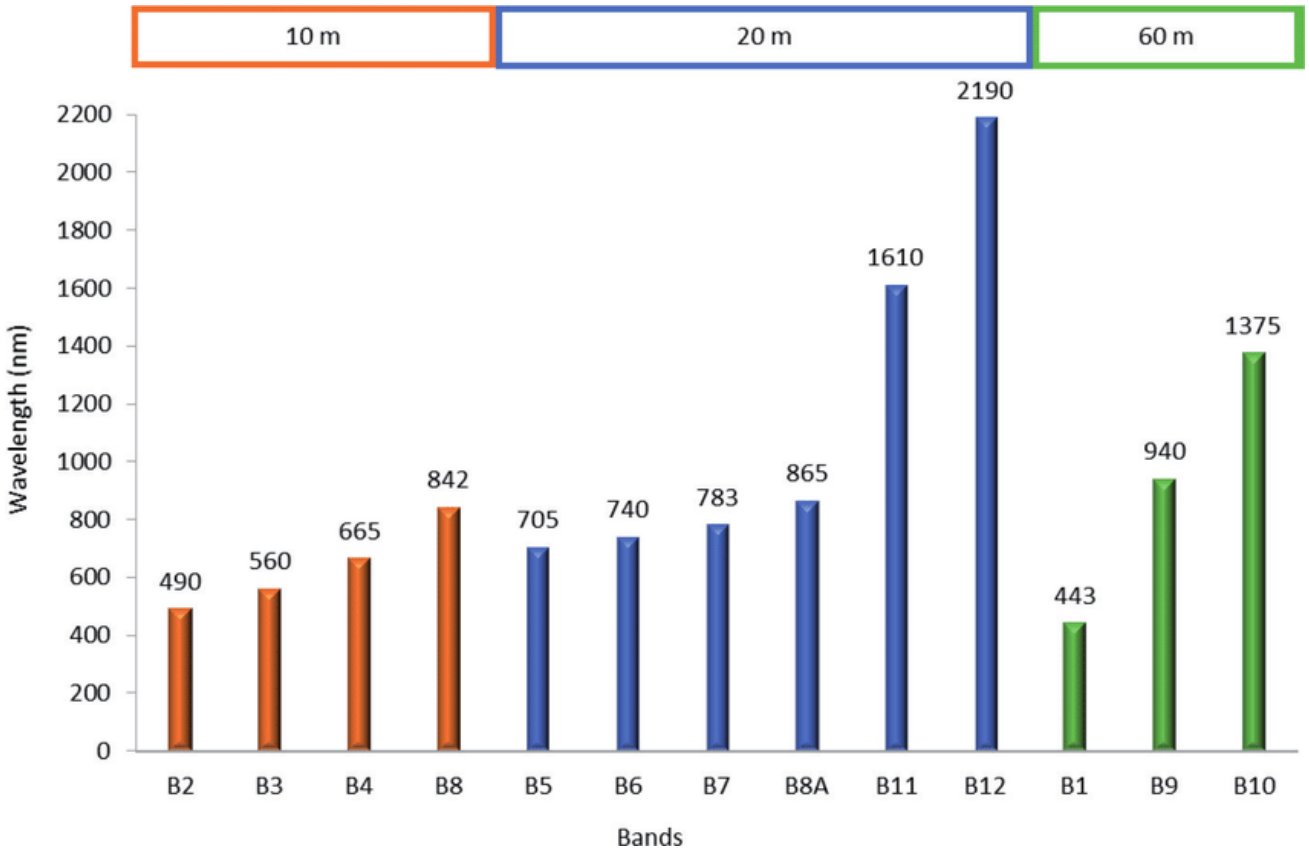


Рис. 2 - Довжина хвилі та роздільна здатність спектральних каналів Sentinel-2[9]

Запропонований інтелектуальний підхід

На рис. 3 зображено підхід, що використовувався при навчанні моделей. Розроблений підхід можна розділити на 3 етапи:

1. Знаходження набору інформативних ознак;
2. Перетворення вхідного зображення в простір ознак;
3. Навчання моделі.

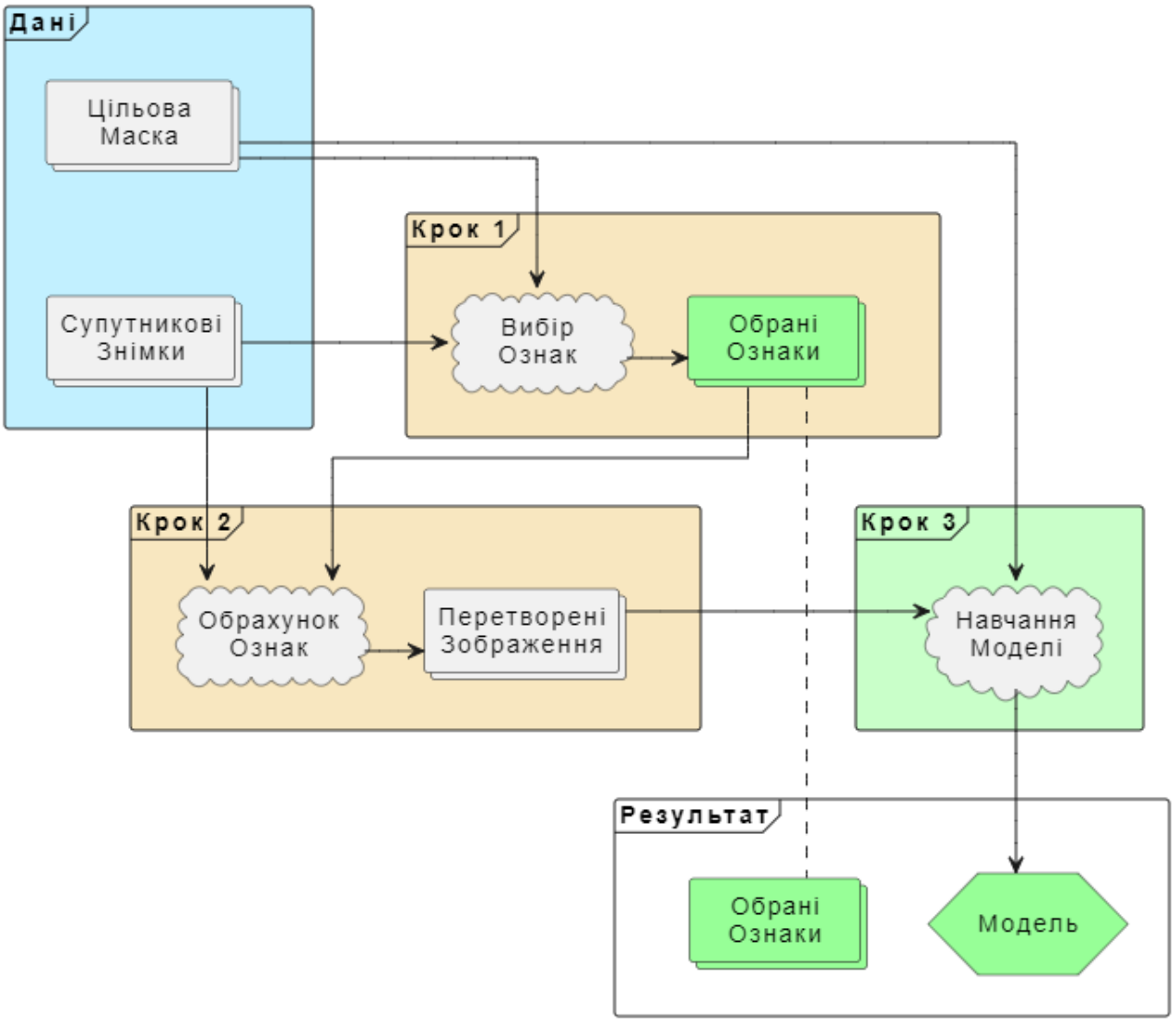


Рис. 3 - Схематичне представлення запропонованого підходу

Використаний підхід приймає на вхід набір ділянок. Кожна ділянка складається з багатоспектральних супутникових зображень, цільової маски та за необхідністю маски підстилаючої поверхні.

На першому етапі, на основі супутникових зображень та цільової маски обирається набір ознак, що буде використовуватися для навчання моделі.

На другому етапі, для кожного супутникового зображення рахуються обрані ознаки. Для кожної окремої ділянки отримані матриці ознак об’єднуються в одну матрицю ознак.

Фінальним етапом є навчання моделі на основі обрахованих матриць ознак та цільової маски. У результаті отримуємо модель, яку можна застосовувати для вирішення поставленої задачі сегментації.

Вегетаційні індекси як ознаки

Вегетаційні індекси [10] - це кількісні показники, які використовуються для вимірювання та аналізу рослинного розвитку та здоров’я на основі спектральних властивостей рослинності. Вони базуються на аналізі відбиття світла, яке рослини поглинають та відбивають в різних спектральних діапазонах.

Вегетаційні індекси дають змогуоцінити кількісні та якісні характеристики рослинного покриву, такі як фотосинтетична активність, біомаса, фенологічний стан та стресові умови рослин. Вони використовуються для вивчення змін в рослинному покриві на різних просторових та часових шкалах, від локального до глобального рівня.

Вегетаційні індекси широко використовуються в наукових дослідженнях, агрономія, екології, географії та інших дисциплінах. Вони дозволяють отримати об'єктивні дані про стан та динаміку рослинного покриву, що є важливим для розуміння та управління екосистемами Землі.

Зазвичай, обрахунок вегетаційних індексів потребує простих алгебраїчних операцій та використовує лише 2-3 канали зображення.

Оскільки вегетаційні індекси є кількісними оцінками стану та можуть бути спроектовані для вирізнення об’єктів серед інших, то вони виспупають (разом із спектральними канали) ознаками на основі яких відбувається сегментація.

Формування класів вегетаційних індексів

Оскільки вегетаційні індекси є математичними функціями, то їх можна узагальнити класами функцій.

Наприклад, розглянемо NDVI, що розраховується за формулою $NDVI = \frac{B8-B4}{B8+B4}$. Узагальнимо її, замінивши B4 та B8 на змінні канали A та B, отримаємо функцію вигляду $NORMP(A,B) = \frac{A-B}{A+B}$. На основі неї можемо

побудувати іншу функцію, присвоюючи кожній позиції змінних окрему змінну: $NORMP4(A,B,C,D) = \frac{A-B}{C+D}$.

Аналогічні перетворення виконаємо для спектральних каналів (результат яких буде функцією ідентичності) та кількох вегетаційних індексів й запишемо цей процес у вигляді ланцюга перетворень:

$$B4 \to B(A) = A \tag{1}$$

$$NDVI = \frac{B8-B4}{B8+B4} \to NORMP(A,B) = \frac{A-B}{A+B} \to NORMP4(A,B,C,D) = \frac{A-B}{C+D} \tag{2}$$

$$CIgreen = \frac{B8A}{B4} \to FRAC(A,B) = \frac{A}{B} \tag{3}$$

$$DSWI = \frac{B8+B3}{B4+B11} \to NORPP(A,B,C,D) = \frac{A+B}{C+D} \tag{4}$$

$$\begin{gathered} CVI = \frac{B8A \cdot B5}{B3^2} \to CVIbased_2(A,B,C) = \frac{A \cdot B}{C^2} \to \\ \to CVIbased(A,B,C,D) = \frac{A \cdot B}{C \cdot D} \end{gathered} \tag{5}$$

$$\begin{gathered} DRS = \sqrt{B4^2 + B8^2} \to DIST_2(A,B) = \sqrt{A^2 + B^2} \to \\ \to DIST_n(A_1,\ldots,A_n) = \sqrt{\sum_{i=1}^{n} A_i^2} \end{gathered} \tag{6}$$

Класи нечутливих до константного шуму вегетаційних індексів

Однією з проблем, яка часто з’являється на супутникових знімках є різний рівень яскравості на різних зображеннях. Деякі знімки світліші - інші темніші. У результаті цього, моделі, що працюють з пікселями супутникових знімків незалежно один від одного можуть показувати неочікувані результати.

Для подолання цієї проблеми пропонується використати вегетаційні індекси, що не чутливі до констатного шуму. Проте такі вегетаційні індекси відсутні серед відомих, тому потрібно їх розробити.

Нехай у нас є багатоспектральне зображення IMG, що має такий вигляд: $IMG = \{B1,\ B2,\ \ldots,\ B12\} = \{B1^{*} + V,\ B2^{*} + V,\ \ldots,\ B12^{*} + V\}$, де $V = const$ - для кожного пікселя.

Звідси можна побудувати такі класи вегетаційних індексів:

$$NORMP3(A,B,C) = \frac{A-B}{2+A+B-2\cdot C} = \frac{A^{*}+V-B^{*}-V}{2+A^{*}+V+B^{*}+V-2\cdot(C^{*}+V)} = \frac{A^{*}-B^{*}}{2+A^{*}+B^{*}-2\cdot C^{*}} \quad (7)$$

$$FRAC3(A,B,C) = \frac{A-C}{1+B-C} = \frac{A^{*}+V-C^{*}-V}{1+B^{*}+V-C^{*}-V} = \frac{A^{*}-C^{*}}{1+B^{*}-C^{*}} \quad (8)$$

Оскільки кожний канал зображення містить значення в діапазоні $(0,\ 1]$, а в знаменнику формул 7-8 є операція віднімання, то для запобігання ділення на 0 до знаменника додається константа.

Визначення цільової функції

Нехай потрібно з множини ознак $F = \{F_1,\ F_2,\ \ldots,\ F_N\}$ ($|F| = N$), де $F_i\colon R^b \to R$ - деяка ознака, знайти підмножину ознак L ($|L| = k \le N$, $L_i \in F$), яка дають змогу якомога краще розділити цільові класи.

Іншими словами, потрібно знайти таку підмножину F потужності k, що мала б найбільшу інформативність: $max\, I_k(L)$, де $I_k\colon F^k \to R$ - функція інформативності. Отже, перед нами стоїть задача оптимізації (максимізації).

Відстань Бгаттачар'я

Відстань Бгаттачар'я [11] використовується для визначення відносної подібності двох вибірок.

Для розподілів ймовірностей H, S відстань Бгаттачар'я визначається таким чином:

$$D_B(H,S) = -\ lnBC(H,S),$$

де $BC(H,S)$ - коефіцієнт Бгаттачар'я, який у випадку дискретного розподілу (чи гістограми) з n значень визначається таким чином: $BC(H,S) = \sum_{i=1}^{n} \sqrt{H_i \cdot S_i}$, де H_i, S_i - ймовірності i-го значення (висота i-го стовпця гістограм).

Якщо вибірки H, S є лінійно роздільними, то $D_B(H, S) = \infty$, якщо ж вони збігаються, то $D_B(H, S) = 0$.

Зазначимо, що значення відстані Бгаттачар'я чутливе до значення n. За надто малого – значення буде недооцінене, за надто великого – переоцінено.

Запропонована функція інформативності

Нехай маємо деякий набір з n ділянок $\Lambda = \{A_1, \ldots, A_n\}$, $A_i = (X, M_S, M_F)$, і необхідно знайти множину $L \subseteq F$ потужності k.

Введемо функцію $G(A, p) = \bigcup_{i=1}^{w} \bigcup_{j=1}^{h} X_{1ij} \mathbb{1}\{M_{Sij} = p, M_{Fij} = 0\}$, яка повертає множину пікселів зображення A, які відповідають цільовому класу підстилаючої поверхні ($M_{Fij} = 0$) та стану даної поверхні p ($M_{Sij} = p$).

Введемо допоміжні множини:

$d_S = \bigcup_{i=1}^{n} G(A_i, 1)$ - множина пікселів цільового стану в межах необхідної підстилаючої поверхні;

$d_H = \bigcup_{i=1}^{n} G(A_i, 0)$ - множина пікселів не цільового стану в межах необхідної підстилаючої поверхні.

Побудуємо функцію інформативності:

У випадку $k = 1$, пропонується використовувати запропонований в роботі [12] підхід з використанням відстані Бгаттачар'я в якості функції інформативності: $I_1(L) = D_B(L_1(d_H), L_1(d_S))$.

У випадку $k > 1$:

$$I_k(L) = I_1(L_1) + \sum_{i=2}^{k} I_1(L_i) \cdot \prod_{j=1}^{i-1} c_I(L_i, L_j), \quad (9)$$

де c_I: $F^2 \to R$ - коефіцієнт незалежності.

Запропонований коефіцієнт незалежності

Коефіцієнт незалежності $c_I(X, Y)$ має бути визначений так чином, щоб:

1. $c_I(X, X) = 0$
2. $c_I(X, Y) = c_I(Y, X)$
3. $c_I(X, Y) \geq 0$
4. $c_I(X, Y)$ має вказувати на ступінь залежності між X та Y

Зважаючи на це та на швидкість обчислень, в якості коефіцієнту незалежності пропонується використовувати таку функцію:

$$c_I(X, Y) = C - |r(X, Y)|, \quad (10)$$

де $C = const \geq 1$, $r(X, Y)$ - коефіцієнт кореляції Пірсона [13].

За замовчуванням $C = 1$, проте при збільшенні його значення, функція інформативності буде повертати більші значення тим наборам, окремі ознаки яких будуть більш інформативними.

Опис експерименту

При навчанні моделей для кожної їх конфігурації використовувалася 5-кратна перехресна валідація [14]. На кожному етапі якої, на основі ділянок, що потрапляють в навчальні дані, вирішувалася задача оптимізації для пошуку набору інформативних ознак. Для того, щоб було можливим оцінити ефективність обраного набору ознак, на його основі навчалося 5 моделей. Отже,

для кожної конфігурації загалом було навчено по 25 моделей; отримано 5 наборів ознак.

Обробка та аналіз супутникових знімків [15] у цій роботі виконувалися з використанням можливостей хмарної платформи CREODIAS, що є основною обчислювальною технологією в рамках ініціативи європейської підтримки EO4UA [16]. Ця платформа надає доступ до великих обсягів даних дистанційного зондування Землі та потужностей для роботи з ними.

Оптимізація набору ознак

Для вирішення задачі оптимізації використовувався генетичний алгоритм [17], реалізований в бібліотеці для Python - PyGAD [18] версії 3.2.0.

У всіх експериментах використовувалися однакові параметри генетичного алгоритму (табл. 1). Кількість генів (num_genes) дорівнює необхідній кількості ознак в цільовому наборі. Як цільова (fitness) функція використовувалася запропонована функція інформативності (формула 9). Функцією ж незалежності виступала формула 10, водночас розглядалися випадки, коли $C \in \{1, 1.6\}$. Всі не згадані параметри приймають значення за замовчуванням.

Табл. 1 - Гіперпараметри генетичного алгоритму

Гіперпараметр	Опис	Значення
num_generations	Кількість поколінь у генетичному алгоритмі	150
num_parents_mating	Кількість батьків, які будуть використані для створення наступного покоління	2
parent_selection_type	Метод відбору батьків для спарювання	tournament
keep_parents	Вказує, чи слід зберігати батьків у наступному поколінні.	0
keep_elitism	Вказує, чи слід зберігати найкращих	1

	особин у новому поколінні	
gene_type	Тип генів	int
sol_per_pop	Кількість особин у популяції	50
mutation_probability	Ймовірність мутації гена	0.2
allow_duplicate_genes	Параметр, що вказує, чи дозволяється дублювання генів в популяції	False

Модель машинного навчання

У даній роботі для вирішення завдання класифікації зображень лісової місцевості за наявністю заражених осередків використовувався багатошаровий перцептрон (MLP) [19, 20] – один з методів навчання штучних нейронних мереж. Як зазначено в оглядовій статті [21], такі методи є ефективним інструментом для інтелектуального аналізу даних дистанційного зондування Землі та прийняття рішень. На відміну від простіших класифікаторів, MLP може автоматично вивчати складні нелінійні залежності в даних, що робить його ефективним інструментом для розв'язання завдань семантичної сегментації у подібних областях.

MLPClassifier є реалізацією багатошарового перцептрону в бібліотеці scikit-learn (версії 1.3.1), яка дозволяє легко побудувати, тренувати та використовувати модель. Гіперпараметри (табл. 2) були обрані після ретельного емпіричного аналізу та оптимізації, спрямованої на досягнення найкращих результатів. Решта параметрів приймає значення за замовчуванням.

Таблиця 2 - Гіперпараметри використаної моделі

Гіперпараметр	Опис	Значення
tol	Параметр, що визначає, коли алгоритм зупиняється на підставі зміни функції втрат	5

alpha	Коефіцієнт регуляризації, який допомагає уникнути перенавчання	2e-5
activation	Функція активації для шарів нейронів	relu
hidden_layer_sizes	Кількість нейронів у прихованому шарі.	20
early_stopping	Забезпечує ранню зупинку навчання, якщо функція втрат на перевірочному наборі не покращується.	True
max_iter	Максимальна кількість ітерацій навчання	80

Конфігурації навчання

Було розглянуто випадки, коли модель навчається лише на основі зображення, що відповідає масці захворювання; та з додаванням зображення з минулого, яка не містить захворювання.

Для кожного з цих випадків було розглянуто 7 конфігурацій моделей (табл. 3).

Конфігурація №0 відповідає моделі, що просто використовує всі спектральні канали супутникового зображення.

Конфігурації №1-2 будуються на основі класу найбільш поширених вегетаційних індексів, а єдиною відмінністю між №1 та №2 - значення константи C, яка має впливати на фокус запропонованої функції інформативності.

У конфігурації №3 набір ознак формується таким чином, щоб додати щонайбільше нової інформації до набору, що складається зі спектральних каналів супутникового знімку. Тобто, перші 12 ознак є фіксованими, а інші 12 шукаються.

Конфігурації №4-6 використовують лише класи вегетаційних індексів, що не чутливі до константного шуму (зміни яскравості зображення).

Табл. 3 - Розглянуті конфігурації на основі 1 та 2 зображень

№	Назва конфігурації	Класи ознак	Кількість ознак	C
0	BANDS	B	12	-
1	NORMP_10	NORMP	12	1.0
2	NORMP_16	NORMP	12	1.6
3	B_NORMP_10	B + NORMP	12 + 12	1.0
4	PROP_10	NORMP3, FRAC3	12	1.0
5	PROP_16	NORMP3, FRAC3	12	1.6
6	24_PROP_10	NORMP3, FRAC3	24	1.0

Метрики оцінки моделі

Для того, щоб оцінити наскільки вдало моделі змогли передбачити належність того чи іншого пікселя зображення лісу до відповідного класу було обраховано наступні показники: Precision, Recall, F1 Macro, Intersection over Union (IoU), logistic loss (Log Loss) та Area Under the Receiver Operating Characteristic Curve (ROC AUC).

Precision вимірює точність класифікації для конкретного класу, і визначається як відношення правильно класифікованих пікселів даного класу до загальної кількості пікселів, які модель визначила як цей клас. Висока точність свідчить про те, що модель рідко неправильно визначає клас, але може не розпізнати деякі пікселі цього класу.

Recall (повнота) визначається як відношення правильно класифікованих пікселів даного класу до загальної кількості пікселів цього класу на зображенні. Тобто, для нашої задачі, як багато з усіх пікселів окремо здорового і хворого лісу модель змогла визначити правильно. Висока повнота свідчить про те, що модель добре визначає всі пікселі даного класу, але може помилитися, визначаючи деякі інші пікселі як цей клас.

F1 score об'єднує precision і recall узгодженим способом, даючи компромісне між ними значення. F1 Macro обчислює середнє значення F1 score по всім класам і є важливим показником, коли класи мають незбалансовану кількість пікселів, як у нашому випадку, коли кількість пікселів хворого лісу значно менше кількості здорових пікселів.

Intersection over Union (IoU) вимірює ступінь перекриття між прогнозованими і справжніми регіонами. IoU є вдалою метрикою для задач семантичної сегментації, з якою ми і маємо справу, адже вона дає нам оцінити наскільки повно реальна область хворого лісу покривається запропонованою.

Logistic loss (Log Loss) визначається як функція втрат для оцінки ймовірностей класифікації. Ця метрика показує наскільки сильно помиляється в своїх передбаченнях. Низький Log Loss свідчить про те, що модель видає впевнені й точні ймовірності класифікації.

Area Under the Receiver Operating Characteristic Curve (ROC AUC) вимірює якість роботи моделі відносно компромісу між чутливістю та специфічністю. Високий ROC AUC свідчить про те, що модель добре розділяє класи.

Аналіз результатів

Показники на валідаційному наборі даних

Розглядаючи випадок з моделями на основі одного зображення (табл. 4), доволі впевнено можна говорити про те, що в рамках індивідуальної точності найкращою є модель, отримана з конфігурації №2. Проте в середньому випадку доволі складно сказати, що якась конфігурація однозначно краща від іншої, оскільки середні значення кожної з метрик достатньо близькі між усіма конфігураціями.

Моделі, що навчалися на основі двох зображень (табл. 5), в середньому показали такі ж результати, як і моделі на основі одного зображення. Це може свідчити про те, що друге зображення є надлишковим, і що хворий ліс достатньо сильно відрізняється від здорового.

Говорячи про індивідуально найкращу конфігурацію з табл. 5, легко відмітити модель з конфігурацією №3. Такий результат не є несподіваним, оскільки дана конфігурація містить всі спектральні канали супутникових зображень та ще 12 вегетативних індексів. У результаті в теорії вона має мати показники не гірші, ніж у моделях з конфігураціями 0 та 1. Проте в середньому дана конфігурація показує схожі на них показники.

З табл. 4 та табл. 5 також видно, що моделі №4-6, що використовують розроблені (формули 7, 8) класи вегетаційних індексів показують конкуренті значення метрик у порівнянні з моделями №0-3, що говорить про перспективність їхнього використання.

Табл. 4 - Валідаційні метрики моделей, навчених на основі одного зображення

№		Precision	Recall	F1 Macro	log-loss	IoU	AUC
0	Max	0.860	0.761	0.869	0.158	0.611	0.970
	Mean	0.767	0.649	0.840	0.111	0.540	0.957
	Min	0.669	0.550	0.819	0.082	0.489	0.933
	Std	0.056	0.055	0.017	0.025	0.042	0.014
1	Max	0.894	0.723	0.864	0.311	0.653	0.966
	Mean	0.772	0.650	0.838	0.143	0.546	0.949
	Min	0.691	0.604	0.822	0.088	0.496	0.923
	Std	0.064	0.024	0.011	0.081	0.045	0.016
2	Max	0.847	**0.799**	**0.888**	0.140	**0.676**	0.974
	Mean	0.748	0.650	0.835	0.117	0.534	0.950
	Min	0.629	0.541	0.784	0.077	0.413	0.924
	Std	0.073	0.079	0.030	0.022	0.078	0.019
3	Max	0.857	0.729	0.878	0.181	0.628	**0.980**
	Mean	0.766	0.639	0.836	0.116	0.536	0.952
	Min	0.665	0.512	0.779	0.077	0.408	0.925
	Std	0.068	0.065	0.030	0.034	0.066	0.017
4	Max	**0.896**	0.714	0.845	0.146	0.547	0.970
	Mean	0.794	0.611	0.832	0.116	0.522	0.955
	Min	0.697	0.532	0.818	0.085	0.494	0.926
	Std	0.062	0.058	0.008	0.019	0.016	0.015
5	Max	0.817	0.710	0.854	0.143	0.585	0.965
	Mean	0.769	0.611	0.828	0.111	0.516	0.955
	Min	0.699	0.433	0.766	0.067	0.374	0.939
	Std	0.035	0.088	0.030	0.025	0.071	0.009
6	Max	0.889	0.670	0.864	0.213	0.596	0.975
	Mean	0.773	0.635	0.835	0.127	0.533	0.956
	Min	0.591	0.605	0.796	**0.062**	0.442	0.922
	Std	0.096	0.018	0.020	0.048	0.046	0.018

Табл. 5 - Валідаційні метрики моделей, навчених на основі двох зображень

№		Precision	Recall	F1 Macro	log-loss	IoU	AUC
0	Max	**0.905**	0.738	0.859	0.188	0.609	0.970
	Mean	0.783	0.647	0.840	0.127	0.543	0.956
	Min	0.636	0.573	0.819	0.084	0.491	0.923
	Std	0.089	0.049	0.011	0.044	0.034	0.016
1	Max	0.812	0.764	0.851	0.209	0.584	0.974
	Mean	0.745	0.663	0.838	0.124	0.538	0.955
	Min	0.662	0.557	0.793	0.078	0.437	0.921
	Std	0.041	0.062	0.015	0.041	0.039	0.018
2	Max	0.867	0.731	0.859	0.206	0.585	0.968
	Mean	0.751	0.643	0.834	0.125	0.529	0.955
	Min	0.659	0.551	0.792	0.075	0.431	0.939
	Std	0.061	0.053	0.018	0.044	0.046	0.010
3	Max	0.897	**0.797**	**0.908**	0.138	**0.728**	**0.978**
	Mean	0.772	0.649	0.840	0.117	0.545	0.951
	Min	0.624	0.448	0.772	0.099	0.392	0.899
	Std	0.081	0.100	0.039	0.012	0.096	0.022
4	Max	0.832	0.793	0.851	0.142	0.623	0.971
	Mean	0.762	0.635	0.838	0.113	0.527	0.956
	Min	0.660	0.494	0.793	0.095	0.446	0.930
	Std	0.049	0.079	0.015	0.017	0.054	0.013
5	Max	0.825	0.698	**0.908**	0.159	0.578	0.976
	Mean	0.760	0.641	0.840	0.117	0.532	0.954
	Min	0.649	0.582	0.772	0.073	0.497	0.928
	Std	0.041	0.032	0.039	0.031	0.024	0.013
6	Max	0.847	0.717	0.854	0.200	0.570	0.961
	Mean	0.768	0.625	0.832	0.122	0.524	0.953
	Min	0.673	0.546	0.802	**0.072**	0.462	0.936
	Std	0.054	0.043	0.014	0.039	0.034	0.008

Аналіз обраних ознак

Розглянемо, які вегетативні індекси були обрані в найкращих моделях.

У випадку з моделлю на основі одного зображення при оптимізації був знайдений набір із 12 ознак (табл. 6). 8 з 12 обраних ознак мають високий рівень індивідуальної інформативності ($I_1(L_i) > 0.4$).

Табл. 6 - Обрані вегетаційні індекси в найкращій моделі на основі одного зображення з конфігурацією №2

№	Вегетаційний індекс	$I_1(L_i)$
1	$NORMP(B4, B3) = \frac{B4-B3}{B4+B3}$	0.573
2	$NORMP(B3, B4) = \frac{B3-B4}{B3+B4}$	0.573
3	$NORMP(B11, B6) = \frac{B11-B6}{B11+B6}$	0.556
4	$NORMP(B6, B4) = \frac{B6-B4}{B6+B4}$	0.551
5	$NORMP(B4, B6) = \frac{B4-B6}{B4+B6}$	0.551
6	$NORMP(B8, B4) = \frac{B8-B4}{B8+B4}$	0.550
7	$NORMP(B6, B5) = \frac{B6-B5}{B6+B5}$	0.504
8	$NORMP(B7, B5) = \frac{B7-B5}{B7+B5}$	0.479
9	$NORMP(B9, B3) = \frac{B9-B3}{B9+B3}$	0.158
10	$NORMP(B2, B11) = \frac{B2-B11}{B2+B11}$	0.025
11	$NORMP(B5, B1) = \frac{B5-B1}{B5+B1}$	0.009
12	$NORMP(B8, B7) = \frac{B8-B7}{B8+B7}$	0.002

Проте, як можна побачити з рис. 4, ці 8 ознак достатньо сильно корелюють одна з одною, а ознаки №1-2 та №4-5 так взагалі повністю

корельовані. Якщо ж поглянути на табл. 6, то ці повністю корельовані ознаки відповідають парам ознак, формули яких відрізняються лише знаком. У такий спосіб, в кожній із пар одну з ознак можна прибрати.

Разом із продубльованими ознаками в табл. 6 можна помітити, що ознаки №10-12 мають майже нульову індивідуальну інформативність, а тому можливо від них теж можна позбутися. Але для того, щоб бути впевненим в доцільності їхнього відкидання, краще скористатися методами зниження розмірності, які у рамках цього дослідження не розглядалися.

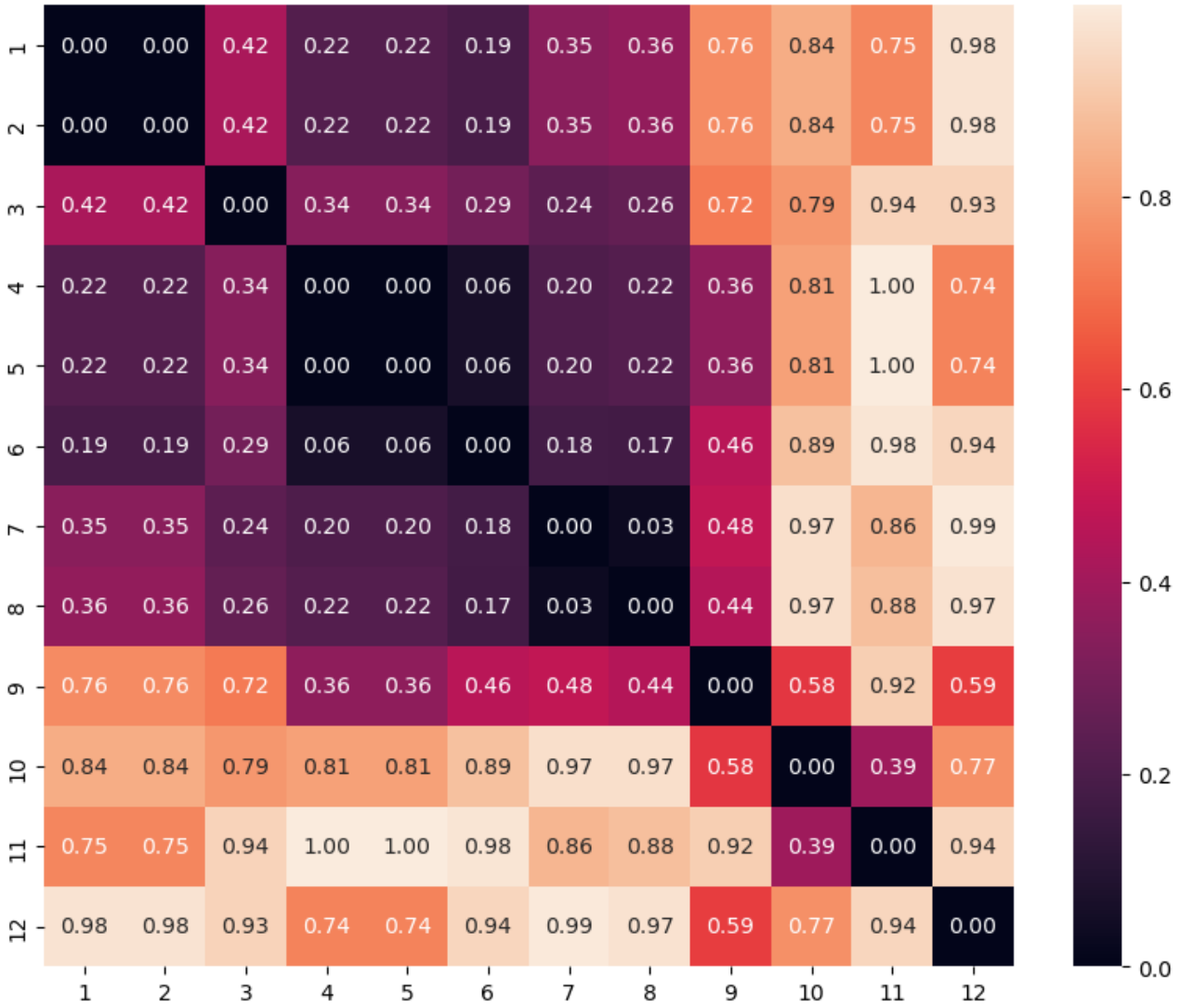


Рис. 4 - Матриця незалежності (C = 0) для набору ознак з табл. 6

Набір ознак (табл. 7), отриманий для моделі на основі двох зображень, має схожі проблеми. Так само, як і в попередньому наборі (табл. 6), він містить 8 ознак з інформативністю понад 0.4. Ознаки №2-3 відрізняються лише знаком, тому так само, як і в попередньому випадку, можна лишити лише один із них.

Аналізуючи рис. 5, можна помітити, що обрані вегетаційні індекси (з високим рівнем інформативності) значною мірою незалежні від наявних в наборі спектральних каналів.

Табл. 7 - Обрані вегетаційні індекси в найкращій моделі на основі двох зображень з конфігурацією №3

№	Вегетаційний індекс	$I_1(L_i)$	№	Спектральний канал	$I_1(L_i)$
1	$NORMP(B6, B12) = \frac{B6-B12}{B6+B12}$	0.582	9	$B4$	0.328
2	$NORMP(B3, B4) = \frac{B3-B4}{B3+B4}$	0.579	10	$B12$	0.291
3	$NORMP(B4, B3) = \frac{B4-B3}{B4+B3}$	0.579	11	$B11$	0.178
4	$NORMP(B4, B7) = \frac{B4-B7}{B4+B7}$	0.551	12	$B7$	0.126
5	$NORMP(B4, B8A) = \frac{B4-B8A}{B4+B8A}$	0.536	14	$B8$	0.108
6	$NORMP(B6, B5) = \frac{B6-B5}{B6+B5}$	0.515	16	$B8A$	0.097
7	$NORMP(B5, B7) = \frac{B5-B7}{B5+B7}$	0.492	17	$B2$	0.097
8	$NORMP(B11, B9) = \frac{B11-B9}{B11+B9}$	0.409	18	$B6$	0.095
13	$NORMP(B12, B1) = \frac{B12-B1}{B12+B1}$	0.113	19	$B5$	0.075
15	$NORMP(B8A, B7) = \frac{B8A-B7}{B8A+B7}$	0.098	20	$B9$	0.055
21	$NORMP(B3, B2) = \frac{B3-B2}{B3+B2}$	0.048	22	$B3$	0.048
24	$NORMP(B8, B7) = \frac{B8-B7}{B8+B7}$	0.002	23	$B1$	0.025

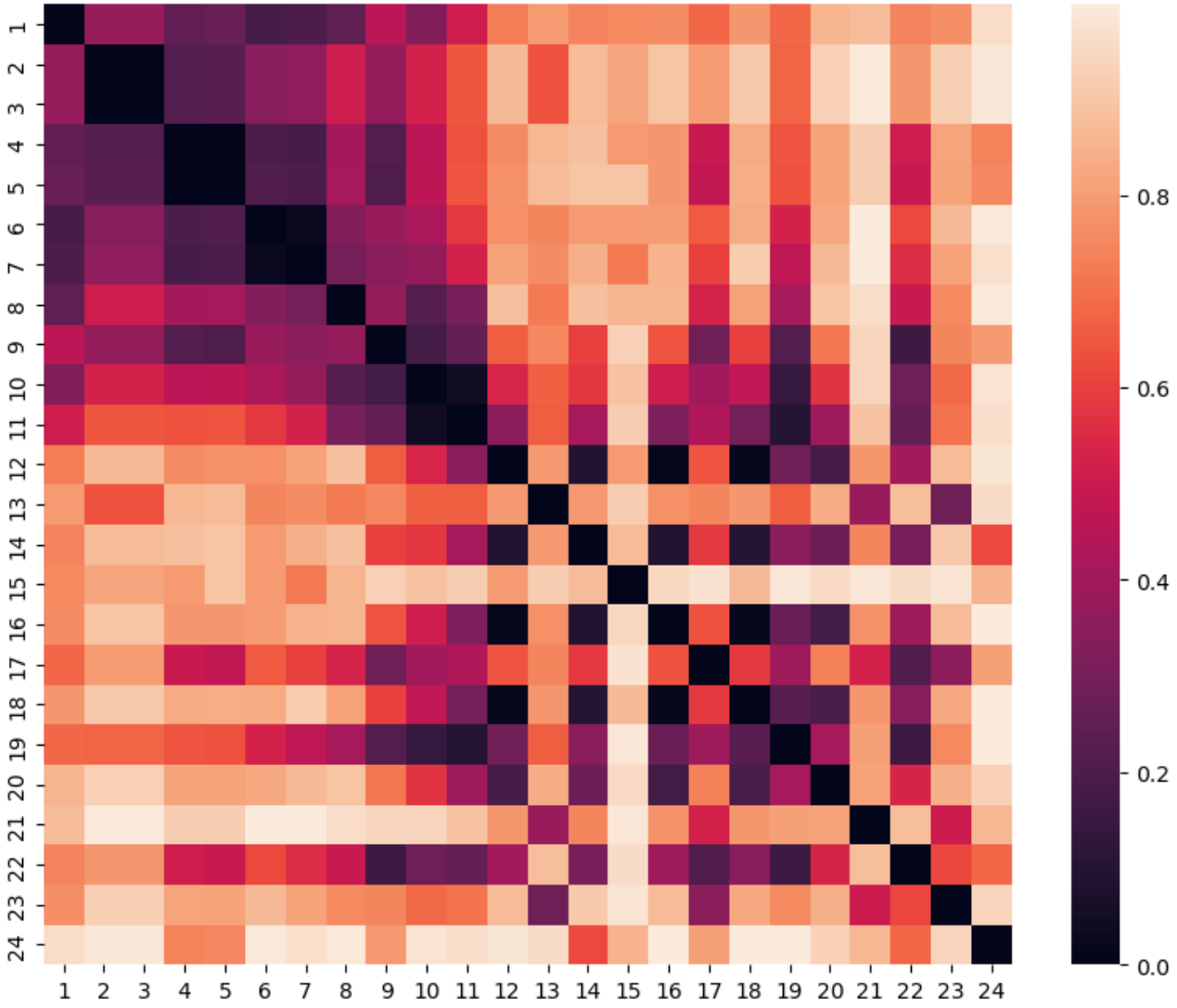


Рис. 5 - Матриця незалежності (C = 0) для набору ознак з табл. 7

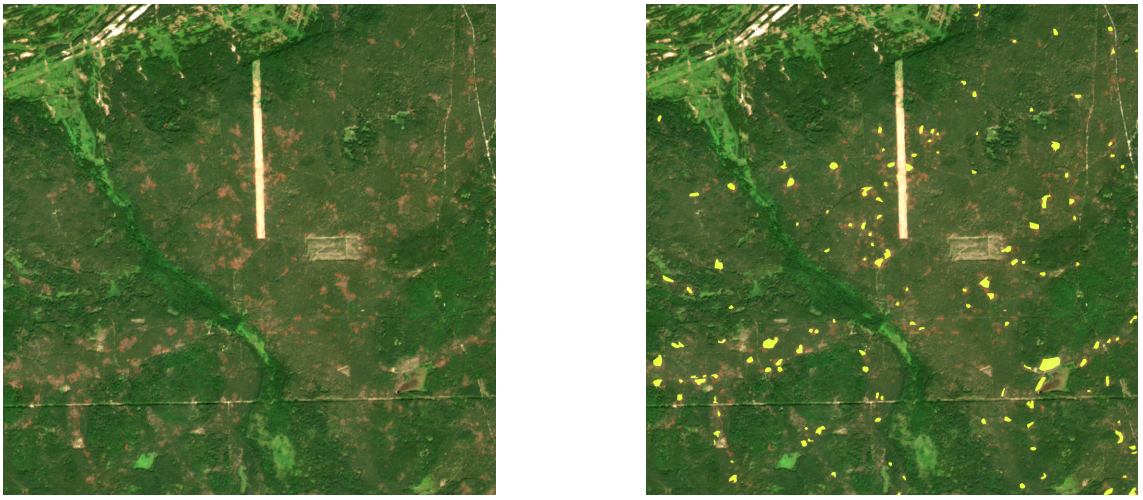
Огляд результуючої сегментації тестової ділянки

Для перевірки можливості узагальнення отриманих моделей проведемо огляд результатів сегментації території, яка не використовувалася при навчанні та валідації моделі й була розташована далеко від навчальної ділянки. Для цього було взято ділянку (рис. 1 (б)) на півночі Київської області, поблизу Чорнобиля.

Для даної території карта сегментації лісу на хвойний та листяний, отримана за допомогою моделі зі статті [22] мала проблему, що фрагменти хворого лісу вона часто класифікувала як не ліс. Через це було прийняте рішення не використовувати карти сегментації лісу та використати навчені

моделі на не відфільтрованих зображеннях (не відкидаючи пікселі з не хвойним лісом).

Надана лісниками карта хворого лісу для даної ділянки (рис. 6) не покриває всіх хворих фрагментів, а тому не може використовуватися для оцінки чисельних параметрів точності сегментації, проте її можна використати для налаштування порогового значення моделі.



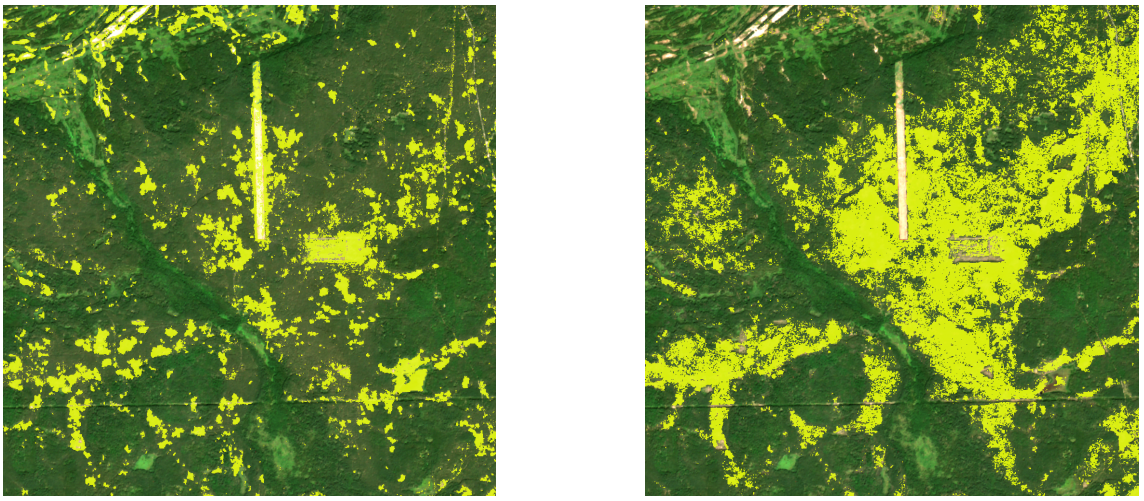
(а) (б)

Рис. 6 - Літнє зображення фрагменту тестової ділянки. (а) - без накладеної маски захворювання, (б) - з накладеною маскою захворювання

На рис. 7 (а, б) продемонстровано результати сегментації ділянки з рис. 7 (а) за допомогою моделей з використанням ознак з табл. 6 та табл. 7 відповідно. Якщо порівняти отриману сегментацію з наданою маскою захворювання лісу (рис. 7 (б)), то відразу можна побачити, що обидві моделі є надчутливими. Це може бути пов’язане з тим, що:

1. Критерії визначення лісу хворим відрізняються між французькими та українськими лісниками.
2. Моделі справді є надчутливими - якщо ліс не виглядає однозначно здоровим, то модель буде вважати його хворим.

Обидві наведені причини в більшості випадків можуть бути компенсовані при налаштуванні порогового значення ймовірності, після якого ліс буде вважатися хворим. Після налаштування порогового значення (рис. 8 (а, б)), результати моделей стали більш наближені до наданої маски.

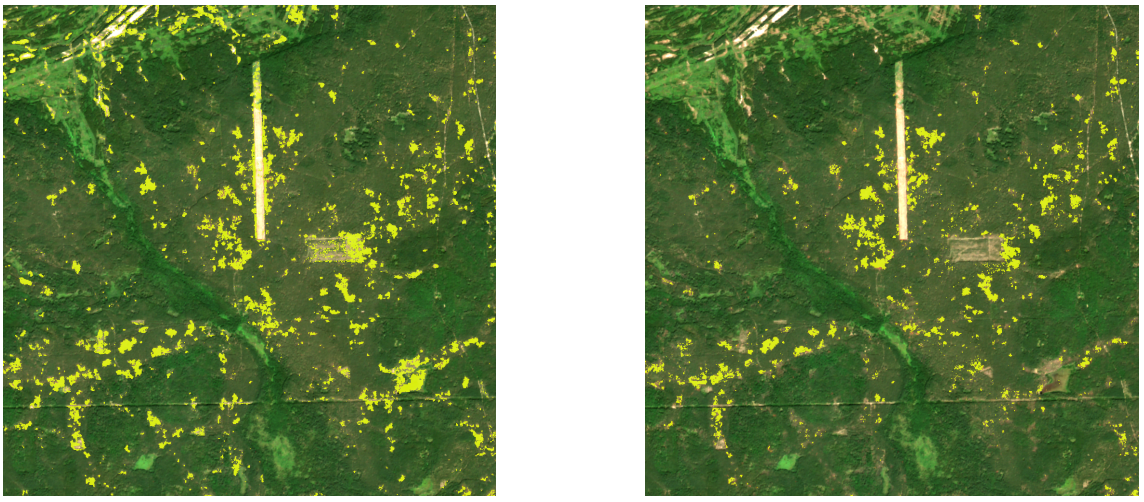


(а) (б)

Рис. 7 - Результати сегментації фрагменту тестової ділянки з пороговим значенням 0.5 накладені на його зображення. Жовтий - хворий ліс.

(а) - результати моделі на основі 1 зображення; (б) - результати моделі на основі 2 зображень.

Якщо порівнювати безпосередньо результати моделей між собою, то модель на основі 2х зображень (рис. 8 (б)) має значно менше проблем з листяного лісу та не лісу як хворого. Також дана модель при початковому пороговому значенні значно краще виділяє фрагменти лісу, що зазнають стресу (майже весь ліс на рис. 7 (а) має сірувато-коричневий колір).



(а) (б)

Рис. 8 - Результати сегментації фрагменту тестової ділянки накладені на його зображення. Жовтий - хворий ліс. (а) - результати моделі на основі 1 зображення з пороговим значенням 0.65; (б) - результати моделі на основі 2 зображень з пороговим значенням 0.85.

Висновки

У даній роботі було запропоновано автоматизований підхід до вирішення задачі моніторингу стану лісу, а саме виявлення уражених шкідниками хвойних лісів, на основі методів супутникового інтелекту. Запропонований підхід може бути використаний для вирішення інших задач сегментації на основі супутникових даних.

Підтверджена можливість використання відстані Бгаттачар'я та коефіцієнту кореляції Пірсона для побудови цільової функції для генетичного алгоритму, яка дозволяє знаходити оптимальний набір ознак для вирішення задачі сегментації.

Показано, що запропонована функція інформативності набору ознак дозволяє за допомогою методів оптимізації знаходити набори ознак, що можуть ефективно вирішувати реальні задачі. Як множину, з якої відбувається вибір ознак, запропоновано використовувати класи вегетаційних індексів, кілька з яких було представлено.

Також запропонована функція інформативності дозволяє виявляти вегетативні індекси, які найкраще розділяють класи, що може стати у нагоді при аналізі причин та наслідків захворювання лісів.

Запропоновано класи вегетаційних індексів, що є не чутливими до констатного шуму. Результати аналізу моделей, побудованих на їх основі, показали, що вони спроможні показувати не гірші результати у порівнянні з моделями на спектральних каналах чи інших класах вегетаційних індексів. Проте отримані моделі гарантовано показуватимуть однаковий результат незалежно від яскравості зображення.

Аналіз реальних даних показав, що використання запропонованого підходу дозволяє якісно вирішувати задачу моніторингу стану лісу на основі супутникових знімків Sentinel-2.

Результати проведених експериментів та їх аналіз показують, що найкращі показники точності та можливості узагальнення були

продемонстровані моделлю, що грунтується на двох зображеннях та комбінації спектральних каналів з отриманими за допомогою оптимізації вегетативними індексами. Дана модель змогла показати IoU на рівні 0.728.

Дослідження розвивало ідеї з робіт [12, 15, 21-22]. Отримані результати впроваджені в проекті SWIFTT (https://swiftt.eu/) програми Horizon Europe.

Отримані результати можуть використовуватися для оперативного моніторингу лісів та планування заходів щодо їх захисту і відновлення. Інформація про осередки ураження дозволяє ефективно спрямовувати ресурси для лікування хворих дерев. Розроблений підхід на основі штучного інтелекту значно пришвидшує процес аналізу стану навколишнього середовища порівняно з традиційними методами.

Отже, дослідження демонструє ефективність технологій штучного інтелекту для автоматизованого моніторингу лісових територій та може бути застосоване для вирішення актуальних завдань охорони лісів.

Використана література

1. Vojtěch Bárta, Petr Lukeš, Lucie Homolová, Early detection of bark beetle infestation in Norway spruce forests of Central Europe using Sentinel-2, International Journal of Applied Earth Observation and Geoinformation, Volume 100, 2021, 102335, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2021.102335.
2. Hoeser, T., Bachofer, F., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. Remote Sensing, 12(18), 3053.
3. Yuan, X., Shi, J., & Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications, 169, 114417.

4. Zhang J, Cong S, Zhang G, Ma Y, Zhang Y, Huang J. Detecting Pest-Infested Forest Damage through Multispectral Satellite Imagery and Improved UNet++. Sensors. 2022; 22(19):7440. https://doi.org/10.3390/s22197440
5. N. N. Kussul, N. S. Lavreniuk, A. Y. Shelestov, B. Y. Yailymov та I. N. Butko, "Land Cover Changes Analysis Based on Deep Machine Learning Technique," Journal of Automation and Information Sciences, т. 48, № 5, с. 42—54, 2016. doi: 10.1615/ jautomatinfscien.v48.i5.40. url: https://doi.org/10.1615/ jautomatinfscien.v48.i5.40.
6. Langning Huo, Henrik Jan Persson, Eva Lindberg, Early detection of forest stress from European spruce bark beetle attack, and a new vegetation index: Normalized distance red & SWIR (NDRS), Remote Sensing of Environment, Volume 255, 2021, 112240, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2020.112240.
7. Štursová, M., Šnajdr, J., Cajthaml, T., Bárta, J., Šantrůčková, H., & Baldrian, P. (2014). When the forest dies: The response of forest soil fungi to a bark beetle-induced tree dieback. ISME Journal, 8(9). https://doi.org/10.1038/ismej.2014.37
8. Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., & Ranagalage, M. (2020). Sentinel-2 data for land cover/use mapping: A review. Remote Sensing, 12(14), 2291.
9. Isbaex, Crismeire & Coelho, Ana. (2021). The Potential of Sentinel-2 Satellite Images for Land-Cover/Land-Use and Forest Biomass Estimation: A Review. 10.5772/intechopen.93363.
10. Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. Journal of sensors, 2017.
11. A. Ilnitskiy and O. Burba, "Statistical criteria for assessing the informativity of the sources of radio emission of telecommunication networks and systems in their recognition," Cybersecurity: Education, Science, Technique, 5 2019. doi: 10.28925/2663-4023.2019.5.8394.

12. Салій Є. В., Лавренюк А. М. Пошук значущих ознак для виявлення захворювань лісу на основі супутникових знімків // Матеріали XXI Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Теоретичні і прикладні проблеми фізики, математики та інформатики», (11 − 12 травня 2023 р., м.Київ, Україна). - 2023 р.- с. 419-422. (електронне видання)
13. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126(5):p 1763-1768, May 2018. | DOI: 10.1213/ANE.0000000000002864
14. A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Remote Sensing, 11(2), 185.
15. Andrii Shelestov, Yevhenii Salii, Nataliia Hordiiko, and Hanna Yailymova Current Advances on Cloud-Based Distributed Computing for Forest Monitoring in Ukraine Springer book series Lecture Notes in Networks and Systems “Advanced Approaches and Innovations in Up-to-Date Networks and Systems”, 2023 (Accepted, in print).
16. Kuzin, V., Musial, J., & Shelestov, A. (2022, December). EO4UA Initiative: Scientific European Support of Ukrainian Scientific Community. In 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 1-5). IEEE.
17. Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, 80(5). https://doi.org/10.1007/s11042-020-10139-6
18. Python genetic algorithm!. PyGAD. (n.d.). https://pygad.readthedocs.io/en/latest/#
19. P. M. Atkinson & A. R. L. Tatnall (1997) Introduction Neural networks in remote sensing, International Journal of Remote Sensing, 18:4, 699-709, DOI: 10.1080/014311697218700

20. Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.
21. Nataliia Kussul, Volodymyr Kuzin, Andrii Shelestov "Deep Learning for Remote Sensing, Earth Intelligence and Decision Making", Springer book series Lecture Notes in Electrical Engineering "Digital Ecosystems: Interconnecting Advanced Networks with AI Applications", 2024 (Accepted, in print).
22. Salii, Yevhenii & Kuzin, Volodymyr & Hohol, Anton & Kussul, Nataliia & Yailymova, Hanna. (2023). Machine Learning Models and Technology for Classification of Forest on Satellite Data. 93-98. 10.1109/EUROCON56442.2023.10199006.