УДК 004.932:004.85:528.85 **DOI:** 10.31673/2412-9070.2026.022704

**П. О. ПРИСТАВКА**[1,2], д-р техн. наук, професор;
ORCID: 0000-0002-0360-2459
**О. Г. ЧОЛИШКІНА**[1], канд. техн. наук, доцент;
ORCID: 0000-0002-0681-0413
**О. С. ПОДСКРЕБКО**[1], канд. екон. наук, доцент;
ORCID: 0000-0001-5282-4691
**М. І. БОРИШКЕВИЧ**[2], студентка,
ORCID: 0009-0008-9407-3994
[1]Київський національний університет імені Тараса Шевченко
[2]Державний університет «Київський авіаційний інститут»

# ДОСЛІДЖЕННЯ ВПЛИВУ АНОМАЛІЙ ТА ДУБЛІКАТІВ У НАБОРАХ ДАНИХ АЕРОЗЙОМКИ НА ЯКІСТЬ ГЛИБОКОГО НАВЧАННЯ

***Якість навчальних наборів даних є одним із ключових чинників ефективності моделей глибокого навчання у задачах автоматизованої обробки аерозображень. Наявність аномальних та дубльованих спостережень у вибірках аерозйомки призводить до спотворення статистичних характеристик даних, зниження ентропії розподілу та погіршення узагальнювальної здатності нейромережевих моделей. У статті досліджено вплив таких спотворень на результати багатокласової класифікації зображень і проаналізовано ефективність різних статистичних методів виявлення аномалій у компактному латентному просторі.***

***Для формування інформативного представлення зображень використано згортковий автоенкодер, що відображає дані у низьковимірний латентний простір, придатний для подальшого статистичного аналізу. Виявлення аномальних спостережень здійснювалося із застосуванням методу трьох сигм, методу асиметрії та ексцесу, а також багатовимірного варіаційного ряду. Окремо проаналізовано вплив вилучення дублікатів зображень. Оцінювання якості навчальних даних виконувалося на основі ентропійних характеристик латентного простору та показників точності згорткової нейронної мережі ResNet50.***

***Експериментальні результати показали, що вилучення аномальних спостережень позитивно впливає на точність класифікації на тестовій вибірці, при цьому найефективнішим виявився метод трьох сигм, який забезпечив приріст точності до 2,1 %. Встановлено, що очищення даних супроводжується підвищенням ентропії латентного простору, що свідчить про зростання інформаційної насиченості та рівномірності розподілу спостережень і корелює з покращенням узагальнювальної здатності моделі. Отримані результати підтверджують доцільність використання статистичного аналізу латентних представлень для підвищення якості навчальних наборів аерозйомки.***

**Ключові слова:** аерозйомка, аномальні дані, дублікати зображень, глибоке навчання, згорткові нейронні мережі, автоенкодер, латентний простір, ентропія даних, класифікація зображень, узагальнювальна здатність.

## *Вступ*

Стрімке зростання обсягів цифрових даних, отриманих із засобів дистанційного зондування Землі, створює нові виклики для систем автоматичної обробки аерозображень. За прогнозами дослідницьких організацій, загальний світовий обсяг даних у найближчі роки продовжить збільшуватися експоненційно, що вимагає застосування високоефективних методів підготовки та очищення навчальних вибірок. Однією з ключових проблем у цьому контексті є наявність аномальних спостережень у навчальних наборах даних, які можуть суттєво погіршувати

точність роботи моделей глибокого навчання, зокрема у задачах багатокласової класифікації об'єктів аерозйомки.

Аномальні спостереження виникають з різних причин: похибки сенсорів, помилки анотації, атмосферні завади, відмінності у масштабі чи ракурсі знімання, а також через появу у вибірці рідкісних або нетипових об'єктів. Наявність таких даних призводить до підвищення рівня шуму, зсуву розподілів ознак, зменшення ентропії даних та погіршення узагальнюючої здатності моделей. Особливо чутливими до аномалій є згорткові нейронні мережі, яким притаманна висока ємність і схильність до запам'ятовування некоректних прикладів.

У зв'язку з цим актуальним є питання автоматизації процесу виявлення та вилучення аномалій з великих навчальних наборів даних. Традиційні статистичні методи - такі як правило трьох сигм, аналіз асиметрії та ексцесу або методи на основі варіаційного ряду – дозволяють виявляти крайові та малоймовірні спостереження, однак є недостатньо ефективними у випадку складних, високовимірних структур зображень. Натомість нейронні мережі-автоенкодери, здатні кодувати дані у компактний простір представлень, відкривають можливість поєднання статистичних та глибинних підходів, забезпечуючи точніше виявлення структурних аномалій.

Додатковою проблемою є наявність дублікатів зображень, які спотворюють статистику вибірки, знижують ентропію розподілу та можуть негативно впливати на результати навчання, спричиняючи перевагу одних класів над іншими.

Виходячи з цих міркувань, метою даної роботи є розроблення та експериментальна перевірка інформаційної технології автоматизованого виявлення аномалій у навчальних наборах даних аерозйомки із застосуванням мереж-автоенкодерів та статистичних методів. У роботі досліджено вплив різних методів очищення даних на структуру простору ознак, ентропію розподілу та точність класифікації зображень за допомогою згорткової нейромере-жевої архітектури ResNet50.

Результати дослідження підтверджують доцільність комбінування нейромережевих і статистичних підходів, а також демонструють здатність запропонованої технології підвищувати якість моделей класифікації за рахунок вилучення аномальних та дубльованих спостережень. Запропоновані методи можуть бути інтегровані у сучасні системи обробки аерозйомки та застосовані у задачах моніторингу територій, агровиробництва, картографування та інтелектуального аналізу зображень.

### *Аналіз останніх досліджень і публікацій*

Проблема виявлення аномальних спостережень у багатовимірних наборах даних залишається однією з ключових у сучасних системах машинного навчання, комп'ютерного зору та автоматизованої обробки зображень. Якість навчальної вибірки безпосередньо впливає на узагальнюючу здатність моделей, стабільність процесу навчання та поведінку алгоритмів у реальних умовах. Особливо це стосується задач класифікації аерозображень, де присутні значні варіації текстур, масштабу та умов знімання.

Однією з найпоширеніших груп методів ідентифікації аномалій є статистичні підходи. Методи оцінювання крайових спостережень за допомогою аналізу відхилень від основного розподілу даних, зокрема через побудову варіаційних рядів та оцінку параметрів щільності, продовжують активно застосовуватися у сучасних дослідженнях. У роботі [1] наведено сучасний підхід до застосування методу головних компонент (PCA) для аналізу багатовимірних даних, що дозволяє зменшити розмірність та виявити напрямки максимальної варіабельності. Методи щільнісної оцінки та статистичного аналізу розподілів також залишаються дієвими інструментами для пошуку малоймовірних точок.

З розвитком глибинного навчання статистичні методи дедалі частіше комбінуються з нейромережевими моделями. Сучасні огляди глибоких згорткових нейронних мереж (CNN) [2] підкреслюють їхню високу ефективність у задачах класифікації зображень, однак також акцентують на їхню чутливість до шумових та аномальних даних. Водночас систематичний аналіз робіт з виявлення аномалій у часових рядах і зображеннях [3] засвідчує, що глибокі моделі

мають значні переваги у виявленні складних структурних відхилень, які неможливо ідентифікувати класичними статистичними критеріями.

Особливе місце у виявленні аномалій займають **автоенкодери** - нейронні мережі, що здійснюють нелінійне зменшення розмірності та реконструкцію вхідних даних. Сучасні дослідження [4] демонструють їхню ефективність у задачах anomaly detection завдяки здатності формувати компактні латентні представлення, у яких відстані між нормальними та аномальними спостереженнями є більш інформативними. Подальший розвиток цієї ідеї представлено у [5], де варіаційні автоенкодери (VAE) застосовано до реальних промислових наборів даних, показавши високу точність виявлення відхилень за рахунок моделювання ймовірнісного простору латентних ознак. Із ростом складності даних та обсягів потокового трафіку автоенкодери також зарекомендували себе у задачах мережевої безпеки та обробки потоків [6].

Методи виявлення аномалій у високовимірних просторах представлені у сучасних оглядах [7, 8], де наголошується на проблемі «прокляття розмірності» та зниженні ефективності класичних відстаневих критеріїв. Автори пропонують комбіновані схеми, які включають попереднє зменшення розмірності (PCA, автоенкодери), а також використання щільнісних і топологічних моделей для визначення аномальних структур.

Суттєвий вплив на якість даних мають дублікати та надлишкові спостереження. Дослідження [9, 10, 11] демонструють, що дублікати зменшують ентропію розподілу даних, спричиняють штучну концентрацію вибірки та призводять до переобтяження моделі окремими класами. Значну увагу приділено методам обчислення ентропійних характеристик зображень та оцінюванню їх впливу на інформаційний вміст вибірки. Зокрема, робота [10] представляє сучасні алгоритми ентропійного аналізу для обробки зображень, а у [11] представлено великомасштабне дослідження ефективності методів порогової обробки та оцінювання інформаційної насиченості.

Виявлення near-duplicate та duplicate зображень активно розвивається у контексті великих наборів даних. Найсучасніші методи базуються на перцептивному хешуванні, сиамських мережах і Vision Transformers, що дозволяють ефективно знаходити дублікати у великих колекціях [12, 13]. Дослідження [14] також показує, що надлишковість у великих наборах даних негативно впливає на алгоритми машинного навчання, а її зменшення покращує продуктивність моделей.

Таким чином, сучасний стан досліджень свідчить про високу ефективність комбінованих підходів, що поєднують глибинні нейронні мережі (зокрема автоенкодери) та статистичні методи аналізу даних. Така інтеграція дозволяє виявляти як крайові, так і структурні аномалії, зменшувати надлишковість та покращувати якість даних для подальшого навчання моделей класифікації.

**Метою статті** є дослідження впливу аномальних та дубльованих зображень у навчальних наборах даних аерозйомки на узагальнювальну здатність і точність моделей глибокого навчання, а також оцінювання ефективності статистичних методів виявлення аномалій у латентному просторі, сформованому згортковим автоенкодером.

### *Основна частина*

У дослідженні використано навчальний набір аерофотознімків [15], що містить понад 12000 зразків, поділених на 11 тематичних класів, які включають природні та антропогенні об’єкти (будівлі, лісові ділянки, водні поверхні, дороги, поля тощо). Зображення мають роздільність 64×64 пікселів і характеризуються високим рівнем внутрішньокласної варіативності через зміну масштабу, освітлення, кута зйомки та локальних текстур. Водночас у даних спостерігається наявність повторюваних фрагментів місцевості, що зумовлює появу дубльованих зображень і підвищує ризик перенавчання моделей. Для оцінки ефективності очищення даних та виявлення аномалій у подальших експериментах як базову модель класифікації було обрано згорткову нейронну мережу ResNet50. Архітектура ResNet50 є загальновизнаним стандартом для задач класифікації зображень із неперервними текстурними структурами завдяки використанню залишкових блоків, які забезпечують стійкість моделі до збільшення глибини та стабільність процесу оптимізації [16]. Навчання моделі проводилося на початковому, неочищеному наборі даних, що дозволило отримати базову оцінку її якості. Згідно з графіком на рис.1, після десяти епох точність на тренувальній вибірці досягла значення 0.9567, тоді як тестова точність

становила близько 0.87. Така різниця між тренувальною та тестовою вибірками свідчить про наявність шумів, аномальних точок і дублікованих зображень у початкових даних, що підтверджує необхідність застосування спеціальних методів очищення.

Для переходу від сирого піксельного простору до компактного й більш інформативного представлення було створено згортковий автоенкодер. Енкодер моделі включає послідовність згорткових шарів із поступовим зменшенням просторової роздільності та збільшенням кількості каналів, що дозволяє виділити стійкі просторово-текстурні ознаки. У фіналі енкодера дані відображаються у 16-вимірний латентний простір - оптимальний компроміс між компактністю та збереженням ключової інформації. Декодер є симетричним відображенням енкодера й відповідає за реконструкцію вихідного зображення. Схема переходу від вхідного зображення до латентного простору та подальше зниження розмірності до тривимірного простору методом

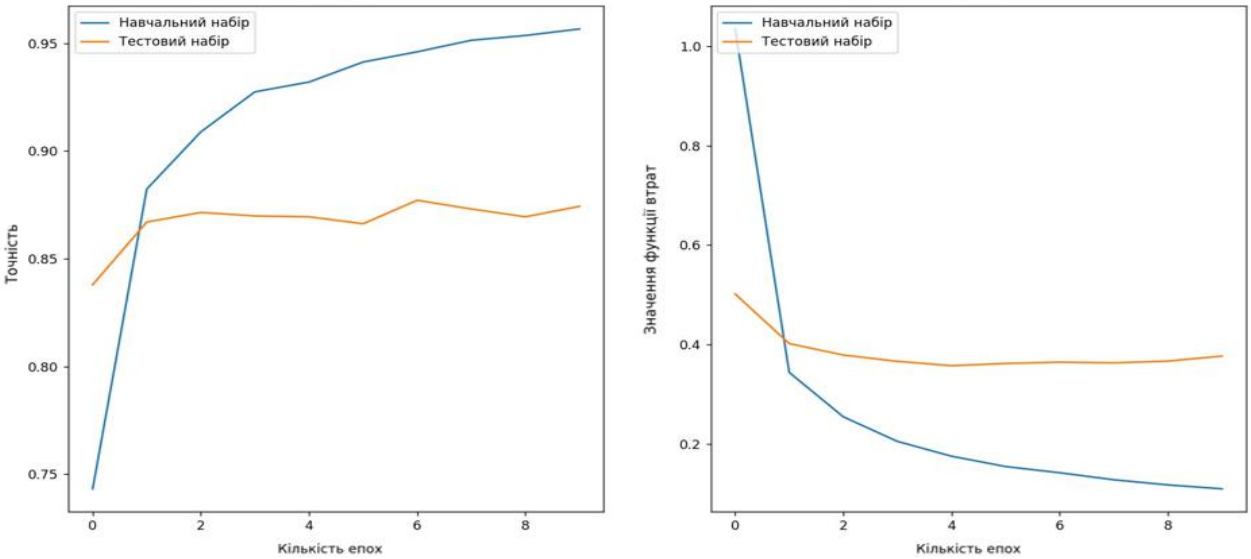


**Рис. 1. Графік залежності точності розпізнавання від кількості епох навчання мережі-класифікатора**

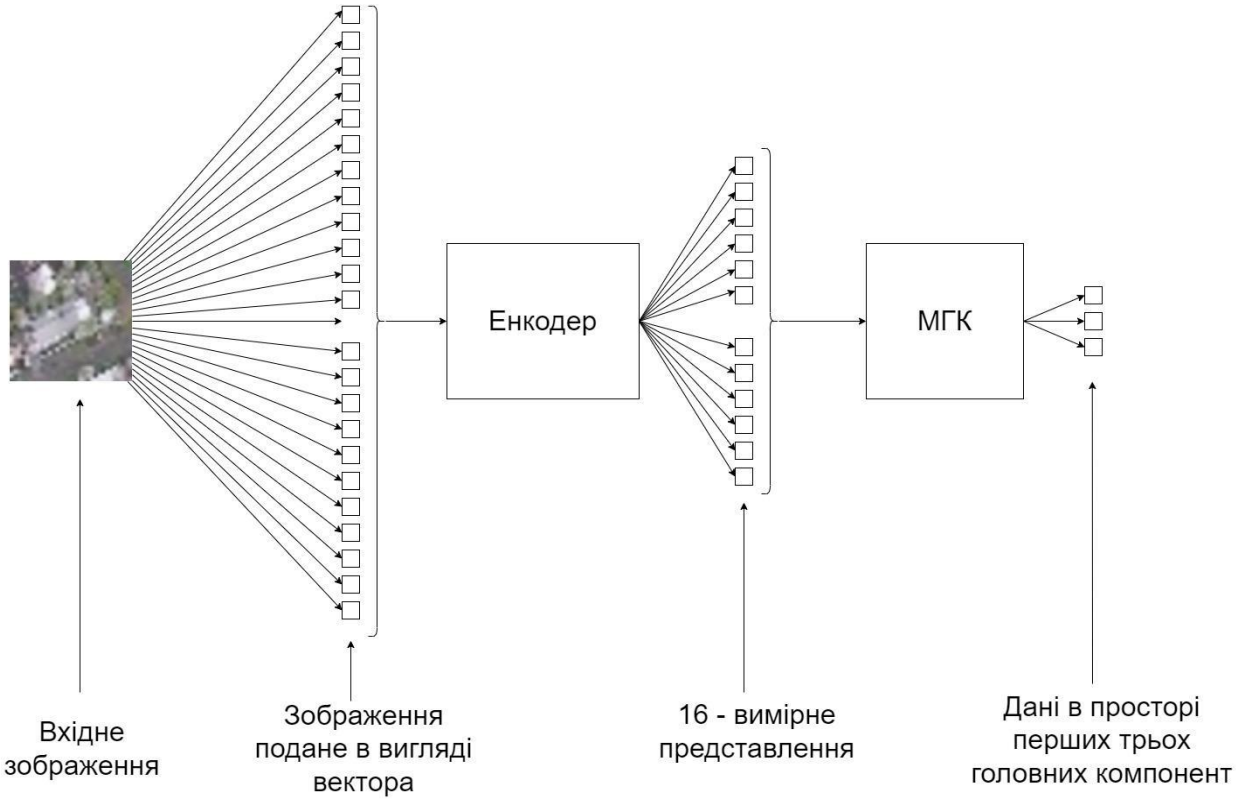


**Рис. 2. Схема переходу від початкових даних до 16 - вимірного та 3 - вимірного представлення**

головних компонент представлено на рис. 2, що ілюструє здатність моделі компактно структурувати дані, зберігаючи їхню внутрішню геометрію. Латентні представлення зображень є критично важливими для подальшого застосування статистичних методів виявлення аномалій, оскільки зменшують вплив шумів і дозволяють оцінювати структуру даних у вимірному та добре розділеному просторі.

Однією з важливих характеристик отриманого латентного простору є інформаційна ентропія**,** яка кількісно описує ступінь різноманітності та насиченості даних у компактному представленні. Диференціальна ентропія обчислюється на основі функції щільності розподілу $f(x)$:

$$h(x) = -\int f(u)\, ln \ln f(u)\, du\,. \quad (1)$$

Для варіаційного ряду розбитого на класи ентропію можна знайти наступним чином:

$$h(x) = -\sum_{i=1}^{M} p_i\, ln\, ln\, p_i\,. \quad (2)$$

Розподіл з максимальною ентропією є рівномірний, оскільки якщо ймовірність усіх подій однакова, то і невизначеність розподілу максимальна. Ентропія вказує, наскільки рівномірно розподіленими є спостереження у багатовимірному просторі та чи присутні серед них суттєві скупчення або, навпаки, порожні області. Висока ентропія вказує на різноманітність даних і відсутність домінування окремих фрагментів, що є бажаною властивістю для навчання моделей глибокого навчання. Низька ентропія, навпаки, свідчить про надлишковість даних (дублікати), наявність шумових скупчень, структурну нерівномірність розподілу. Такі фактори погіршують узагальнювальну здатність класифікатора, оскільки модель отримує менш інформативну та менш різноманітну вибірку.

Для виявлення аномальних зображень у латентному просторі було застосовано три статистичні підходи, кожен з яких дозволяє ідентифікувати різні типи відхилень.

Метод трьох сигм, який ґрунтується на припущенні про нормальність розподілу ознак. В основі цього методу лежить правило трьох сігм, згідно з яким 99.7% нормально розподілених спостережень буде знаходитися в межах трьох стандартних квадратичних відхилень від середнього арифметичного. Оскільки кількість 99.7% є дуже наближеною до 100%, маємо наступні нерівності:

$$\begin{gathered} x_{ij} > \bar{x_\iota} + 3 * \sigma_i \\ x_{ij} < \bar{x_\iota} - 3 * \sigma_i \\ i = \overline{1,n}; j = \overline{1,N}\,. \end{gathered} \quad (3)$$

Якщо одна з них справджується, то спостереження вважаємо аномалією, тобто, усі спостереження які віддалені від середнього арифметичного далі ніж на 3 середніх квадратичних відхилення, будемо позначати як аномалії. Попри простоту реалізації, метод здатний виявляти лише крайові відхилення та не враховує структурної складності розподілу даних.

Другим застосовано **метод асиметрії та ексцесу**, який спирається на незсунені оцінки відповідних коефіцієнтів. Коефіцієнт асиметрії характеризує асиметричність функції щільності відносно середнього, він буває:

- зсунений

$$\hat{A} = \frac{1}{N\hat{\sigma}^3}\sum_{l=1}^{N}(x_l - \bar{x})^3\,,$$

- незсунений

$$\overline{\mathrm{A}} = \frac{\sqrt{N(N-1)}}{N-2}\hat{A}\,.$$

Асиметрія буде рівною нулю при симетричності функції щільності. Асиметрія додатня, якщо функція щільності має ліву асиметрію та від'ємна, якщо функція щільності правоасиметрична.

Коефіцієнт ексцесу характеризує гостровершинність функції щільності вибіркового розподілу (гістограми) відносно теоретичного нормального розподілу. Коефіцієнт ексцесу може бути:

- зсунений

$$\hat{E} = \frac{1}{N\hat{\sigma}^4}\sum_{l=1}^{N}(x_l - \bar{x})^4 ,$$

- незсунений

$$\bar{E} = \frac{N^2-1}{(N-2)(N-3)}\left((\hat{E}-3) + \frac{6}{N+1}\right).$$

У цьому методі аномалії визначаються з нерівностей:

$$x_l \geq b\ ,\ x_l \leq a,$$

де $a, b$ – граничні значення на осі. Якщо хоч одна з цих нерівностей справджується, то $x_l$ позначають як аномалію.

Значення a та b знаходять із співвідношень:

$$\begin{aligned} a &= \bar{x} - t_2 S, \quad b = \bar{x} + t_1 S, \quad \text{якщо } \bar{A} < -0{,}2 \\ a &= \bar{x} - t_1 S,\ b = \bar{x} + t_2 S,\ \text{якщо } \bar{A} > 0{,}2 \\ a &= \bar{x} - t_1 S, \quad b = \bar{x} + t_1 S, \quad \text{якщо } |\bar{A}| \leq 0{,}2\ , \end{aligned} \tag{4}$$

де

$$t_1 = 2 + 0{,}1\ lg\ lg\ (0{,}04N)\ ;\ \ t_2 = (19(\bar{E}+2)^{0{,}5} + 1)^{0{,}5}.$$

Цей підхід є більш чутливим до форми розподілу латентних ознак і дозволяє враховувати його асиметричність і «гостровершинність». Даний метод показав кращу здатність виявляти структурні аномалії порівняно з трьома сигмами.

Третім застосовано метод багатовимірного варіаційного ряду. На відміну від попереднiх методів, він дає змогу виявляти аномалії не лише на периферії, а й у центральній частині розподілу — у вигляді рідкісних комбінацій ознак.

Для його використання необхідно побудувати для вхідних даних варіаційний ряд (посилання):

$$\left\{\left(x_{1,i_1}, \ldots x_{n,i_n}\right), n_{i_1 \ldots i_n}, p_{i_1 \ldots i_n}; i_k = \overline{1, M_k}\ , k = \overline{1, n}\right\} \tag{5}$$

з відносними частотами $p_{i_1 \ldots i_n}$ та задати граничне значення відносної частоти для малоймовірної події $p_{\text{гр}}$.

Тоді, маємо нерівність:

$$p_{i_j} \leq p_{\text{гр}}\ ,$$

якщо вона справджується, то позначаємо усі спостереження у відповідному елементі розбиття як аномалії. Цей метод виявився найбільш чутливим до внутрішньокласових спотворень і шумів.

Усі три методи застосовувалися як окремо, так і в комбінації з етапом вилучення дублікатів, що дозволило отримати декілька варіантів очищених навчальних наборів для подальшого порівняння.

У табл. 1 наведено зведені результати про зміну якості розпізнавання по окремим класам.

Таблиця 1

**Зміна якості розпізнавання зображень по кожному з класів в залежності від використаних методів вилучення аномалій**

| | Будівлі | Будівлі поміж дерев | Ліси | Земляні поля | Не вегетаційні поля | Стовпи | Дороги між будівлями | Дороги між деревами | Вегетаційні поля | Вода | Широкі грунтові дороги | Загальна точність |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Звичайне навчання | 0,89 | 0,91 | 0,98 | 0,84 | 0,97 | 0,89 | 0,77 | 0,8 | 0,86 | 0,87 | 0,8 | 0,87 |
| 3 сігма | 0,98 | 0,99 | 0,98 | 0,81 | 0,99 | 0,96 | 0,64 | 0,85 | 0,96 | 0,85 | 0,86 | 0,9 |
| Асиметрія та ексцес | 0,89 | 0,86 | 0,98 | 0,72 | 0,92 | 0,89 | 0,77 | 0,91 | 0,88 | 0,95 | 0,92 | 0,88 |
| Варіаційний ряд | 0,96 | 0,97 | 0,99 | 0,8 | 0,98 | 0,9 | 0,75 | 0,88 | 0,86 | 0,82 | 0,85 | 0,88 |
| Дублікати | 0,92 | 0,91 | 0,92 | 0,81 | 0,89 | 0,89 | 0,77 | 0,92 | 0,85 | 0,86 | 0,8 | 0,87 |
| Дубл. + 3сігма | 0,87 | 0,86 | 0,98 | 0,84 | 0,93 | 0,96 | 0,64 | 0,87 | 0,95 | 0,89 | 0,86 | 0,88 |
| Дубл. + асиметрія та ексцес | 0,92 | 0,91 | 0,92 | 0,81 | 0,89 | 0,89 | 0,77 | 0,82 | 0,85 | 0,86 | 0,92 | 0,87 |
| Дубл. + варіаційний ряд | 0,87 | 0,93 | 0,91 | 0,89 | 0,89 | 0,9 | 0,75 | 0,84 | 0,85 | 0,86 | 0,85 | 0,87 |

Отримані результати свідчать, що вилучення аномальних спостережень практично не впливає на точність класифікації на тренувальній вибірці: значення точності залишаються стабільними незалежно від застосованого методу очищення. Це є очікуваним, оскільки модель із високою потужністю апроксимації здатна ефективно підлаштовуватися під структуру наявних даних, навіть якщо в них зберігаються окремі відхилення.

Водночас, аналіз точності на тестовій вибірці демонструє суттєво інформативні результати. Показово, що вилучення дублікатів не лише не сприяло покращенню здатності моделі до узагальнення, а в окремих випадках навіть призвело до погіршення якості класифікації, зокрема й тоді, коли після видалення дублікатів застосовувалися додаткові процедури виявлення аномалій. Це свідчить про те, що дублікати, попри свою надлишковість, інколи виконують стабілізуючу роль для моделі, підсилюючи найбільш типові зразки класів та зменшуючи варіативність у межах тренувальної вибірки.

На відміну від цього, вилучення аномальних спостережень у початковому наборі даних продемонструвало помітне покращення результатів класифікації на тестових даних. Зокрема:

1) застосування методу асиметрії та ексцесу забезпечило приріст точності приблизно на 0.6%;

2) використання методу варіаційного ряду підвищило точність приблизно на 1%;

3) метод трьох сигм виявився найбільш результативним, збільшивши точність розпізнавання приблизно на 2.1%.

Перевагу методу трьох сигм можна пояснити його здатністю ефективно вилучати найбільш віддалені від середнього значення спостереження. Такі крайові точки часто не відображають типових ознак класів, а репрезентують локальні або випадкові особливості окремих зображень. Оскільки коректні зразки класу зазвичай згруповані ближче до центру багатовимірного латентного простору, видалення значних відхилень дозволяє зберегти найбільш інформативні,

репрезентативні дані, що й призвело до максимального покращення якості класифікації серед усіх досліджених методів.

### *Висновки*

У роботі досліджено вплив аномальних та дубльованих зображень у навчальних наборах аерозйомки на результати багатокласової класифікації із застосуванням згорткової нейронної мережі ResNet50. Показано, що наявність таких спостережень суттєво впливає на структуру латентного простору та узагальнювальну здатність моделі, незважаючи на стабільно високу точність на тренувальній вибірці.

Встановлено, що видалення аномальних спостережень у початковому наборі даних позитивно впливає на точність класифікації на тестовій вибірці. Найефективнішим серед досліджених підходів виявився метод трьох сигм, який забезпечив приріст точності до 2,1 % за рахунок вилучення найбільш віддалених від середнього спостережень, що не несуть суттєвої інформації про узагальнені ознаки відповідних класів. Методи, засновані на аналізі асиметрії та ексцесу, а також на багатовимірному варіаційному ряді, продемонстрували помірне, але стабільне покращення результатів, зокрема у випадках структурних внутрішньокласових відхилень.

Окремо проаналізовано вплив процедур очищення на ентропійні характеристики даних. Показано, що видалення аномальних та надлишкових спостережень приводить до підвищення ентропії латентного простору, що відображає зростання інформаційної насиченості та рівномірності розподілу даних. Збільшення ентропії корелює з покращенням узагальнювальної здатності класифікаторів, оскільки модель навчається на більш репрезентативній та менш структурно спотвореній вибірці.

Отримані результати підтверджують доцільність попереднього статистичного аналізу латентних представлень аерозображень та обґрунтованого вилучення аномальних спостережень як етапу підготовки даних для задач глибокого навчання.

Запропонований підхід може бути інтегрований у сучасні системи обробки аерофотознімків і використаний для підготовки даних у задачах моніторингу територій та інтелектуального аналізу зображень. У подальших дослідженнях доцільно розширити підхід за рахунок використання варіаційних автоенкодерів, методів щільнісного моделювання та метрик, чутливих до локальної геометрії латентного простору.

**Внесок авторів**

Пилип ПРИСТАВКА – концептуалізація, методика; Ольга ЧОЛИШКІНА – аналітичне опрацювання джерел, методика, підготовка первинного тексту; Оксана ЗОЛОТУХІНА – аналітичне опрацювання джерел, методика, візуалізація, редагування та доопрацювання; Олександр ПОДСКРЕБКО – збір і перевірка емпіричних даних, систематизація, підготовка та організація джерел і даних для подальшого аналізу, емпіричне дослідження; Марія БОРИШКЕВИЧ – підготовка первинного тексту, програмне забезпечення; емпіричне дослідження.

**Декларація про штучний інтелект**

Автори декларують, що штучний інтелект не використовувався для генерування наукового змісту, результатів дослідження, інтерпретації даних або формулювання висновків. Усі наукові положення статті є результатом самостійної роботи авторів.

**Конфлікт інтересів**

Автори заявляють про відсутність конфлікту інтересів та підтверджують, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Робота виконана відповідно до принципів академічної доброчесності, етичних норм проведення наукових досліджень та вимог редакційної політики щодо запобігання конфлікту інтересів.

***Список використаної літератури***

*1. Gewers, F. L., Ferreira, G. R., de Arruda, H. F., Silva, F. N., Comin, C. H., Amancio, D. R., Costa, L. F. Principal Component Analysis: A Natural Approach to Data Exploration. ACM Computing Surveys, 54(4), 1–34, 2021. https://doi.org/10.1145/3447755*

*2. Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., Ilono, P. Deep Convolutional Neural Networks: A Comprehensive Review. Preprints, 2024, Article 202408.1288. https://doi.org/10.20944/preprints202408.1288.v1*

*3. Darban, Z. Z., Las-Heras, I., G-Berdonces, M., Valero, M., Barambones, O. Deep Learning for Time Series Anomaly Detection: A Survey. ACM Computing Surveys, 57, 1–42, 2024. https://doi.org/10.48550/arXiv.2211.05244*

*4. Neloy, A. A., Turgeon, M. A Comprehensive Study of Auto-Encoders for Anomaly Detection: Efficiency and Trade-Offs. Machine Learning with Applications, 17, 100572, 2024. https://doi.org/10.1016/j.mlwa.2024.100572*

*5. Ji, T., et al. Variational Autoencoder Based Anomaly Detection in Real Operational Data. Energies, 18(11), 2770, 2025. https://doi.org/10.3390/en18112770*

*6. Korniszuk, K., Sawicki, B. Autoencoder-Based Anomaly Detection in Network Traffic. In: Proc. CPEE 2024 (International Conference on Compatibility, Power Electronics and Power Engineering), 2024. https://doi.org/10.1109/CPEE64152.2024.10720411*

*7. Souiden, I., Omri, M. N, Brahmi, Z. A Survey of Outlier Detection in High Dimensional Data Streams. Computer Science Review, 44, 100463, 2022. https://doi.org/10.1016/j.cosrev.2022.100463*

*8. Ginni, G. R., Chakravarthy, S. L. Efficient Outlier Detection in High-Dimensional Data Using Unsupervised Machine Learning. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15(4), 192-212, 2024. https://doi.org/10.58346/JOWUA.2024.I4.013*

*9. Li, Y., Yang, J., Wen, J. Entropy-Based Redundancy Analysis and Information Screening. Reliability Engineering & System Safety, 214, 107742, 2021. https://doi.org/10.1016/j.dcan.2021.12.001*

*10. Jeon, G. (Ed.). Information Entropy Algorithms for Image, Video, and Signal Processing. Special Issue of Entropy, 23(8), 926, 2021. https://doi.org/10.3390/e23080926*

*11. Mohammadi, H., Gupta, S., Sharma, S. A Large-Scale Performance Study of Entropy-Based Image Thresholding Techniques Using a New Sum of Absolute Value of Differences Metric. Pattern Analysis and Applications, 26, 2023. https://doi.org/10.1007/s10044-022-01121-z*

*12. Thyagharajan, K. K., Kalaiarasi, G. Image Near-Duplicate Detection Using Computer Vision Techniques: An Engineering Perspective. Archives of Computational Methods in Engineering, 28, 897–916, 2021. https://doi.org/10.1007/s11831-020-09400-w*

*13. Jakhar, Y., Borah, M. D. Effective Near-Duplicate Image Detection Using Perceptual Hashing, Siamese Network, and Vision Transformer. Information Processing & Management, 62(4), 2025. https://doi.org/10.1016/j.ipm.2025.104086*

*14. Li, K., Persaud, D., Choudhary, K,, DeCost, B., Greenwood, M. & Hattrick-Simpers, J. Exploiting Redundancy in Large Materials Datasets for Machine Learning. Nature Communications, 14, 7283, 2023. https://doi.org/10.1038/s41467-023-42992-y*

*15. V. Zivakin, O. Kozachuk, P. Prystavka, O. Cholyshkina, Training set AERIAL SURVEY for Data Recognition Systems From Aerial Surveillance Cameras // CEUR Workshop Proceedings. 2022. Vol. 3347. P. 246–255. [Online]. Available: https://ceur-ws.org/Vol-3347/Paper_21.pdf.*

*16. Sumit, S. S., Anavatti, S., Tahtali, M., Mirjalili, S., Turhan, U. ResNet-Lite: On Improving Image Classification with a Lightweight ResNet50 Approach, Procedia Computer Science, 2024. https://doi.org/10.1016/j.procs.2024.09.597*

*P. Prystavka, O. Cholyshkina, O. Podskrebko, M. Boryshkevich*

**STUDYING THE IMPACT OF ANOMALIES AND DUPLICATIONS IN AERIAL SURVEYING DATASETS ON THE QUALITY OF DEEP LEARNING**

*The quality of training datasets is a critical factor affecting the performance of deep learning models in automated aerial image analysis. The presence of anomalous and duplicate samples in*

*aerial imagery datasets leads to distortion of statistical properties, reduction of distribution entropy, and degradation of model generalization capability. This paper investigates the impact of such data imperfections on multi-class image classification performance and analyzes the effectiveness of statistical anomaly detection methods applied in a compact latent space.*

*A convolutional autoencoder is employed to generate low-dimensional latent representations of aerial images, providing an informative and noise-resistant space for subsequent statistical analysis. Anomalous samples are identified using the three-sigma rule, skewness and kurtosis-based analysis, and a multidimensional variation series approach. The effect of duplicate image removal is examined separately. Dataset quality is evaluated through entropy-based characteristics of the latent space and classification accuracy obtained using a ResNet50 convolutional neural network.*

*Experimental results demonstrate that removing anomalous samples has a positive effect on classification accuracy on the test dataset. Among the considered approaches, the three-sigma method proved to be the most effective, providing an accuracy improvement of up to 2.1% by eliminating samples that are highly distant from the distribution center and do not represent typical class characteristics. It is shown that data cleaning leads to an increase in latent space entropy, indicating higher information richness and a more uniform data distribution. This increase in entropy correlates with improved generalization performance of the classifier. The obtained results confirm the relevance of statistical analysis of latent representations as an effective stage in preparing aerial imagery datasets for deep learning applications.*

**Keywords:** aerial imagery, anomalous data, image duplicates, deep learning, convolutional neural networks, autoencoder, latent space, data entropy, image classification, generalization performance.

Надійшла до редакції: 29.12.2025
Прийнята до друку: 21.04.2026
Опубліковано: 27.04.2026