DOI 10.31891/2307-5732-2025-347-64
УДК 621.3:681.34

РИБНИЦЬКИЙ МАКСИМ
Національний аерокосмічний університет «ХАІ»
https://orcid.org/0009-0000-1299-1604
e-mail: m.a.rybnytskyi@khai.edu
КРИВЕНКО СЕРГІЙ
Національний аерокосмічний університет «ХАІ»
https://orcid.org/0000-0001-6027-5442
e-mail: s.kryvevnko@khai.edu
ЛУКІН ВОЛОДИМИР
Національний аерокосмічний університет «ХАІ»
https://orcid.org/0000-0002-1443-9685
e-mail: v.lukin@khai.edu

ОГЛЯД ЗАСТОСУВАННЯ ТА МЕТОДІВ КЛАСИФІКАЦІЇ ДАНИХ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ

Наведено результати аналізу літературних джерел щодо методів класифікації даних, отриманих за допомогою дистанційного зондування, які мають значний вплив на різні галузі, включаючи медицину, господарство та урбаністику. В роботі проведено систематичний огляд літературних джерел, класифіковано за категоріями їхнього внеску у розвиток класифікації даних: від важливості дистанційного зондування до специфічних підходів класифікації та її проблематики. Особлива увага приділяється порівнянню традиційних методів класифікації та методів, що базуються на нейронних мережах, їхнім перевагам та обмеженням. Визначено ключові проблеми сучасних методів та окреслено перспективні напрямки подальших досліджень у цій області.

Ключові слова: дистанційне зондування, класифікація даних, нейронні мережі, машинне навчання.

RYBNYTSKYI MAKSYM,
KRYVENKO SERGII,
LUKIN VOLODYMYR
National Aerospace University – KhAI

OVERVIEW OF APPLICATIONS AND METHODS OF REMOTE SENSING DATA CLASSIFICATION

In this comprehensive study, we explore the multifaceted challenges and limitations associated with classification methods in machine learning, particularly focusing on image classification. Despite the remarkable achievements of machine learning models in this domain, there is a tendency to prioritize accuracy while overlooking other critical factors such as data preparation, the impact of noise on accuracy, computational power requirements, and more. A significant portion of the study addresses the issues of model interpretability and the inherent complexities of deep learning models, often referred to as "black box" models. These models, while highly accurate, pose significant challenges in terms of understanding and explaining their decision-making processes. This is particularly problematic in high-stakes environments such as healthcare and judicial systems, where decisions can have profound implications. Adversarial attacks represent another critical challenge discussed in this study. These attacks involve manipulating input data to deceive models into making incorrect predictions. We highlight the need for robust optimization techniques in decision tree thresholds to minimize potential losses in worst-case scenarios of data perturbations. Furthermore, the study delves into the computational demands of training deep learning models, emphasizing the environmental and economic constraints posed by increasing computational loads. The research suggests a shift towards more computationally efficient methods to mitigate these challenges without compromising performance. Additionally, the study discusses the challenges of applying machine learning in scenarios where only a limited number of labeled examples are available. Neural networks require large volumes of labeled data to perform effectively, a condition not always feasible, especially in specialized fields with limited data availability. This limitation necessitates the development of alternative strategies that can learn effectively from smaller data sets. In conclusion, while machine learning, particularly deep learning, continues to advance and achieve high accuracy in image classification, this study underscores the importance of addressing interpretability, adversarial robustness, computational efficiency, and the ability to perform under constrained data conditions. Developing models that are not only accurate but also interpretable, less resource-intensive, and capable of learning from limited data could lead to more sustainable and ethically responsible applications of machine learning technologies.

Keywords: remote sensing, data classification, neural networks, machine learning.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Дистанційне зондування (ДЗ) розширює можливості людини по вивченню та спостереженню навколишнього середовища, це пов'язано з тим що людське око обмежене невеликою частиною електромагнітного спектру. ДЗ сьогодні відіграє важливу роль у багатьох екологічних дисциплінах, таких як географія, геологія, зоологія, сільське господарство, лісове господарство, ботаніка, метеорологія, океанографія та цивільне будівництво [1], а також застосовується в медицині з використанням різноманітних сенсорів [2].

Класифікація даних ДЗ – це критично важливий процес, який використовується для розпізнавання різних типів даних, отриманих із супутників або за допомогою медичного обладнання. Цей процес передбачає аналіз різних спектральних та інших сигнатур, отриманих із даних, щоб віднести кожен піксель, їх групу або зображення в цілому до певного класу чи категорії на основі певних атрибутів (ознак). Класифікація може бути виконана за допомогою різних методів, починаючи від традиційних,

таких як метод максимальної правдоподібності чи кластеризація K-середніх, до більш просунутих та сучасних методів, що використовують машинне навчання і, зокрема, глибинне навчання.

Основною проблемою у використанні ДЗ є велика кількість даних, які потрібно ефективно обробляти для вирішення наукових та практичних завдань. Сучасні методи обробки даних мають великий потенціал для підвищення ефективності використання ДЗ, але ідеальне рішення ще не знайдено, що вимагає подальших досліджень у цій галузі. Це підкреслює важливість розвитку нових технологій та методів, які допоможуть краще розуміти та використовувати об'єми даних, що зростають, для вирішення критично важливих завдань.

Аналіз досліджень та публікацій

Після аналізу основних джерел публікації по темі можна розподілити наступним чином:

• Загальні джерела інформації (книги, навчальні посібники), де надана загальна інформація по ДЗ - від процесу отримання хвиль, їх реєстрації [3] до етапу класифікації отриманих даних для різноманітних галузей [4].

• Статті, у яких розглядаються питання прикладного застосування класифікації даних ДЗ. До таких питань належить моніторинг міст [5, 6] та сільське господарство [7], керування та подолання наслідків катастроф, як природнього так і техногенного характеру [8], а також моніторинг кліматичних змін [9]. Сюди ж відносяться техніки аналізу та інтерпретації медичних зображень, таких як рентгенівські, дентальні та знімки отримані за допомогою спеціального обладнання, яке дозволяє отримувати зображення високої роздільної здатності та виявляти навіть невеликі об’єкти [10].

• Джерела, де розглядають техніки та методи класифікації зображень [11-21], а також надається інформація про оцінку ефективності методів класифікації та деякі рекомендації щодо їх покращення.

• Джерела, в яких порівнюються різні методи класифікації на заданих наборах даних (не лише даних ДЗ), та метрики для оцінки цих методів [22-29].

• Джерела, в яких розглядаються проблеми методів класифікації, їхні особливості роботи та можливі опції покращення [30]. У роботі [31] розглядається питання змагальних (adversarial) атак та можливі шляхи боротьби з ними. Важливою проблемою є інтерпретація моделей прийняття автоматичних рішень [32]. У роботі [33] автори розглядають використання простих моделей та векторів ознак, що вимагає менше обчислювальних ресурсів та прискорює процеси тренування та навчання моделей в порівнянні з моделями глибокого навчання. Питання впливу зменшення якості на точність класифікації аналізується в [34]. Питання, які стосуються проблем пов’язаних з обчислювальними ресурсами, висвітлені у [35, 36].

Формулювання цілей статті

Метою цієї статті є огляд і аналіз існуючих методів класифікації у ДЗ з метою виявлення особливостей та визначення слабких місць при вирішенні завдань аналізу та інтерпретації даних. Цей огляд дозволить виявити потенційні напрямки для подальших досліджень та розробки нових, більш ефективних підходів до класифікації, які можуть покращити точність та ефективність обробки різних обсягів даних, отриманих від ДЗ.

Виклад основного матеріалу

Застосування даних дистанційного зондування. На рис. 1 зображена «Хмара слів», побудована на основі текстів [5-10]. Більший розмір тексту відповідає більшій кількості згадувань про слово. Із «Хмари слів» слідує, що роботи головним чином концентруються на застосуванні зображень (image, data) дистанційного зондування (Remote Sensing), моніторингу міської місцевості (urban), аналізу земельного покриву (land cover), навколишнього середовища (Environment) та класифікації поверхонь, які не пропускають (impervious surface) деякі види випромінювання.

Рис. 1. Хмара слів для джерел, що описують застосування методів класифікації для даних дистанційного зондування

Зображення ДЗ міської місцевості застосовуються для аналізу якості води в річках, плануванні підтоплень у випадку повеней, ризику пожеж, аналізу та плануванні транспортної інфраструктури міста та забудови. У роботі [5] проводять класифікацію типу поверхні за допомогою гіперспектральних зображень міської місцевості з використанням методу B-дистанції. Автори підкреслили важливість конкретних спектральних смуг для розрізнення міських матеріалів та відзначили, що гіперспектральні дані можуть покращити картографування міських районів. Однак відзначено і обмеження, які виникають через шум гіперспектральних сенсорів. У [6] розглянуто головні підходи та методи до класифікації зображень міської місцевості. По-піксельний, коли кожному пікселю присвоюється мітка класифікації, субпіксельний, коли допускається, що піксель може включати декілька класів. Субпіксельний підхід використовується, коли недоступні зображення високої роздільної здатності, для обробки таких зображень застосовують нейронні мережі. Існує також об'єкто-орієнтований підхід, який став можливим саме завдяки сучасним системам обробки та супутниковим зображенням високої роздільної здатності. В роботі [6] також відзначено, що використання додаткових спектральних смуг покращує точність систем класифікації.

Автор [7] розглядає важливість застосування даних ДЗ в сільському господарстві. Відмічено важливість застосування декількох джерел інформації від різних сенсорів (віддалених та наземних) та автоматизація обробки даних. Увагу приділено боротьбі з завадами та особливостям обробки даних в залежності від пори року чи зміни погоди.

У роботі [8] запропоновано використовувати оптичні, термальні, мікрохвильові, радіолокаційні та лідарні дані для виявлення наслідків стихійних лих та катастроф. Автори відмічають, що використання лише одного типу даних неможливе для роботи за різних умов. Також класичні підходи інтерпретації даних не підходять для швидкої реакції на надзвичайні випадки. Мультиспектральні дані ДЗ, отримані з різних супутникових систем, дозволяють оцінювати стан атмосфери, парниковий ефект на поверхні Землі, наземні та океанічні екосистеми. Автори [9] наголошують на обмеженнях таких систем, пов'язаних з різними типами даних та старінням сенсорів, тобто навіть один і той же сенсор, встановлений на різних системах у різні періоди часу, надає різні дані через старіння елементів та появу внутрішніх шумів, що в свою чергу ускладнює використання архівних даних.

В медичній сфері дані ДЗ використовуються для виявлення пошкоджень та захворювань. Проблеми, які намагаються вирішити, пов'язані з сегментацією зображень легенів, виявленням пухлин та структур у мозку, а також змін в кістковій тканині чи у клітинах тканин та органів. Використання глибоких нейронних мереж (НМ) та згорткових нейронних мереж (ЗНМ) показає високу ефективність в сегментації медичних зображень [10]. Головними складнощами в даній області залишається наявність якісних даних у достатніх кількостях, що пов'язано з регулюванням доступу до них.

Задача класифікації зображень у машинному у навчанні. У роботі [11] автори використали алгоритм Support Vector Machine (SVM) для по-піксельної класифікації типу покриву земної поверхні в східному Меріленді, США. Було використано два типи ядра: поліноміальне та на основі радіальних базисних функцій (radius basis function - RBF). Встановлено, що в залежності від типу ядра, форми областей класів різнились. Відмічено, що використання поліномів високих рівнів необхідне при зменшенні кількості параметрів класифікатора. При використанні 7 змінних точність класифікації покращувалась для поліномів до 4 порядку, подальше збільшення мало впливало на точність. Для RBF ядра зміна головного параметру γ (який визначає вплив одного навчального прикладу) на проміжку від 1 до 7,5 вносила лише невелике покращення точності класифікації (на 2-4%, в залежності від розміру тренувальних даних), і не вносила видимого покращення на проміжку від 5 до 20. Автори порівняли 4 алгоритми: SVM, MLC (maximum likelihood classifier), NNC (neural network classifier) та DTC (decision tree classifier). SVM показав найкращі результати при використанні 7 змінних у класифікаторі. Також зазначено, що на покращення в точності класифікації мали великий вплив кількість змінних для моделі, в той час, як тренувальні дані чи вибір алгоритму не приніс значного виграшу.

На рис 2. зображена "Хмара слів", побудована на основі текстів джерел [11-21]. Окрім уже згаданого дистанційного зондування (Remote Sensing), у роботах найчастіше згадуються, різні алгоритми (algorithm) класифікації (SVM, RF, decision tree), задачі аналізу покриву земної поверхні (land cover), піксельна класифікація (pixel), а також підходи до оцінки точності моделей (accuracy).

SVM застосовують у медицині для виявлення хвороби Альцгеймера, базуючись на знімках мозку людини (eigenbrains) у різних площинах. У роботі [12] провели перевірку трьох різних моделей, що базувались на SVM: лінійна, яка не використовує ядро, RBF-KSVM та POL-KSVM, які використовують RBF та поліноміальну функції відповідно. Запропоновані методи дали наступні результати, POL-KSVM показала вищу точність класифікації на рівні 92,36 ± 0,94, ніж лінійна SVM (91,47 ± 1,02), що очікувано для даного типу задач. Проте RBF-KSVM показала точність на рівні 86,71 ± 1,93, що є неочікуваним, оскільки RBF найчастіше використовується як ядро для SVM класифікаторів. POL-KSVM показав точність, яка не поступається найкращим з сучасних (state-of-the-art або SOTA) методам. Такі результати свідчать про те, що вибір класифікатора та конфігурації для нього являється важливою задачею і варто проводити оцінку ефективності декількох класифікаторів для вибору найкращого з них для вирішенні поставлених завдань.

Рис. 2. Хмара слів для джерел, що описують задачу класифікації зображень у машинному у навчанні

Автори [13] розглядають використання НМ з двовимірними та тривимірними згортками для моніторингу стану лісових масивів. Використано дві моделі НМ, одна з яких була зосереджена на аналізі послідовності зображень, а інша – на окремих зображеннях. З роботи слідує, що при наявності достатньої кількості (5 та більше) послідовних у часі зображень, модель на тривимірних згортках показує кращі результати, ніж при використанні трьох та меншої кількості зображень. Також авторами зазначено, що використання моделей, які базуються на згортках, вимагає великої кількості розмічених даних та значних обчислювальних ресурсів.

У праці [14] автори запропонували шестиступеневу систему класифікації гіперспектральних даних. Процес починається з попередньої обробки даних, це може бути фільтрація чи корекція зображень, наступним виступає етап – вибір інформативних каналів спектру, так як не всі канали несуть однакову кількість інформації, тому використання зайвих каналів не призведе до покращення результатів класифікації, але буде вимагати більше обчислювальних ресурсів. Окремою задачею являється виділення різних ознак, при використанні машинного навчання – цей процес можливо автоматизувати або виконувати в напівавтоматичному режимі. Виділення ознак, які не корелюють між собою, покращує якість фінальної класифікації. В роботі як приклад наведене використання спектральних та просторових ознак об'єктів. Даний концепт може бути використаний як основа для подальших досліджень складних класифікаторів, які включають декілька окремих рівнів підготовки або обробки даних.

Автори [15] розглядають Random forest (RF) (ліс випадкових рішень) для задач класифікації. Це ансамблевий класифікатор, який використовує множинні дерева рішень. Ансамблеві класифікатори - це методи машинного навчання, які поєднують кілька моделей для покращення точності прогнозування. Вони використовують різні техніки, такі як баггінг, бустинг і стекінг, щоб зменшити варіативність і підвищити стійкість моделі. Виходи декількох індивідуальних класифікаторів стають входом класифікатора "вищого рівня", який приймає рішення як їх поєднати для отримання результату [16]. При порівнянні класифікаторів SVM та RF для попіксельної класифікації класифікатор на базі RF зміг досягти схожої точності класифікації з SVM. Також RF потребував лише два параметра – кількість ознак (Mtry) та дерев (Ntree), тоді як SVM вимагав більшої кількості параметрів від користувача. Згадана загальна рекомендація [17] використовувати nTree рівним 500 та Mtry рівним кореню квадратному з кількості вхідних змінних. Класифікатор на базі RF надає змогу оцінити важливість різних ознак, що в перспективі може допомогти з вибором ознак для побудови вектору ознак. RF класифікатори менш чутливі, ніж інші класифікатори машинного навчання, до якості тренувальних даних та перенавчання. RF класифікатор підходить для класифікації гіперспектральних даних, де проблема розмірності та висококорельованих даних ставить серйозні виклики для інших класифікаційних методик. Зазначено, що RF менш чутливий до зміни параметра Ntree. У роботі [18] наявна інформацію про те, що при збільшенні кількості дерев помилка узагальнення завжди збігається і перенавчання не є проблемою завдяки «закону великих чисел». З іншого боку, зменшення кількості прогнозованих змінних (ознак) призводить до того, що кожне окреме дерево моделі стає менш сильним, але це також зменшую кореляцію між деревами, що підвищує загальну точність моделі. Враховуючи це, вибір кількості дерев та кількості ознак вимагає оптимізації для мінімізації похибки узагальнення. Встановлено, що моделі з більшою кількістю дерев та малим значенням параметра Mtry краще справляються зі зменшенням помилки узагальнення та мають меншу кореляцію між деревами. RF стійкий до зменшення кількості тренувальних даних та шуму. Так, зменшення розміру тренувальних даних на 50% та внесення шуму до 20% не мало значного в пливу на точність класифікації [18].

Ще один ансамблевий класифікатор, який використовує дерева рішень, носить назву Gradient Boosting Tree (GBT). Головна відмінність від RF полягає в тому, що дерева тренуються послідовно один

за одним. Алгоритм побудований таким чином, що кожне наступне дерево коректує помилки попереднього. Завдяки цьому GBT може бути більш точним, ніж RF. У [19] запропоновано застосування GBT для класифікації покриву земної поверхні та аналізу наявності достатньої кількості рослинності на поверхні. Незважаючи на невелику точність (F1-score знаходився в межах 68,56-69,04 %) запропонований підхід дозволив оцінити архівні дані та визначити загальну тенденцію зміни покриву земної поверхні, відповідно до отриманої задачі. Автори [20] використовують Boosted Regression Tree (BRT) для класифікації типу біомаси на півночі Австралії. Цей алгоритм схожий до попереднього і відрізняється лише незначними змінами в його реалізації на програмному рівні. Джерелом даних автори обрали знімки, що отримані з Global Information System (GIS) та Department of Science, Information Technology and Innovation (DSITI). Дані від різних сенсорів різняться за типом та роздільною здатністю. Тому важливим етапом перед навчанням моделі була підготовка даних. Автори поєднали дані, які отримані з різних джерел, та відійшли від попіксельної класифікації. Натомість дані об'єднали та отримали новий агрегований тип даних. В ході роботи автори звернули увагу на складнощі з обробкою великого об'єму даних. Кожен знімок Landsat покриває площу 185x185км та має розмір близько 300 Мб, а за рік доступно 22 знімки тієї ж області зйомки. Сюди ж додається проблема, пов'язана з різними типами даних, які потрібно обробляти спільно. Великий об'єм даних викликає складнощі з вибором потрібних тренувальних даних, а наявність інформації від різних сенсорів ускладнює їх агрегацію через різну просторову розрізнювальну здатність та різні кути погляду. Незважаючи на складнощі, авторам вдалось показати ефективність роботи BRT з великим об'ємом даних, а також його здатність до обробки інформації з неповними даними.

Окрім уже згаданих алгоритмів, для класифікації гіперспектральних зображень також використовують такі алгоритми як: Minimum Distance (MD), K-Nearest Neighbor (K-NN), а також K-mean Clustering та ISODATA для класифікації без нагляду [21].

Порівняння та оцінка ефективності алгоритмів класифікації. Шляхом аналізу джерел [22-29] була побудована хмара слів, зображена на рис. 3.

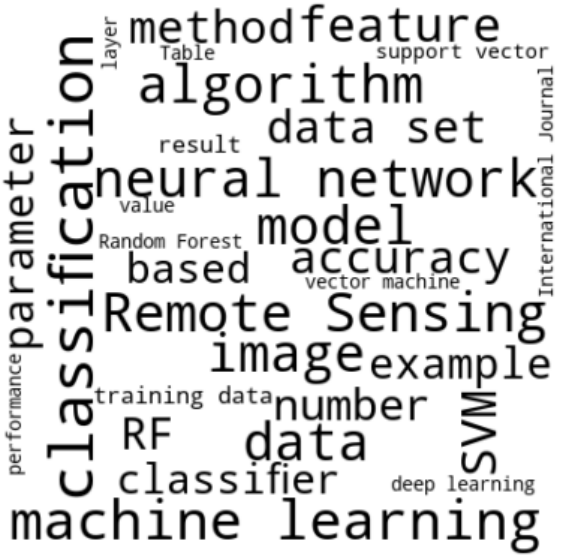

Рис. 3. Хмара слів для джерел, де проводять оцінку ефективності методів класифікації

З рис. 3 слідує, що окрім слів, які описують задачу класифікації в цілому (classification, Remote Sensing, machine learning), найбільш часто згадуються слова, пов'язані з різними методами класифікації (neural network, SVM, RF, feature, data set). Також можна виділити групу слів, пов'язану з оцінкою ефективності (accuracy, performance). У роботі [22] проведено детальний аналіз таких класифікаторів як RF, SVM, DT, Boosted DT, ANN та k-NN. Для оцінки ефективності роботи класифікаторів було обрано два набори даних: Indian Pines data set та GEOBIA data set. Вибір найкращого класифікатора відповідно до завдання є складною задачею. В різних джерелах зустрічається інформація, яка часто протирічить одна одній. Це пов'язано з тим, що роботи використовують різні процедури для конфігурації та оцінки моделей. Найбільш «чисті» результати можна отримати в рамках однієї роботи, коли різні алгоритми використовуються за однакових умов. Методи керованого навчання сильно залежать від тренувальних даних. Виявлено, що розмір та якість тренувальних даних мають більший вплив на точність класифікації, ніж вибір алгоритму.

У роботі [22] при оцінці згаданих наборів даних, SVM показав найкращі результати для Indian Pines data set (Таблиця 1), в той час як RF був найточнішим класифікатором для GEOBIA data set (Таблиця 2). Indian Pines включає 220 спектральних каналів, а у GEOBIA data set дані згруповані у об'єкти і кожен з них описується 147 змінними. Для видалення неінформативних змінних та зменшення їх загальної

кількості для оптимізації задачі класифікації було використано рекурсивний алгоритм видалення ознак (Виділення ознак (A)) описаний у [23], вдалось зменшити кількість змінних з 220 до 171 для Indian Pines data та з 147 до 121 змінної для GEOBIA data set. Також було використано ще один алгоритм (Виділення ознак (B)), описаний у [24], згідно з яким було використано 15% найінформативніших змінних, в результатів отримано 33 змінних для Indian Pines data та 22 для GEOBIA data set. Для оцінки впливу незбалансованості класів, в додатковому експерименті дані було вирівняно використовуючи метод подвоєння екземплярів класів, кількість яких була менша. Для оцінки результатів використано дві метрики, загальна точність класифікації та статистика Каппа. Загальна точність виражається у відношенні правильно класифікованих зарків до їх загальної кількості. Коефіцієнт Kappa [25] (або статистика Каппа) є мірою узгодженості між двома або більше класифікаційними системами, яка враховує випадкову згоду. Це корисний показник для оцінки точності класифікації даних ДЗ, оскільки він надає більш об'єктивну оцінку, ніж проста відсоткова точність.

Надані дані дають чітко зрозуміти, що неможливо визначити найкращий алгоритм машинного навчання для аналізу даних ДЗ. Проте очевидно, що ансамблеві методи більш ефективні ніж ті, що використовують один класифікатор.

Таблиця 1

Результати класифікації для набору даних «Indian pines»

Попередня обробка	**Міра точності**	**SVM**	**DT**	**RF**	**Boosted DT**	**ANN**	**k-NN**
Відсутня	Загальна точність (%) Каппа	89,1 0,844	78,3 0,687	87,1 0,812	87,2 0,817	85,1 0,787	78,6 0,686
Виділення ознак (A)	Загальна точність (%) Каппа	94,2 0,918	78,3 0,687	88,0 0,827	88,3 0,832	91,0 0,871	88,0 0,829
Виділення ознак (B)	Загальна точність (%) Каппа	86,1 0,801	77,2 0,672	83,4 0,761	83,0 0,758	86,3 0,803	84,1 0,772
Збалансовані тренувальні дані	Загальна точність (%) Каппа	89,5 0,850	65,8 0,541	87,4 0,820	87,0 0,814	43,0 0,269	76,3 0,666
Виділення ознак (A) + збалансовані тренувальні дані	Загальна точність (%) Каппа	94,4 0,921	65,8 0,542	87,8 0,826	87,6 0,823	85,9 0,802	87,3 0,820
Виділення ознак (B) + збалансовані тренувальні дані	Загальна точність (%) Каппа	86,6 0,810	71,2 0,594	83,8 0,768	82,3 0,760	83,3 0,763	82,1 0,748

Таблиця 2

Результати класифікації для набору даних «urban GEOBIA»

Попередня обробка	**Міра точності**	**SVM**	**DT**	**RF**	**Boosted DT**	**ANN**	**k-NN**
Відсутня	Загальна точність (%) Каппа	76,3 0,724	68,1 0,629	81,5 0,782	76,9 0,733	67,5 0,621	72,4 0,677
Виділення ознак (A)	Загальна точність (%) Каппа	76,9 0,730	68,8 0,636	81,7 0,785	77,3 0,735	71,4 0,669	75,0 0,706
Виділення ознак (B)	Загальна точність (%) Каппа	77,1 0,732	68,1 0,627	78,3 0,746	75,0 0,708	72,8 0,683	72,8 0,682
Збалансовані тренувальні дані	Загальна точність (%) Каппа	76,1 0,722	69,0 0,641	80,5 0,771	81,9 0,788	70,8 0,661	68,4 0,636
Виділення ознак (A) + збалансовані тренувальні дані	Загальна точність (%) Каппа	75,4 0,713	70,0 0,652	81,3 0,781	75,5 0,715	72,0 0,674	67,1 0,620
Виділення ознак (B) + збалансовані тренувальні дані	Загальна точність (%) Каппа	75,2 0,709	71,0 0,664	76,1 0,722	74,0 0,696	73,6 0,692	64,7 0,589

Автори [26] розглянули застосування нейронних мереж для аналізу даних електрокардіографії. Вони приділили увагу особливостям аналізу даних в умовах їх обмеженої кількості. Одним із рішень є виділення найголовніших ознак та зменшення розмірності даних. Такий підхід дозволяє зменшити вплив перенавчання моделей. Для більшості моделей побудова вектору ознак являється окремою задачею та потребує залучення оператора. На противагу цьому штучні нейронні мережі (ANN) спроектовані таким чином, що вони самі здатні виділити оптимальні ознаки для моделі, але такий підхід призводить до втрати можливості інтерпретувати зв'язок між вхідними даними та результатом. Найпростішою нейронною мережею є персептрон, який складається з функції передачі (зважена сума входів) та функції

активації. Відзначено два типи мереж для роботи з зображеннями – повнозв'язані мережі (FNN) та згорткові (CNN).

Відзначено, що такі алгоритм як RF сильно залежить від якості вектору ознак, але це має і ряд плюсів, так як завдяки тому, що внесок кожної з ознак відомий, користувач може інтерпретувати отримані результати. З іншої сторони ANN сама будує вектор ознак, опираючись на структуру даних, що призводить до неможливості співвідносити вхідні та вихідні дані.

Автори [27] провели порівняння SVM та CNN. CNN має складну внутрішню структуру, яка складається з декількох шарів: вхідний, згортковий, активаційний, об'єднання, повнозв'язаний та вихідний. Варто зазначити, що шар об'єднання може використовувати декілька методів роботи. Метод об'єднання середніх значень може ефективно зменшити вплив шуму на зображеннях, але він руйнує інформацію про структуру зображення. Метод максимальних значень може зменшити помилку згортки та зберегти структуру інформації зображення. У роботі було використано два набори даних: MNIST та Corel1000. MNIST – це великий набір написаних від руки цифр, від 0 до 9, в ньому є 10000 зображень з розміром 28x28. COREL1000 – це малий набір даних (1000 зображень), який складається з зображень, предметів, тварин та іншого, в даному дослідженні використовується для оцінки роботи алгоритмів в умовах обмеженої кількості маркованих даних. При використанні великого набору даних, SVM показала загальну точність 0,88 (виражається у відношенні кількості правильно класифікованих зращків до їх загальної кількості), а CNN – 0,98, час на побудову та тестування моделі, витрачений для SVM, склав 27,6 хвилини, а для CNN – 23,2 хвилини. При використанні малого датасету, точність SVM склала 0,86, а CNN – 0,83, затрачений час на модель SVM склав 1,02 хвилини, а на CNN – 2,01 хвилини. Експериментальні результати роботи показують, що традиційні методи машинного навчання краще підходять для роботи в умовах обмеженої кількості даних.

У роботі [28] автори оцінили ефективність платформи Google Earth Engine (GEE) для задач класифікації земної поверхні. З огляду на загальну точність класифікації, ансамбль нейронних мереж показав кращі результати у порівнянні з SVM, DT та RF. Це ще раз доводить, що ансамблеві класифікатори є перспективними.

При порівнянні [29] KNN, MLP та RF на наборі зображень Fashion-MNIST, MLP та RF показали близькі результати в точності класифікації, тоді як метод KNN показав найгірший результат, 89% проти 86% відповідно. Якщо говорити про час, затрачений на навчання моделі, то для RF було витрачено близько 35 с, тоді як для KNN - близько 106 с, а для ML - більше 500 с. Результати свідчать, що MLP показав найкращі результати точності класифікації, але цей алгоритм вимагає складної конфігурації параметрів та більше обчислювальних ресурсів, тоді як RF співставний по точності, але витрата часу на навчання в більше, ніж 14 разів, менша, що свідчить про можливість використовувати менше обчислювальної потужності для отримання схожих результатів.

Деякі проблеми методів класифікації та відомі обмеження. Шляхом аналізу джерел [30-36] була побудована хмара слів, зображена на рис. 4.

Рис. 4. Хмара слів для джерел, де описують деякі проблеми методів класифікації

Машинне навчання показує чудові результати в класифікації та розпізнанні зображень, проте зазвичай автори беруть до уваги лише точність та не зважають на інші фактори, такі як підготовка даних, вплив шумів на точність, необхідна обчислювальна потужність та ін. У [30] проведено широке опитування експертів в галузі машинного навчання та НМ. Найбільш спірними питаннями виявились: абстракція, генералізація, моделі пояснення, планування та втручання в навчання. В деяких питаннях експерти зійшлись на тому, що й досі неможливо пояснити, як моделі глибокого навчання працюють та

чому приймають ті чи інші рішення. Досі не створено достатньо ефективних підходів для розпізнання об'єктів на відео. Однією з найбільших проблем нейронних мереж являються «змагальні атаки», досі не розроблено ефективних методів для боротьби з ними. Цей термін використовується для опису ситуацій, коли зловмисник використовує спеціально підготовлені вхідні дані для введення в оману моделі машинного навчання. Ці атаки можуть призвести до того, що модель робить помилкові або неправильні прогнози. Особливу увагу цьому питанню приділено у [31]. Автори пропонують ідею, коли процес вибору оптимального порогу розділення в дереві рішень формується як задача робастної оптимізації. Це означає, що при виборі кривої розділення класів враховуються можливі збурення вхідних даних. Замість того, щоб просто максимізувати точність на оригінальних даних, метод прагне знайти криву, яка мінімізує можливі втрати в найгіршому сценарії збурень. Для кожної кривої в дереві рішень розглядається найгірший випадок збурень, і навчання зводиться до задачі оптимізації max-min. Це означає, що ми намагаємося максимізувати мінімальне значення оцінки, яке може бути отримано за всіх можливих збурень.

Поведінку моделі машинного навчання часто важко пояснити. Можна почути таке визначення як «black box», тобто модель, яка не піддається інтерпретації, типової для емпіричних ('white box') моделей. Автори [32] приділили особливу увагу цьому питанню. Вони розглянули ряд підходів до проблеми «black box». У першому допускається створення окремої моделі, яка могла б пояснити, чому інша модель машинного навчання прийняла те чи інше рішення, у другому - це створення моделі, яку можна пояснити та дати відповідь на питання, чому було прийняте те чи інше рішення. Автори відзначили, що існує думка, що чим складніша модель, тим вища її точність, тобто складні «black box» моделі необхідні для високої точності. Проте це не завжди так, особливо при наявності структурованих даних, з чітко вираженими ознаками. Сюди ж можна додати, що, чим точніша модель, тим гірше вона піддається інтерпретації. У роботі [32] також відзначено, що більшість зусиль сьогодні приділенні саме складним моделям машинного навчання, таким як НМ. Їх складність продовжує зростати, що призводить до ще більших труднощів з інтерпретацією результатів. В той час, як роботи, пов'язані з інтерпретованими моделями, складають невелику частин усіх відомих досліджень. Проблема інтерпретації дуже важлива для задач високо-критичних рішень, як наприклад у медицині чи судовій системі, де помилка (навіть якщо вона <1%) призводить до критичних наслідків для конкретної людини. Побудова моделей, які піддаються інтерпретації, має ряд складнощів. Потрібно вирішити задачу оптимізації при побудові моделі та створити модель з мінімальною кількістю умов, щоб вона краще піддавалась інтерпретації. Для таких моделей потрібно оптимізувати просторову систему оцінки, тобто визначити критерії оцінки та їх вагу. Окремою задачею являється комп'ютерний зір, де немає чіткого розуміння, як інтерпретувати результати розпізнання. Проте зустрічаються ідеї з використанням прототипних зображень, які містять визначені ознаки, що дозволять вбудувати в систему розпізнання алгоритм, який буде пояснювати, які ознаки були використанні та яка їх вага в прийнятті рішення. Тренування таких моделей являється окремою складною задачею, в тому числі через необхідність підготовки алгоритму виділення прототипних зображень.

Тренування глибоких моделей навчання комплексна задача, такі моделі вимагають великої кількості даних та спеціалізованого обладнання (GPU). У роботі [33] автори пропонують підхід, відмінний від нейронних мереж. ЗНМ мережі не потребують спеціальної обробки даних і здатні самі виділяти важливі частини інформації. Натомість автори пропонують спочатку підготувати дані, використовуючи попередню обробку, що може включати фільтрацію чи зменшення розмірності даних. Наступним етапом є підготовка вектору ознак, задача виділення високоінформативного вектору характеристик меншої розмірності, ніж оригінальні дані. В результаті класичні алгоритми машинного навчання здатні досягти продуктивності НМ, використовуючи обмежені ресурси. Автори зазначають, що для генерації вектору ознак були використані відомі бібліотеки, а сам процес не вимагав великих обчислювальних ресурсів. А для невеликих наборів даних такий підхід навіть перевершує нейроні мережі.

Автори [34] розглядають вплив обробки даних на продуктивність НМ. Дані ДЗ можуть піддаватись попередній обробці, таким як фільтрації та стиснення. В процесі дослідження виявлено, що для навчання моделі, ефективніше використовувати дані, які були стиснуті тим же алгоритмом, яким будуть стиснуті дані, які потрібно класифікувати потім. Також виявлено, що в процесі класифікації стиснутих зображень, моделі, що треновані із використанням стиснутих зображень, показують кращі результати, ніж моделі, де для тренування було використано зображення без стиснення.

У роботі [35] розглядається залежність прогресу глибокого навчання від зростання обчислювальної потужності. Автори підкреслюють, що нинішні тенденції можуть призвести до технічних та економічних обмежень через збільшення "обчислювального навантаження", що також має серйозні екологічні наслідки. З огляду на ці виклики, спільнота машинного навчання має або значно підвищити ефективність глибокого навчання, або перейти до більш обчислювально ефективних методів. Загалом, зростання обчислювального навантаження, яке супроводжує глибоке навчання, незабаром стане обмежувальним фактором для ряду застосувань, роблячи досягнення важливих етапів розвитку неможливими, якщо поточні тенденції збережуться. Фактично, розробка більш ресурсоефективних підходів у машинному навчанні стає життєво важливою, оскільки продовження нинішніх тенденцій

може не тільки викликати значні економічні та екологічні проблеми, але і обмежити доступ до новітніх технологій для більшості користувачів та малих підприємств. Тому необхідно розвивати та впроваджувати інноваційні рішення, які дозволять зменшити залежність від обчислювальної потужності, забезпечуючи при цьому продуктивність на рівні сучасних вимог.

До згаданої проблеми ще варто додати необхідність передачі великої кількості даних, що в умовах обмеженої пропускної здатності каналів вимагає багато часу та забезпечення достатнім розміром дискового простору для зберігання даних. Автори [36] пропонують до використання сучасні Хмарні рішення від Amazon Web Services(AWS) та Google Earth Engine (GEE). Вони дозволяють користувачам отримати доступ до відкритих супутникових даних Landsat та Sentinel. В той же час ці системи надають можливість використовувати майже необмежені обчислювальні ресурси та дисковий простір. Хмарні рішення дозволяють відкласти вирішення проблем, викликаних вимогами ресурсоємних алгоритмів машинного навчання, проте не вирішують їх.

Висновки

Огляд методів та засобів класифікації даних ДЗ підкреслив важливість і складність цього напрямку. Класифікація зображень і даних ДЗ відіграє критичну роль у різних галузях, зокрема в медицині, урбаністиці та управлінні ресурсами, де вона дозволяє автоматизувати обробку великих обсягів інформації. Використання мульти- та гіперспектральних даних значно покращує результати класифікації, підкреслюючи необхідність інтеграції різних типів даних. Також варто згадати про наявність шуму, який має місце при отриманні даних ДЗ, а також особливості стиснення зображень, що спотворені шумом. Вибір алгоритму класифікації залежить від конкретного застосування та доступних даних, а ансамблеві методи, які комбінують кілька класифікаторів, зазвичай показують кращу точність, ніж окремі класифікатори. Підбір параметрів алгоритму є складною емпіричною задачею, яка може вимагати значних ресурсів та спеціальних знань, а підготовка та розмір тренувальних даних можуть мати вирішальний вплив на точність класифікації. Незбалансовані тренувальні дані можуть спотворювати точність класифікації, і на малих наборах даних ефективніші традиційні методи машинного навчання, тоді як на великих наборах перевагу мають моделі глибокого навчання. Досі проблема "чорного ящика" в моделях глибокого навчання залишається важливою, особливо в критичних застосуваннях. Подальший ріст продуктивності алгоритмів класифікації вимагає створення нових підходів, так швидкість росту обчислювальної потужності в найближчому майбутньому може стати обмежувальним фактором для сучасних алгоритмів машинного навчання.

В результаті можна виділити такі перспективні напрямки досліджень:

- Дослідження методів класифікації на основі дерев рішень, зокрема застосування Gradient Boosting Tree.
- Розробка алгоритмів та підходів для ефективного виділення ознак.
- Розробка методів класифікації, що достатньо ефективно працюють в умовах обмеженого об’єму даних та обмежених обчислювальних ресурсів.
- Використання інтерпретованих моделей, особливо важливих для критичних застосувань, таких як медицина.
- Вивчення методів боротьби з змагальним атаками, які можуть впливати на рішення моделей.

Література

1. De Jong S., Meer F., Clevers J. Basics of Remote Sensing // *Remote Sensing Image Analysis: Including The Spatial Domain.* – 2007.
2. Suzuki S., Takemi M. Remote sensing for medical and health care applications // *Remote Sensing-Applications.* – Norderstedt, Deutschland: BoD–Books on Demand, 2012. – P. 479-492.
3. Lillesand T., Kiefer R., Chipman J. Remote Sensing and Image Interpretation. – 7th ed. – Wiley, 2015.
4. Jia X. Remote Sensing Digital Image Analysis: An Introduction. – Springer, 2006.
5. Herold M., Roberts D., Gardner M., Dennison P. Spectrometry for urban area remote sensing—Development and analysis of a spectral library from 350 to 2400 nm // *Remote Sensing of Environment.* – 2004. – Vol. 91, No. 3-4. – P. 304-319.
6. Weng Q. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends // *Remote Sensing of Environment.* – 2012. – Vol. 117. – P. 34-49.
7. Atzberger C. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs // *Remote Sens.* – 2013. – Vol. 5. – P. 949-981.
8. Joyce K., Belliss S., Samsonov S., McNeill S., Glassey P. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters // *Progress in Physical Geography.* – 2009. – Vol. 33. – P. 183-207.
9. Zhao S., Liu M., Tao M., Zhou W., Lu X., Li F., Wang O. The role of satellite remote sensing in mitigating and adapting to global climate change // *Science of The Total Environment.* – 2023. – Vol. 904.
10. Lee J., Jun S., Cho Y., Lee H., Kim G., Seo J., Kim N. Deep Learning in Medical Imaging: General Overview // *Korean J Radiol.* – 2017. – Vol. 18(4). – P. 570-584.

11. Huang C., Davis L., Townshend J. An assessment of support vector machines for land cover classification // *International Journal of Remote Sensing.* – 2002. – Vol. 23.

12. Zhang Y., Dong Z., Phillips P., Wang S., Genlin J., Yang J., Yaun T.-F. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning // *Frontiers in Computational Neuroscience.* – 2015. – Vol. 9.

13. Shelestov A., Bukhanevych R. Analysis of Satellite Data Time Series for Forest Monitoring Using Neural Networks Based on Three-Dimensional Convolutions // *International Scientific Technical Journal Problems of Control and Informatics.* – 2024. – Vol. 69. – P. 73-82.

14. Stankevich S., Piestona I., Podorvan V. Deep Learning Concept for Hyperspectral Imagery Classification // *Central European Researchers Journal.* – 2016. – Vol. 2. – P. 30-36.

15. Pal M. Random forest classifier for remote sensing classification // *International Journal of Remote Sensing.* – 2005. – Vol. 26(1). – P. 217–222.

16. Domingos P. A Few Useful Things to Know About Machine Learning // *Communications of the ACM.* – 2021. – Vol. 55. – P. 78-84.

17. Belgiu M., Dragut L. Random forest in remote sensing: A review of applications and future directions // *ISPRS Journal of Photogrammetry and Remote Sensing.* – 2016. – Vol. 114. – P. 24-31.

18. Rodriguez-Galiano V., Ghimire B., Rogan J., Chica-Olmo M., Rigol-Sanchez J. An assessment of the effectiveness of a random forest classifier for land-cover classification // *ISPRS Journal of Photogrammetry and Remote Sensing.* – 2012. – Vol. 67. – P. 93-104.

19. Handoko J., Herwindiati D. E., Hendryli J. Gradient Boosting Tree for Land Use Change Detection Using Landsat 7 and 8 Imageries: A Case Study of Bogor Area as Water Buffer Zone of Jakarta // *IOP Conf. Ser.: Earth Environ. Sci. 581.* – 2020.

20. Colin B., Clifford S., Wu P., Rathmanner S., Mengersen K. Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making // *Open Journal of Statistics.* – 2017. – Vol. 7. – P. 859-875.

21. Didore V., Nalawade D., Vaidya R. Remote Sensing Data Classification Technique: A Review // *International Journal of Advanced Research in Science, Communication and Technology.* – 2021. – P. 67-75.

22. Maxwell A., Warner T., Fang F. Implementation of machine-learning classification in remote sensing: an applied review // *International Journal of Remote Sensing.* – 2018. – Vol. 39(9). – P. 2784-2817.

23. Kuhn M., Weston S., Culp M., Coulter N., Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models. – 2015. [Online]. Available: https://cran.r-project.org/web/packages/C50/index.html. [Accessed 3 1 2025].

24. Murphy M. A., Evans J. S., Storfer A. Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics // *Ecology.* – 2010. – Vol. 91, No. 1. – P. 252-261.

25. McHugh M. Interrater reliability: The kappa statistic // *Biochemia medica: časopis Hrvatskoga društva medicinskih biokemičara / HDMB.* – 2012. – Vol. 22. – P. 276-282.

26. Mincholé A., Camps J., Lyon A., Rodríguez B. Machine learning in the electrocardiogram // *Journal of Electrocardiology.* – 2019. – Vol. 57. – P. 61-64.

27. Wang P., Fan E., Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning // *Pattern Recognition Letters.* – 2021. – Vol. 41. – P. 61-67.

28. Shelestov A., Lavrenik M., Kussul N., Novikov A., Skakun S. Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping // *Frontiers in Earth Science.* – 2017. – Vol. 5, No. 1-10. – P. 778 - 782.

29. Chugh R., Bhatia V., Khanna K., Bhatia V. A Comparative Analysis of Classifiers for Image Classification // *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence).* – 2020.

30. Cremer C. Examining expert disagreement over deep learning // *Prog Artif Intell.* – 2021. – Vol. 10. – P. 449-464.

31. Chen H., Zhang H., Boning D., Hsieh C.-J. Robust decision trees against adversarial examples // *36th International Conference on Machine Learning.* – 2019.

32. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead // *Nature Machine Intelligence.* – 2019. – Vol. 1. – P. 206-215.

33. Donckt J., Donckt J., Deprost E., Vandenbussche N., Rademaker M., Vandewiele G., Hoecke S. Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring // *Biomedical Signal Processing and Control.* – 2023. – Vol. 81.

34. Proskura G., Rubel O., Lukin V., Kryvenko S. On Classifier Performance for Remote Sensing Images Compressed by Different Coders // *Aerospace Technic and Technology.* – 2023. – P. 67-77.

35. Thompson N., Greenewald K., Lee K., Manso G. The Computational Limits of Deep Learning // *Ninth Computing within Limits 2023.* – 2023.

36. Shelestov A., Lavreniuk M., Vasyliev V., Shumilo L., Kolotii A., Yailymov B., Kussul N., Yailymova H. Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery // *IEEE Transactions on Big Data.* – 2019. – P. 1-12.