

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота
освітній ступінь – бакалавр

на тему: **«ОЙЛЕРОВІ ХАРАКТЕРИСТИКИ В АНАЛІЗІ
ДАНИХ»**

Виконав: студент 4-го року
навчання
освітньої програми «Прикладна
математика»,
спеціальності 113 Прикладна
математика

Рябченко Нікіта Іванович

Керівник: Крюкова Г. В.
доцент, кандидат фіз.-мат. наук

Рецензент:

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
(підпис)

«_____» _____ 20__р.

Київ - 2025

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

ЗАТВЕРДЖУЮ

Зав.кафедри математики,
доцент, кандидат фіз.-мат. наук

_____ Чорней Р.К.
(підпис)

“ _____ ” _____ 2025

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

для кваліфікаційної роботи
студенту 4-го курсу, факультету інформатики
Рябченко Нікіті Івановичу

Тема: «Ейлерові характеристики в аналізі даних»

Зміст кваліфікаційної роботи:

Анотація

1. Вступ

2. Основи ейлерової характеристики та топологічного аналізу

3. Використання ейлерових характеристичних кривих для аналізу даних

4. Програмна реалізація алгоритмів ейлерової характеристики

Висновки

Список літератури

Дата видачі “ _____ ” _____ 2025 Керівник _____
(підпис)

Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2024р.

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Отримання теми кваліфікаційної роботи.	17.09.2024			
2.	Ознайомлення з темою кваліфікаційної роботи.	24.09.2024			
3.	Розробка плану та структури роботи.	27.09.2024			
4.	Робота з науковою літературою, опис основних означень. Написання вступу та анотації.	7.10.2024			
5.	Дослідження методів топологічного аналізу даних	21.10.2024			
6.	Робота над текстовим оформленням теоретичної частини.	18.02.2025			
7.	Попередній аналіз кваліфікаційної роботи. Виправлення помилок.	04.04.2025			
8.	Попередній захист кваліфікаційної роботи.	14.05.2025			
9.	Захист кваліфікаційної роботи.	04.06.2025			

Науковий керівник _____
(ПІБ)

Виконавець кваліфікаційної роботи _____
(ПІБ)

Зміст

1	Вступ	5
2	Знайомство	7
2.1	Вступ до топологічного аналізу даних	7
2.2	Означення характеристики Ойлера	8
3	Топологічні дескриптори	9
3.1	Основні поняття	10
3.2	CW-комплекси та абстрактні симпліціальні комплекси	10
3.3	Комплекси Чеха та Вісторіса-Ріпса	12
4	Діаграми персистентності та інструменти ойлерової ха- рактеристики	15
4.1	Основні поняття гомології	16
4.2	Побудова персистентної гомології	17
4.3	Інструменти на основі ойлерової характеристики та їх по- рівняння з діаграмами персистентності	18
5	Стійкість	22
5.1	Стійкість діаграм персистентності	22
5.2	Стійкість кривої Ойлерової характеристики	23
5.3	Стійкість профілів Ойлерової характеристики	24
6	Реалізація алгоритму	25
6.1	Класифікація на даних ORBIT5K	25
6.2	Реалізація дескриптора профілю Ойлерової характери- сти	27
6.3	Реалізація дескриптора Перетворення Радона	30
6.4	Реалізація дескриптора Гібридне перетворення	33
6.5	Порівняння результатів	36

7 Висновки	39
Література	41
Додаток	43

Анотація

У цій кваліфікаційній роботі досліджується методи топологічного аналізу даних на основі ойлерової характеристики, яка є важливим інструментом для вивчення структури і властивостей даних. Основна увага приділена реалізації однопараметричних і багатопараметричних топологічних дескрипторів, таких як криві та профілі Ойлерової характеристики, перетворення Радона та гібридні згорткові перетворення .

У теоретичній частині подано основні поняття топологічного аналізу, а саме топологічним дескрипторам, їх стабільності. В практичній частині реалізовано алгоритми обчислення кожного з дескрипторів, створені візуалізації результатів, а також виконано порівняльний аналіз ефективності класифікації із застосуванням машинного навчання.

Ключові слова: топологічний аналіз, ойлерова характеристика, гібридні перетворення, компресія даних, багатопараметрична фільтрація.

1 Вступ

Зараз ми живемо у світі з неймовірно великими обсягами даних з абсолютно різними структурами, і з кожним роком ці обсяги будуть збільшуватися, а структура — ускладнюватися. Звичайні методи статистики та машинного навчання не завжди є ефективними у виявленні прихованих структур даних, у зв'язку з цим виникає нагальна потреба в нових підходах до аналізу інформації. Саме через це в останні роки стрімко набирає обертів новий напрям — топологічний аналіз даних.

TDA базується на алгебраїчній топології, а саме — на її гомологічних інваріантах. Дані інструменти дозволяють досліджувати структуру даних і аналізувати її, адже вона не змінюється при деформаціях, шумі чи часткових спотвореннях. Цей метод демонструє високу точність у задачах машинного навчання і є гідною альтернативою більш традиційним методам.

Метою даної роботи є дослідження та застосування топологічних дескрипторів у задачах класифікації. Розглядаються такі інструменти, як профілі Ойлерової характеристики, профілі Ойлерової характеристики, перетворення Радона та гібридні згорткові перетворення.

У роботі реалізовано побудову вищезгаданих топологічних перетворень, зроблено візуалізацію результатів, а також оцінено точність класифікації кожного методу. Також окрему увагу приділено питанню стійкості дескрипторів до шуму та будь-яких інших деформацій вхідних даних, оскільки ця властивість є критично важливою для практичного застосування в реальних задачах.

Актуальність даного дослідження визначається потребою в ефективних і, що найважливіше, обчислювально незатратних методах аналізу, здатних працювати з високовимірними та складно структурованими видами даних. Наприклад, використання ойлерової характеристики та її гібридних перетворень відкриває нові можливості дослідження топологічних сигналів у таких галузях, як машинне навчання, біоінформатика, медична діагностика, матеріалознавство тощо. Таким чином, моя робота

поєднує математичну основу і практичне застосування, показуючи, що TDA має великий потенціал як інструмент аналітики даних та машинного навчання.

2 Знайомство

Цей розділ присвячено ознайомленню з топологічним аналізом даних як окремою галуззю аналітики, а також розглянуто історію та базові поняття характеристики Ойлера.

2.1 Вступ до топологічного аналізу даних

Топологічний аналіз даних (скорочено TDA) — це доволі новий напрям у прикладній математиці. Він сформувався на початку XXI століття і є сумішню алгебраїчної топології та обчислювальної геометрії. Історія TDA бере свій початок з праць Едельсбруннера (2002) і Зомородіана і Карлссона (2005), які заснували напрям персистентної гомології. Також досить важливим дослідженням для розвитку TDA є робота Карлссона (2009), яке досить сильно підвищило популярність даного напрямку. [17]

Ідея топологічного аналізу даних базується на припущенні, що форма даних містить важливу інформацію про їхню структуру. У реальних задачах, коли йдеться про дані високої розмірності, часто спостерігається розрідженість та наявність значущих ознак нижчої розмірності. Одним із завдань топологічного аналізу є точне виявлення та опис таких властивостей. Наприклад, траєкторія простої моделі типу хижак–жертва, що описується рівняннями Лотки–Вольтерри, утворює замкнутий контур у фазовому просторі, і TDA надає інструменти для виявлення та кількісної оцінки подібних циклічних рухів. [18]

Протягом останніх років було створено кілька бібліотек, які реалізують ефективні алгоритми для побудови топологічних дескрипторів. Далі ми наведемо основні з них. Бібліотека GUDHI (на C++ і Python) є найбільш потужною і вживаною бібліотекою. Вона надає інструменти для створення та аналізу структурних особливостей даних, що мають топологічну природу. Також корисними є бібліотеки Dionysus, PHAT та DIPHA. Вони реалізують ефективні алгоритми для обчислення персистентної гомології. Слід згадати і бібліотеку Giotto-TDA, вона є дуже зручною для

сучасного застосування TDA. Ця Python-бібліотека дозволяє легко імплементувати топологічний аналіз у задачі машинного навчання. [17]

На сьогоднішній день сфера топологічного аналізу даних є доволі популярною галуззю. Вона вже має базовий набір математичних і програмних інструментів, що доволі сильно спрощує роботу. Більше того, ці методи часто люблять комбінувати з класичними методами машинного навчання. Така методика дозволяє покращувати якість аналізу складних даних та підвищувати стійкість моделей до шуму та інших деформацій даних.

2.2 Означення характеристики Ойлера

Наразі характеристика Ойлера є одним із ключових понять сучасної математичної топології. Вперше вона була сформульована в середині XVIII століття швейцарським математиком Леонардом Ойлером.[1] Саме тоді він опублікував формулу, яка пов'язує кількість вершин (V), ребер (E) і граней (F) у геометричній фігурі. Характеристику Ойлера прийнято позначати грецькою літерою χ (хі), і вона має наступний вигляд:

$$\chi = V - E + F$$

Більше того, будь-яка поверхня опуклого багатогранника має ойлерову характеристику:

$$\chi = V - E + F = 2$$

У XIX столітті математики Огюстен Коші та Анрі Пуанкаре почали розвивати ідеї, започатковані Ойлером. Вони узагальнили поняття характеристики Ойлера з опуклих багатогранників на довільні поверхні й з'ясували, що ця характеристика є *топологічним інваріантом* [2] [3]. Тобто при деформації фігури (згинанні, стисканні або розтягуванні) значення характеристики Ойлера залишається незмінним.[4]

У сучасній топології характеристика Ойлера використовується не лише з багатогранниками, а й з більш складними об'єктами, такими як

CW-комплекси і симпліціальні комплекси.[5] Для них ойлерова характеристика може бути визначена як знакозмінна сума:

$$\chi = k_0 - k_1 + k_2 - k_3 + \dots,$$

де k_i — кількість клітинок розмірності i .

Також характеристику Ойлера можна представити в більш загальному вигляді через *числа Бетті* [2] — топологічні інваріанти, які описують кількість незалежних циклів заданої розмірності в геометричному об'єкті. [6] Для многовиду розмірності 2 характеристика Ойлера визначається наступним чином:

$$\chi = \beta_0 - \beta_1,$$

де:

- β_0 — кількість зв'язних компонент (0-вимірні базиси),
- β_1 — кількість отворів у фігурі (1-вимірні базиси).

У загальному випадку ойлерова характеристика для топологічного простору через числа Бетті має вигляд [2]:

$$\chi = \beta_0 - \beta_1 + \beta_2 - \beta_3 + \dots$$

Як підсумок, на сьогоднішній день характеристика Ойлера є ключовою компонентою топології, входить до загальної теорії гомологій і має безліч застосувань у таких сферах, як наука, інженерія та комп'ютерна графіка. Ми ж будемо використовувати її для аналізу даних.

3 Топологічні дескриптори

У даному розділі ми розглянемо ключові поняття для побудови топологічних дескрипторів, які будуть застосовуватись далі в роботі.

3.1 Основні поняття

Перед використанням методів топологічного аналізу даних важливо ознайомитись із базовими математичними поняттями, що лежать в його основі. Основою теорії є загальна топологія, яка формалізує уявлення про неперервність, відкритість і компактність простору. [16]

Означення 1. Топологією на множині X називається колекція підмножин T , яка задовольняє такі умови:

1. $\emptyset \in T$ і $X \in T$;
2. довільна (можливо нескінченна) об'єднана кількість множин з T також належить T ;
3. скінченна кількість перетинів множин з T також належить T .

Пара (X, T) називається *топологічним простором*.

Означення 2. Підмножина $U \subseteq X$ є *відкритою*, якщо $U \in T$.

Означення 3. Підмножина $U \subseteq X$ є *замкненою*, якщо її доповнення $X \setminus U$ є відкритим.

Означення 4. *Покриттям* простору X називається набір підмножин $A = \{A_i\}$ таких, що $X = \bigcup_i A_i$. Якщо всі A_i є відкритими множинами, то таке покриття називається *відкритим покриттям*.

Означення 5. Топологічний простір X називається *компактним*, якщо для кожного відкритого покриття існує його скінченна підмножина, яка теж покриває X .

3.2 CW-комплекси та абстрактні симпліціальні комплекси

Означення 6. **CW-комплекси** — це вид топологічних просторів, який будується за допомогою з'єднання базових блоків, а саме — клітин різних

розмірностей. Спочатку створюється множина нульвимірних клітин (точок) X^0 . Наступним кроком формується n -скелет X^n шляхом приєднання n -вимірних клітин до вже створеного $(n - 1)$ -скелета X^{n-1} через їхні межі [5].

Процес побудови може бути завершений на певному скінченному етапі або продовжуватись нескінченно довго. Підкомплексом A клітинного комплексу X називають підмножину, яка разом із кожною своєю клітиною містить також усі клітини нижчої розмірності.

Однак, поняття CW-комплексу є доволі загальним, що дозволяє описувати різні класи топологічних просторів. Окремим важливим випадком слід виділити абстрактні симпліціальні комплекси. Вони є набагато зручнішими для дискретизації простору та проведення досліджень у ньому. Далі ми наводимо формальне означення даного комплексу.

Означення 7. Абстрактний симпліціальний комплекс — це скінченна множина K симплексів, де симплекс — це елемент $\sigma \in K$, що задовольняє умову: якщо симплекс σ належить K , тоді кожна його підмножина $\tau \subseteq \sigma$ також належить K [7].

Розмірність симплексів визначається за формулою:

$$\dim(\sigma) = |\sigma| - 1,$$

де $|\sigma|$ — кількість елементів у множині σ .

Наприклад:

- 0-симплекси — вершини,
- 1-симплекси — ребра,
- 2-симплекси — трикутники.

Ойлерова характеристика симпліціального комплексу K визначається як ціле число [7]:

$$\chi(K) = \sum_{\sigma \in K} (-1)^{\dim(\sigma)}.$$

У аналізі даних дуже важливо не лише вивчати структуру простору, а й розуміти, як різні параметри впливають на її зміну. Саме для цього вводиться поняття m -параметричної фільтрації.

Означення 8. m -параметричною фільтрацією скінченного симпліціального комплексу K називають сімейство підкомплексів $(F_t)_{t \in \mathbb{R}^m}$, таке, що для будь-яких двох параметрів $t, t' \in \mathbb{R}^m$ виконується наступна умова: якщо $t \leq t'$, то $F_t \subseteq F_{t'}$ [7].

Тобто, якщо за всіма координатами один параметр менший або дорівнює іншому, то відповідний підкомплекс також є підмножиною наступного. Іншими словами, у процесі фільтрації кількість доступних симплексів поступово зростає, що призводить до наповнення структури комплексу новими елементами.

Одним із класичних способів побудови фільтрації є використання субрівневих множин.

Приклад 1. Нехай $f : K \rightarrow \mathbb{R}^m$ — функція, яка не зменшується при переході від симплекса до його надмножини, тобто для будь-яких $\sigma \subseteq \tau$ виконується:

$$f(\sigma) \leq f(\tau).$$

Тоді для кожного значення параметра t будується підкомплекс:

$$F_f(t) = \{\sigma \in K : f(\sigma) \leq t\}.$$

Функцію f у цьому випадку називають фільтром фільтрації [7].

3.3 Комплекси Чеха та Вісторіса-Ріпса

Одним із найпоширеніших прикладів симпліціальних комплексів є комплекс Чеха. Комплекс Чеха — це абстрактний симпліціальний комплекс, побудований у метричному просторі за допомогою множини точок.[8] Завдяки ньому можна відобразити топологічну інформацію про множину точок або про їх розподіл. У нашому випадку він дозволяє відобра-

зити локальну близькість і структуру точок на основі симпліціального комплексу.

Означення 9. Нехай $X \subset \mathbb{R}^d$ — скінченна множина точок. Для кожного радіуса $r \geq 0$ будується комплекс Чеха $\mathcal{C}(X, r)$. У цьому комплексі симплекс утворюється з набору точок $\{x_0, \dots, x_k\}$, якщо відкриті кулі з радіусом r , що побудовані навколо цих точок, мають спільний перетин, який не є порожнім. [9] Тобто:

$$\bigcap_{i=0}^k B(x_i, r) \neq \emptyset,$$

де $B(x_i, r)$ — відкрита куля радіуса r з центром у точці x_i .

Інтуїтивно, комплекс Чеха формує симплекси для точок, які знаходяться досить близько одна до одної. Зі збільшенням радіуса r структура комплексу стає складнішою: з'являються нові ребра, трикутники та вищі симплекси.

Буває, що інформації про розташування точок недостатньо, і виникає необхідність враховувати додаткові властивості даних. Для таких випадків використовується функціональна фільтрація Чеха.

Означення 10. Нехай $X \subset \mathbb{R}^d$ — скінченна множина точок, а $f = (f_1, \dots, f_m) : X \rightarrow \mathbb{R}^m$ — задана функція, яка для кожної точки зіставляє набір числових характеристик (густина точки, вага тощо).

Тоді функціональна фільтрація Чеха — це побудова комплексу Чеха $\mathcal{C}(X, r)$ із додатковим обмеженням. Для параметрів $r \in \mathbb{R}$ і $t = (t_1, \dots, t_m) \in \mathbb{R}^m$ симплекси визначаються умовою:

$$\mathcal{C}(X, f)(r, t) = \left\{ \sigma \in \mathcal{C}(X, r) : \max_{x \in \sigma} f_i(x) \leq t_i \quad \forall i = 1, \dots, m \right\}.$$

У такий спосіб фільтрація одночасно враховує геометричну близькість точок та додаткові властивості кожної з них.[7]

Функціональні фільтрації дають змогу будувати більш гнучкі моделі топологічної структури даних, допомагаючи, зокрема, справлятися з викидами або виділяти області високої густини.

Також, як альтернатива комплексу Чеха існує комплекс Вієторіса–Ріпса. У топології комплекс Вієторіса–Ріпса (також відомий як комплекс Вієторіса або Rips-комплекс) є методом побудови симпліціального комплексу на основі лише метричної інформації, саме відстаней між точками у множині. Даний підхід дозволяє легко описувати топологічну структуру даних, навіть якщо сама геометрична форма множини є невідомою або недоступною.

Rips-комплекс задається параметром $\delta > 0$ та довільним метричним простором M . Симплекс утворюється для кожної скінченної підмножини точок $S \subset M$, у якій відстань між двома будь-якими точками не перевищує δ . Така підмножина входить до комплексу як симплекс розмірності $|S| - 1$ — тобто ребро для двох точок, трикутник для трьох, тетраедр для чотирьох тощо. [10]

Нижче подано формальні означення та властивості комплексу Вієторіса–Ріпса.

Означення 11. Флаговий комплекс — це абстрактний симпліціальний комплекс, у якому будь-яка підмножина вершин, що з'єднана 1-простими симплексами (ребрами), автоматично створює вищий симплекс. Іншими словами, комплекс визначається за допомогою своєї 1-скелетонної структури.

Будь-який симпліціальний комплекс, який задовольняє умови флагового комплексу, може бути повністю описаним графом суміжності своїх вершин.

Означення 12. Нехай X — множина точок у метричному просторі, а $\delta > 0$ — фіксований радіус. Тоді комплекс Вієторіса–Ріпса $VR_\delta(X)$ — це симпліціальний комплекс, вершини якого збігаються з елементами X , а кожна n -мерна грань (симплекс) відповідає $(n + 1)$ -точковій підмножині X , у якій усі попарні відстані менші або рівні δ .

Зазначимо, що побудова комплексу Вієторіса–Ріпса є менш обчислювально затратною, у порівнянні з комплексом Чеха, оскільки не вимагає

побудови перетинів куль, а залежить лише від порогового графа відстаней. Саме тому Rips-комплекси є широко вживаними в обчислювальній топології.

Більше того, такі комплекси можуть бути побудовані в довільному метричному просторі. Саме ця ознака робить їх універсальними для застосування в задачах з нестандартною геометрією (наприклад, простори ознак у машинному навчанні). [10]

На практиці, для спрощення обчислень, часто використовують функціональну альфа-фільтрацію, яка є ще більш обчислювально ефективною.

Означення 13. Альфа-комплекс — це симпліціальний комплекс, побудований на основі множини точок у метричному просторі. Альфа-комплекс є підкомплексом триангуляції Делоне. Він визначається шляхом вибору таких симплексів триангуляції Делоне, для яких описана сфера є порожньою та має квадрат радіуса, що не перевищує заданого порогу α^2 . Якщо описана сфера не є порожньою, то фільтраційне значення симплекса визначається як мінімальне фільтраційне значення серед його граней. [9]

Альфа-комплекс зберігає топологічну інформацію простору подібно до комплексу Чеха, проте є обчислювально ефективнішим та компактнішим.

4 Діаграми персистентності та інструменти ойлерової характеристики

У процесі топологічного аналізу даних важливо не лише виявляти топологічні характеристики простору, а й витягувати багатомасштабну топологічну інформацію з даних. Саме для вирішення таких задач використовується персистентна гомологія.

У топологічному аналізі даних персистентна гомологія є ключовим методом для отримання найважливішої топологічної інформації з даних.

Ця інформація зручно може бути представлена за допомогою діаграм персистентності, які лаконічно узагальнюють геометричні та топологічні властивості на масштабах різної розмірності.

У цьому розділі ми коротко ознайомимося з основними поняттями персистентної гомології і приділимо особливу увагу діаграмам персистентності.

4.1 Основні поняття гомології

Гомологія — це інструмент, що допомагає досліджувати топологічні простори за допомогою кількісного опису їхніх структур: компонент зв'язності, циклів тощо. За означенням, для топологічного простору X та цілого числа k ми позначаємо k -ту гомологічну групу як $H_k(X)$, а її розмірність — як $\beta_k(X) = \dim H_k(X)$, тобто k -те число Бетті. [11]

Також у персистентній гомології розглядаються неперервні функції $f : X \rightarrow \mathbb{R}$, які фільтрують простір через підрівневі множини наступного вигляду: $f^{-1}(-\infty, a]$, де a — це змінна, яка на пряму впливає на топологію цих множин і, відповідно, на їхні гомологічні групи. [12]

Для опису місць, де ці зміни відбуваються, вводиться таке поняття, як *гомологічно критичні значення*. Це такі значення $a \in \mathbb{R}$, на яких структура гомології підрівневих множин $f^{-1}(-\infty, a]$ змінюється: наприклад, з'являється новий цикл або зникає існуюча компонента зв'язності.

Крім того, важливим поняттям є *лема про критичне значення*. Вона стверджує, що якщо на деякому взятому інтервалі значень функції не трапляється жодного гомологічно критичного значення, то гомологічна структура не змінюється. Це дозволяє точно вказувати на ті рівні функції, де відбуваються топологічні переходи — саме така логіка лежить в основі побудови діаграм персистентності, які виявляють появу або зникнення гомологічних ознак різних розмірностей. [12]

Інтуїтивно, k -вимірна гомологічна ознака описує наявність отвору розмірності k у просторі. Наприклад:

- 0-вимірні ознаки — це компоненти зв'язності;

- 1-вимірні — це цикли або петлі;
- 2-вимірні — це порожнини, схожі на внутрішню частину сфери.

Кожна гомологічна ознака має свій власний момент *народження* та момент *смерті*. Моментом народження називається значення параметра, при якому ця ознака вперше з'являється, а моментом смерті — значення, при якому вона зникає. Пара чисел $[a, b)$ описує цей інтервал існування і називається *персистентним інтервалом*. [13]

Множина таких інтервалів утворює *штрих-код фільтрації*, який можна трактувати як життєвий цикл кожної топологічної ознаки.

4.2 Побудова персистентної гомології

Одним із способів побудови персистентної гомології є розгляд множини точок, які лежать у вигляді прихованої геометричної структури у просторі. Мета полягає в тому, щоб зробити висновки про топологію утвореної фігури, базуючись лише на скінченну вибірку точок.

Замість того, щоб фіксувати певний радіус r , навколо кожної точки створюються кулі зі змінюваними радіусами, і аналізується об'єднання цих куль. Для малих r кожна точка є окремим компонентом зв'язності, а для надто великих r — усе зливається в одну велику компоненту. [14]

Аби уникнути вибору одного фіксованого значення r , персистентна гомологія вивчає зміну топологічних властивостей у всьому діапазоні значень r , дозволяючи виділити лише ті особливості, які є стійкими до масштабування. Точки, що лежать далеко від діагоналі $x = y$ на діаграмі персистентності, позначають саме такі стабільні топологічні особливості, а ті, що лежать поблизу діагоналі, зазвичай інтерпретуються як шум.

На даному малюнку ми можемо спостерігати зміну топологічної структури простору під час збільшення радіуса t . Для малих значень радіуса точки утворюють окремі компоненти зв'язності, які пізніше починають об'єднуватись, а при достатньо великому t утворюються 1-вимірні петлі.

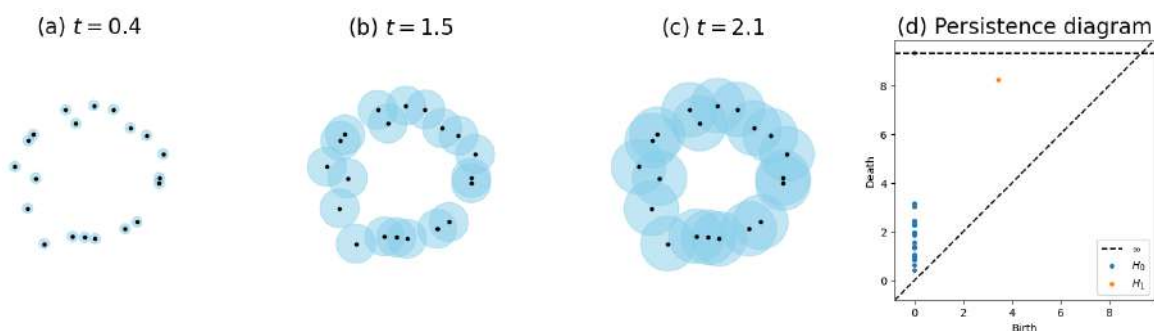


Рис. 1: Візуалізація персистентної гомології на основі вибірки точок. (а–с) — об’єднання куль з центрами у вибірці точок для трьох різних радіусів $t = 0.4$, $t = 1.5$, $t = 2.1$. (d) — відповідна діаграма персистентності.

Остання візуалізація (d) відображає діаграму персистентності, де сині точки відповідають компонентам зв’язності (H_0), а помаранчева — стійкому 1-вимірному циклу (H_1). Чим далі точка від діагоналі, тим стійкішою і тим важливішою вона вважається.

4.3 Інструменти на основі ойлерової характеристики та їх порівняння з діаграмами персистентності

Цей підрозділ має реферативний характер і викладає теоретичні засади з роботи [7]

Ойлерова характеристика є одним з базових компонентів топологічного аналізу даних. Вона допомагає описувати складність топологічних структур у фільтрованих симпліціальних комплексах. Особливо актуальним їхнє застосування стає у випадках багатопараметричних фільтрацій, де нам важлива не стільки структура на конкретному кроці фільтрації, скільки динаміка її змін в просторі з усіх параметрів.

Перед переходом до більш складних конструкцій, треба сформулювати основне означення профілю ойлерової характеристики:

Означення 14. *Профіль ойлерової характеристики* (скорочено ЕСР) фільтрованого комплексу K над множиною параметрів P — це функція,

яка кожному значенню $p \in P$ ставить у відповідність ойлерову характеристику відповідного підкомплексу K_p [15]:

$$\text{ECP}(K, p) = \chi(K_p).$$

У випадку багатопараметричної фільтрації $\mathcal{F}$ з m параметрами, можна ввести таку функцію, що буде описувати зміну ойлерової характеристики в залежності від значення параметра $t \in \mathbb{R}^m$:

$$\chi_{\mathcal{F}} : t \in \mathbb{R}^m \mapsto \chi(\mathcal{F}_t).$$

Така функція називається кривою ойлерової характеристики, якщо $m = 1$, і поверхнею ойлерової характеристики, у випадку коли $m = 2$. [15]

Однак робота з багатьма параметрами значно ускладнює роботу. Щоб спростити її, часто використовують *pushforward*-операцію, яка зводить багатовимірну фільтрацію до одновимірної:

Означення 15. Нехай $\mathcal{F}$ — однокритична m -параметрична фільтрація, а $\xi \in \mathbb{R}_+^{m*}$ — фіксований напрям у параметричному просторі. Тоді *pushforward* фільтрації $\mathcal{F}$ вздовж ξ визначається як однопараметричне сімейство:

$$(\xi_* \mathcal{F})_s = \bigcup_{\xi \cdot t \leq s} \mathcal{F}_t, \quad s \in \mathbb{R}.$$

Функція, що відображає ойлерову характеристику цієї фільтрації, називається *проштовзнutoю кривою ойлерової характеристики* та позначається як:

$$\xi_* \chi_{\mathcal{F}} = \chi_{\xi_* \mathcal{F}}.$$

Цю конструкцію можна застосовувати до фільтрацій, які задаються функціями.

Лема 1. Нехай $f : \mathcal{K} \rightarrow \mathbb{R}^m$ — невізростаюча функція, і нехай $\xi \in \mathbb{R}_+^{m*}$. Тоді фільтрація, породжена композицією $\xi \circ f$, задовольняє:

$$\xi_* \mathcal{F}_f = \mathcal{F}_{\xi \circ f}, \quad \text{та} \quad \xi_* \chi_f = \chi_{\xi \circ f}.$$

Для покращення інтерпретації кривої ойлерової характеристики в напрямку ξ застосовують *гібридне згорткове перетворення*, що поєднує топологічні інваріанти з аналітичними засобами згладження.

Означення 16. Нехай $\mathcal{F}$ — однокритична m -параметрична фільтрація, а $\kappa \in L^1(\mathbb{R})$ — функція ядра. Тоді гібридне перетворення ойлерової характеристики $\chi_{\mathcal{F}}$ у напрямку $\xi \in \mathbb{R}_+^{m*}$ визначається як:

$$\psi_{\mathcal{F}}^{\kappa}(\xi) = \int_{\mathbb{R}} \kappa(s) \cdot \xi_* \chi_{\mathcal{F}}(s) \, ds.$$

Звісно існує формула для ойлерової характеристики у випадку так званих *однокритичних фільтрацій*. Вона записується через характеристичні функції супермножин параметричного простору:

Означення 17. Припустимо, що кожному симплексу $\sigma \in \mathcal{K}$ відповідає параметр $t(\sigma) \in \mathbb{R}^m$ такий, що $\sigma \in \mathcal{F}_{t(\sigma)}$. Тоді ойлерова характеристика фільтрації має вигляд:

$$\chi_{\mathcal{F}} = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \mathbf{1}_{Q_{t(\sigma)}}, \quad (2.1)$$

де $Q_u := \{t \in \mathbb{R}^m : t \geq u\}$ — супермножина у параметричному просторі.

Якщо $\kappa \in L^1(\mathbb{R})$ — ядро згортки, а $\bar{\kappa}$ — його первісна функція, що прямує до нуля при $+\infty$, то гібридне перетворення можна записати як:

$$\psi_{\mathcal{F}}^{\kappa}(\xi) = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \bar{\kappa}(\xi \cdot t(\sigma)). \quad (2.2)$$

Ці формули дозволяють ефективно застосовувати ойлерову характеристику для аналізу складної топологічної інформації в багатопараметричних фільтраціях.

Для випадків з однопараметричною фільтрацією спеціально існує зв'язок між діаграмами персистентності та інструментами ойлера, такими як криві ойлерової характеристики та гібридні перетворення. Їх зв'язок можна подати за допомогою наступної формули:

$$\chi_{\mathcal{F}}(t) = \sum_{k \geq 0} \sum_{i=1}^{n_k} (-1)^k \cdot \mathbf{1}_{[a_i^k, b_i^k)}(t),$$

де (a_i^k, b_i^k) — інтервали діаграми персистентності k -го степеня. Можна зазначити, що ойлерова характеристика є знакоперемінною сумою всіх ознак, і не залежить від їхньої індивідуальної тривалості.

Аналогічно, формулу для гібридних перетворень можна записати як:

$$\psi_{\mathcal{F}}^{\kappa}(\xi) = \sum_{k \geq 0} \sum_{i=1}^{n_k} (-1)^k (\overline{\kappa}(\xi \cdot b_i^k) - \overline{\kappa}(\xi \cdot a_i^k)) ,$$

що знову спрощує багатовимірну інформацію до скалярної ознаки.

Перехід від повної діаграми персистентності до ойлерової кривої чи гібридного перетворення призводить до втрати деякої інформації:

- Зникає розподіл за гомологічними степенями, оскільки сума виконується по всіх k .
- Зникає зв'язок між моментом народження та моментом смерті гомологічних ознак, оскільки у кривій враховуються лише набори усіх a_i та b_i , але не пари (a_i, b_i) .
- Криві більше не здатні розрізняти об'єкти з однаковою кількістю коротких та довгих життєвих циклів, адже діаграми $\{(0, 1)\}$ і $\{(0, 1/2), (1/2, 1)\}$ мають однакову криву.

Незважаючи на втрату точності, ойлерові інструменти мають свої переваги :

- Вони є легкообчислювальними і не потребують повної побудови гомологій.
- На відміну від діаграм персистентності, вони легше узагальнюються на багатопараметричних фільтраціях.
- Завдяки стислому числовому представленню вони показують велику ефективність у задачах машинного навчання

Отже, незважаючи на те, що діаграми персистентності є потужнішим джерелом топологічної інформації, досить часто у прикладних задачах

використовуються саме ойлерові інструменти. Вони є більш вживаними завдяки своїй швидкості роботи, інтерпретованості та здатності узагальнення.

5 Стійкість

Одним із ключових аспектів при застосуванні топологічних дескрипторів у задачах машинного навчання є їхня стійкість, тобто здатність зберігати інформативність за наявності шуму або інших спотворень у вхідних даних. У цьому розділі ми перевіримо найбільш уживані дескриптори (діаграми персистентності, профілі ойлерової характеристики та криву ойлерової характеристики) на стабільність.

5.1 Стійкість діаграм персистентності

Процес формування діаграм персистентності побудований на фільтраціях, під час яких нові симплекси послідовно приєднуються до комплексу. Дуже важливо, що їхня основна структура залишається незмінною при геометричних змінах вхідної хмари точок. Візуально це проявляється у збереженні головних патернів діаграми (діагональних кластерів та їх відносності до діагоналі). Це підтверджує те, що навіть за наявності змін у вхідних даних діаграми персистентності є досить надійними дескрипторами.

Теоретично це можна обґрунтувати за допомогою властивості, яка забезпечує стабільність і формалізується у вигляді нерівності:

$$d_B(D(f), D(g)) \leq \|f - g\|_\infty,$$

де $D(f)$ і $D(g)$ — діаграми персистентності двох фільтрацій, а d_B — відстань Боттлнека. Це означає, що при незначаній вплив на вхідну функцію (у нашому випадку — координат точок хмари) призводить лише до несуттєвих змін у відповідній діаграмі. [12]

З геометричної точки зору, діаграма персистентності кодує інформацію про народження і смерті топологічних структур. Кожна віддалена від діагоналі точка, вказує на значущу особливість, а точки, що є близькими до діагоналі — на шум або неважливість структури. Таким чином, стійкість гарантує, що розташування основних (далеких від діагоналі) точок залишається стабільним навіть у присутності зашумлених даних.

5.2 Стійкість кривої Ойлерової характеристики

Крива Ойлерової характеристики (скорочено ЕСС) є однією з найважливіших однопараметричних топологічних дескрипторів. Вона відображає зміну топологічної складності фільтрованого симпліціального комплексу під час фільтрації. На відміну від діаграм персистентності, ЕСС об'єднує топологічну інформацію з усіх розмірностей і представляє її як алгебраїчну суму чисел Бетті:

$$\text{ЕСС}(K, t) = \sum_{k=0}^n (-1)^k \beta_k(K_t),$$

де K_t — підкомплекс симпліціального комплексу K , який містить всі симплекси з фільтраційним значенням, меншим або рівним t , а $\beta_k(K_t)$ — число k -вимірних гомологічних класів у K_t . [15]

Означення 18. Нехай K_1 та K_2 — два фільтровані симпліціальні комплекси. Тоді відстань між відповідними ЕСС можна визначити через норму L^1 :

$$\|\text{ЕСС}(K_1, t) - \text{ЕСС}(K_2, t)\|_1 = \int_{\mathbb{R}} |\text{ЕСС}(K_1, t) - \text{ЕСС}(K_2, t)| dt.$$

Ця формула чисельно відображає розбіжність між топологічною еволюцією двох різних фільтрацій. Експериментальні дослідження демонструють що, ця відстань залишається малою навіть при значних пертурбаціях вхідних точок, що вказує на високу стійкість ЕСС. [15]

Стійкість кривої Ойлерової характеристики безпосередньо впливає зі стабільності кривих Бетті, які є складовими ЕСС. Для кожної розмірності k можна оцінити відстань між відповідними кривими Бетті двох

фільтрацій через Васерштейн-відстань між їхніми діаграмами персистентності.

Нехай X і Y — два фільтровані симпліціальні комплекси, побудовані на основі схожих вхідних даних. Тоді з використанням лінійності та нерівності трикутника одержується наступна оцінка для відстані між кривими Ойлерової характеристики:

$$\|\text{ECC}(X, t) - \text{ECC}(Y, t)\|_1 \leq \sum_k 2W_1(Dgm_k(X), Dgm_k(Y)),$$

де W_1 позначає першу Васерштейн-відстань між відповідними k -вимірними діаграмами персистентності.

Цей результат підтверджує, що ЕСС є стабільним дескриптором: невеликі зміни у вхідних даних (наприклад, шум або деформації) не призводять до суттєвих змін у формі чи значеннях кривої. [15]

5.3 Стійкість профілів Ойлерової характеристики

Профілі Ойлерової характеристики є узагальненням кривих Ойлерової характеристики і використовуються у випадках мультифільтрацій. Завдяки ним з'являється можливість аналізувати зміну топології за наявності декількох фільтраційних параметрів, наприклад, координат або щільностей. Застосування таких профілів у задачах класифікації є особливо перспективним у контексті двовимірних структур, як-от зображення або траєкторії в просторі.

Означення 19. Нехай K_1 і K_2 — два мультифільтровані симпліціальні комплекси, фільтровані за n -вимірним параметром $v \in \mathbb{R}^n$. Тоді відстань L^1 між відповідними ЕСП визначається наступною формулою:

$$\|\text{ЕСР}(K_1, v) - \text{ЕСР}(K_2, v)\|_1 = \int_{\mathbb{R}^n} |\text{ЕСР}(K_1, v) - \text{ЕСР}(K_2, v)| dv.$$

Ця метрика є узагальнюючим поняттям звичайної L^1 -відстані між ЕСС на багатовимірний випадок, що дозволяє кількісно зіставляти мультифільтровані об'єкти. [15]

Для практичного використання ЕСР зазвичай обмежують діапазон фільтрації до скінченного інтервалу $[0, f_\infty]$ в кожному напрямку. Таке обмеження гарантує, що обчислення будуть скінченними і дозволяє отримати адекватні оцінки стійкості.

Означення 20. Нехай K — скінченний симпліціальний комплекс з n -вимірною мультифільтрацією $F : K \rightarrow \mathbb{R}^n$. Припустимо, що збурення фільтраційних значень усіх клітин не перевищує ε в l_∞ -нормі, і обчислення проводиться на обрізаному домені $[0, f_\infty]^n$. Тоді:

$$\|\text{ЕСР}(K, v) - \text{ЕСР}(K_\varepsilon, v)\|_1 \leq |K| \cdot n \cdot \varepsilon^{n-1} \cdot f_\infty,$$

де $|K|$ — кількість клітин в комплексі.

Цей вираз чітко оцінює стійкість ЕСР щодо малого збурення фільтрації. Об'єм зміненої області можна оцінити сумою n n -вимірних паралелепіпедів з основою ε^{n-1} і висотою f_∞ . [15]

Отримані результати також підтверджують, що ЕСР є стабільним дескриптором у практичних задачах аналізу даних. За умови контролюваних відхилень фільтраційних параметрів, зміну профілю можна з легкістю передбачувати та обмежувати.

6 Реалізація алгоритму

6.1 Класифікація на даних ORBIT5K

Набір даних **ORBIT5K** є класичним бенчмарком у топологічному аналізі даних і широко використовується для задач класифікації топологічно неоднорідних траєкторій. Кожен об'єкт у цьому наборі представляє собою траєкторію довжини $N = 1000$ в межах одиничного квадрата $[0, 1]^2$, згенеровану за допомогою детермінованої динамічної системи:

$$\begin{aligned} x_{n+1} &= (x_n + \rho y_n (1 - y_n)) \bmod 1, \\ y_{n+1} &= (y_n + \rho x_{n+1} (1 - x_{n+1})) \bmod 1, \end{aligned}$$

де параметр $\rho \in \{2.5, 3.5, 4.0, 4.1, 4.3\}$ контролює динамічну поведінку. Початкові умови (x_0, y_0) вибираються рівномірно випадково в $[0, 1]^2$. [7]

Для кожного з п'яти класів (ρ) було згенеровано по 1000 траєкторій. Ми розбили дані на тренувальну вибірку (70%) та тестову (30%). Кожну траєкторію розглядали як точкову хмару, до якої застосовувалась **альфа-фільтрація**. На основі цієї фільтрації ми обчислювали топологічні дескриптори у вигляді ознак фіксованого розміру 30×30 (тобто 900 ознак на об'єкт):

- **Профілі ейлерової характеристики** — двопараметричне узагальнення ЕСС;
- **Перетворення радона** (скорочено RT) — згортка фільтраційної функції вздовж напрямку;
- **Гібридне перетворення** (скорочено HT) — згладжена проєкція топологічного сигналу з використанням ядра.

Для кожного з дескрипторів ми навчали класифікатор **XGBoost** з оптимізацією гіперпараметрів за допомогою крос-валідації. Всі моделі були протестовані на однаковій тестовій вибірці, що дозволяє коректно порівнювати їхню якість.

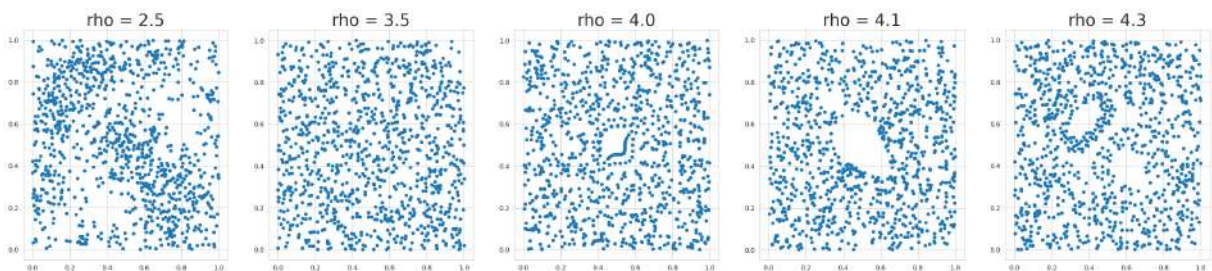


Рис. 2: Приклади траєкторій з ORBIT5K для різних значень параметра ρ : (a) $\rho = 2.5$, (b) $\rho = 3.5$, (c) $\rho = 4.0$, (d) $\rho = 4.1$, (e) $\rho = 4.3$.

Цей набір є складним з точки зору класифікації, оскільки траєкторії побудовані за єдиною формулою, а різниця між класами полягає виклю-

чно в зміні топологічної складності траєкторій. Це робить ORBIT5K ідеальним середовищем для тестування топологічних дескрипторів. У наступних підрозділах представлено результати реалізації кожного методу з детальним аналізом візуалізацій та метрик точності.

6.2 Реалізація дескриптора профілю Ойлерової характеристики

Профіль Ойлерової характеристики— це двовимірне розширення кривої Ойлерової характеристики, яке фіксує зміну топологічної структури при варіаціях двох параметрів фільтрації. У рамках нашої роботи ми реалізували обчислення ЕСР для кожної траєкторії з набору даних ORBIT5K, використовуючи функціональну фільтрацію.

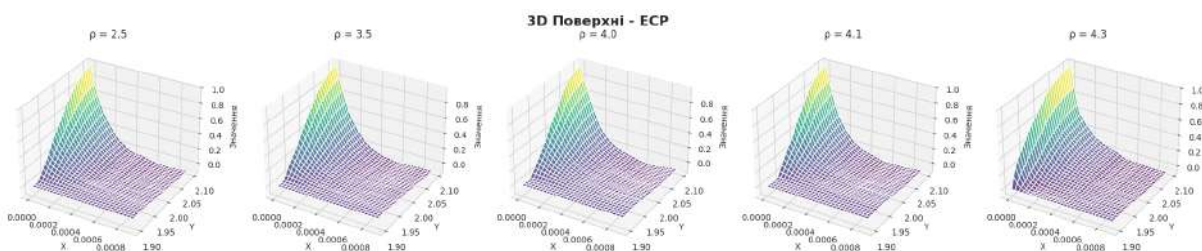


Рис. 3: 3D-поверхні дескриптора ЕСР для п'яти класів траєкторій з різними значеннями параметра ρ . По осях відкладено параметри фільтрації, по вертикалі — значення χ .

На рис. 3 подано тривимірні поверхні значень ЕСР, де помітно, що зростання параметра ρ змінює характер топологічних особливостей. Зокрема, максимальні значення χ спостерігаються на початку фільтрації, а далі спостерігається поступове спадання. Такі властивості свідчать про наявність складної внутрішньої структури у траєкторіях, що добре відображається через дескриптор.

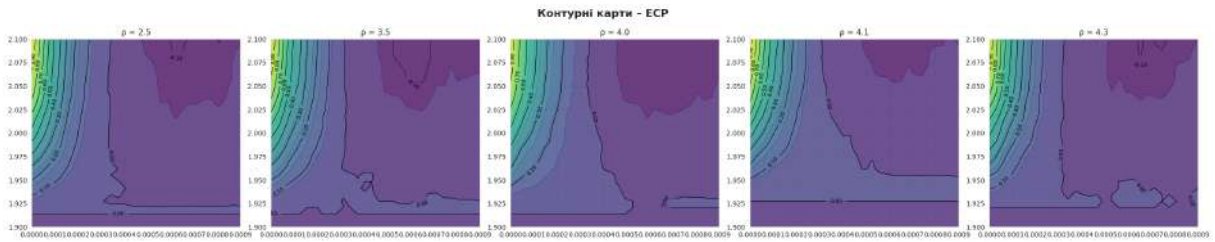


Рис. 4: Контурні карти для дескрипторів ЕСР, що відповідають значенням $\rho = 2.5, 3.5, 4.0, 4.1, 4.3$. Яскравість та рівні позначають інтенсивність ойлерової характеристики.

Контурні карти на рис. 4 деталізують інформацію про поведінку дескриптора в кожній точці фільтраційного простору. Вони демонструють чітке розмежування зон з високими та низькими значеннями χ , що є важливим для подальшої класифікації.

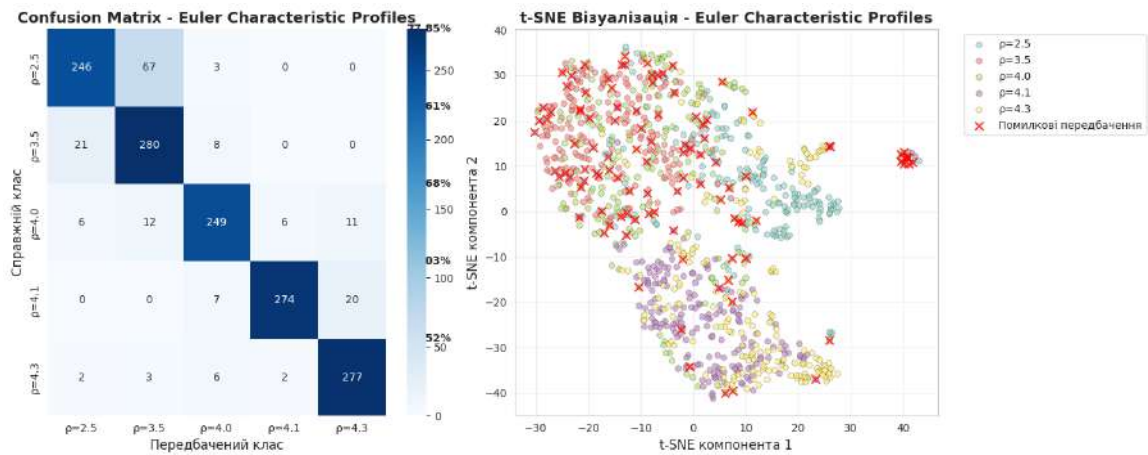


Рис. 5: Ліва частина: матриця неточностей для класифікації з використанням ЕСР (точність 88.4%). Права частина: t-SNE візуалізація ознак ЕСР, де червоним позначено помилкові передбачення.

На рис. 5 зліва показано матрицю неточностей класифікатора, який працює з ознаками ЕСР. Найкраще розпізнаються класи з великим ρ , тоді як найбільше помилок спостерігається між близькими значеннями $\rho = 2.5$ і $\rho = 3.5$. t-SNE візуалізація праворуч підтверджує кластеризацію даних за класами, а червоні мітки ілюструють приклади помилкових класифікацій.

Табл. 1: Детальний звіт класифікації для дескриптора ЕСР.

Клас	Precision	Recall	F1-score	Кількість
$\rho = 2.5$	0.89	0.78	0.83	316
$\rho = 3.5$	0.77	0.91	0.83	309
$\rho = 4.0$	0.91	0.88	0.89	284
$\rho = 4.1$	0.97	0.91	0.94	301
$\rho = 4.3$	0.90	0.96	0.93	290
Середнє (macro)	0.89	0.89	0.89	—
Зважене (weighted)	0.89	0.88	0.88	1500

Таблиця 1 підтверджує, що модель з ЕСР-дескрипторами демонструє високу якість розпізнавання у всіх класах, з найвищими значеннями F1-метрики при $\rho = 4.1$ та $\rho = 4.3$.

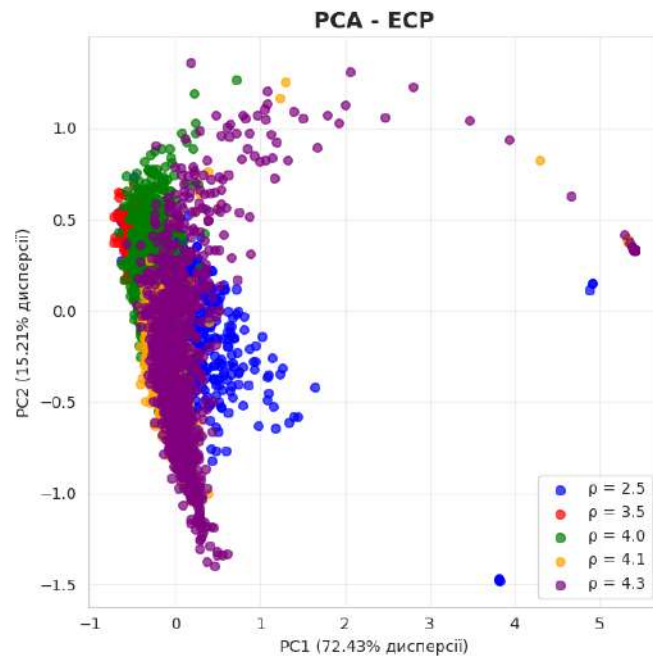


Рис. 6: PCA-візуалізація ЕСР-ознак. Добре помітна роздільність класів у проєкції на дві головні компоненти.

На завершення, PCA-візуалізація (рис. 6) демонструє, що більшість варіації в даних пояснюється першою головною компонентою, а також

чітко виявляються скупчення точок, що відповідають кожному з класів. Це підтверджує, що ECP-дескриптор ефективно захоплює релевантну інформацію для класифікації.

Таким чином, використання профілю Ойлерової характеристики є обґрунтованим і ефективним підходом у задачах аналізу траєкторій.

6.3 Реалізація дескриптора Перетворення Радона

Radon-перетворення є аналітичним інструментом, який інтегрує функцію вздовж гіперплощин і широко застосовується в задачах комп'ютерної томографії та аналізу зображень. У контексті топологічного аналізу даних ми використовуємо RT як дескриптор, що відображає інтегральну інформацію про структуру фільтрованих топологічних сигналів.

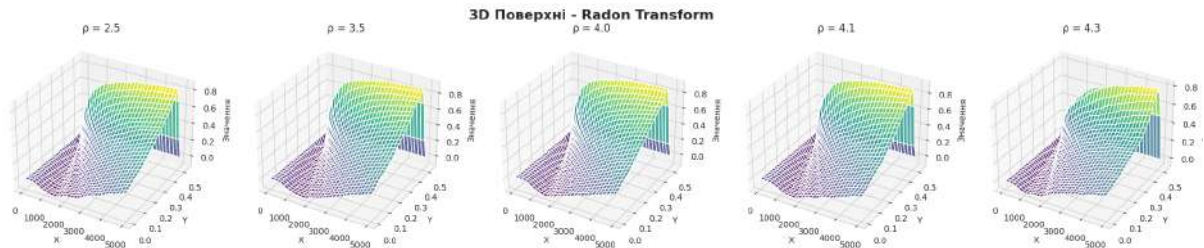


Рис. 7: Radon-перетворення у вигляді 3D поверхонь для різних ρ .

На рис. 7 подано тривимірні поверхні RT-дескрипторів. Можна спостерігати подібність загальної форми поверхонь у різних класах, однак різниця в локальних перепадах значень є достатньо вираженою для використання у задачах класифікації. Висота поверхні вказує на інтегровану характеристику фільтрованого сигналу, яка відображає локальні топологічні особливості.

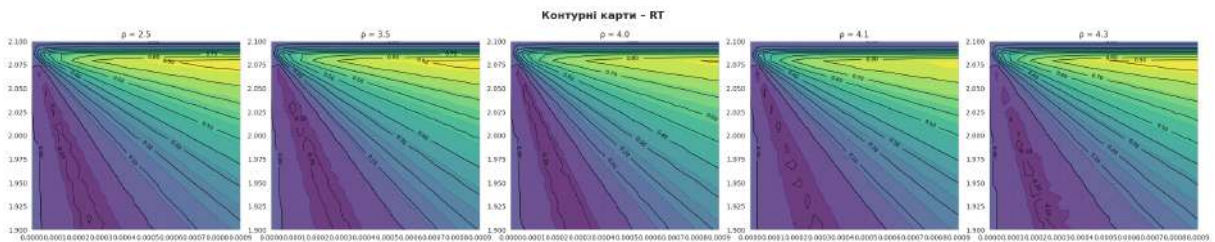


Рис. 8: Контурні карти RT-дескрипторів.

Контурні карти на рис. 8 демонструють зміни інтегральної топологічної інформації у просторі параметрів фільтрації. Для кожного класу можна побачити унікальні контурні шаблони, які відображають різну геометрію розподілу ознак. Це свідчить про наявність дискримінативних властивостей у дескриптора RT.

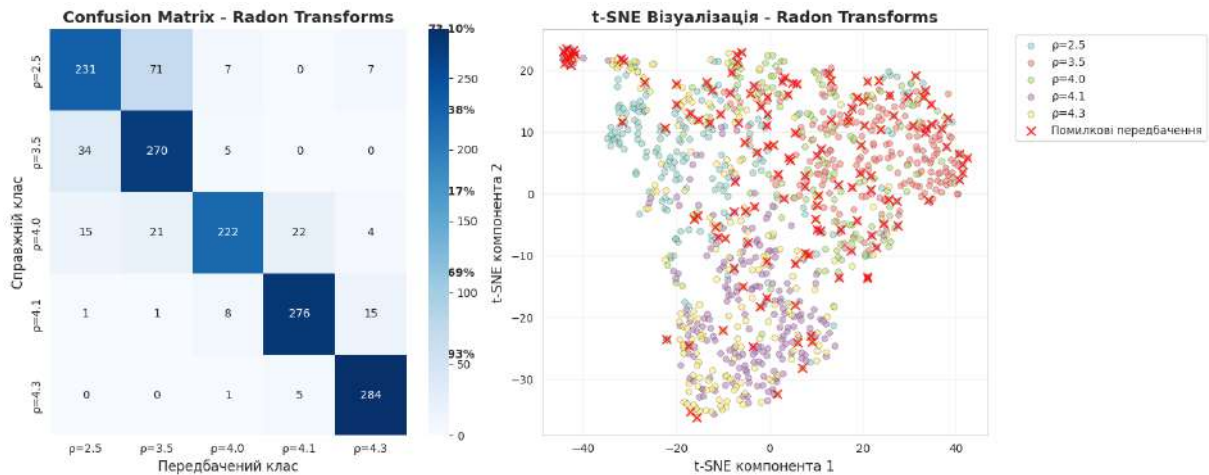


Рис. 9: Ліва частина: матриця змішування. Права частина: t-SNE проєкція Radon-перетворень.

На рис. 9 представлено результати t-SNE візуалізації простору ознак RT (праворуч) та відповідну матрицю неточностей (ліворуч). Видно, що класи розподілені досить компактно у 2D-просторі. Помилкові передбачення (червоні хрестики) в основному припадають на межі між класами $\rho = 2.5$ та $\rho = 3.5$, що також видно з помилок класифікації у матриці змішування.

Табл. 2: Детальний звіт класифікації для дескриптора Radon Transform.

Клас	Precision	Recall	F1-score	Кількість
$\rho = 2.5$	0.82	0.73	0.77	316
$\rho = 3.5$	0.74	0.87	0.80	309
$\rho = 4.0$	0.91	0.78	0.84	284
$\rho = 4.1$	0.91	0.92	0.91	301
$\rho = 4.3$	0.92	0.98	0.95	290
Середнє (macro)	0.86	0.86	0.86	—
Зважене (weighted)	0.86	0.86	0.85	1500

Звіт класифікації в таблиці 2 демонструє, що найвищу точність (F1-score) було досягнуто для класів з вищим значенням ρ , тоді як найнижчі результати спостерігалися для класу $\rho = 2.5$. Проте загальна точність залишається високою і становить понад 85%.

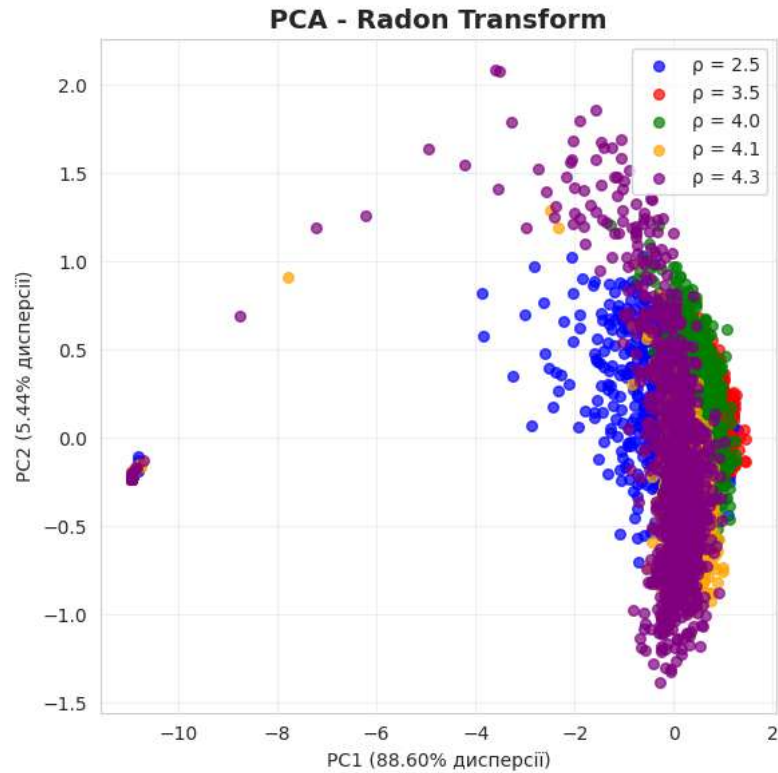


Рис. 10: PCA-візуалізація RT-дескрипторів.

Нарешті, на рис. 10 наведено PCA-візуалізацію, яка показує, що перша головна компонента пояснює понад 89% дисперсії. Кожен клас утворює щільне скупчення, що ще раз підкреслює здатність RT до ефективного відображення топологічної структури даних.

Класифікаційні результати. Загальна точність моделі на основі дескриптора Radon Transform становить **85.53%**, що підтверджує його ефективність у задачах мультикласової класифікації.

Radon-дескриптор показав високу здатність до узагальнення та ефективну класифікацію складних топологічних структур, що підтверджується як візуалізаціями, так і числовими метриками.

6.4 Реалізація дескриптора Гібридне перетворення

Гібридне згорткове перетворення поєднує топологічні інваріанти з методами згладжування, що дозволяє отримати дескриптори, стійкі до шуму, а також здатні виявляти тонку структуру даних. У рамках нашої реалізації ми застосували гібридну фільтрацію до багатовимірного простору параметрів і проєктували її вздовж обраного напрямку ξ . Результати згорткового перетворення були візуалізовані у вигляді тривимірних поверхонь та контурних карт.

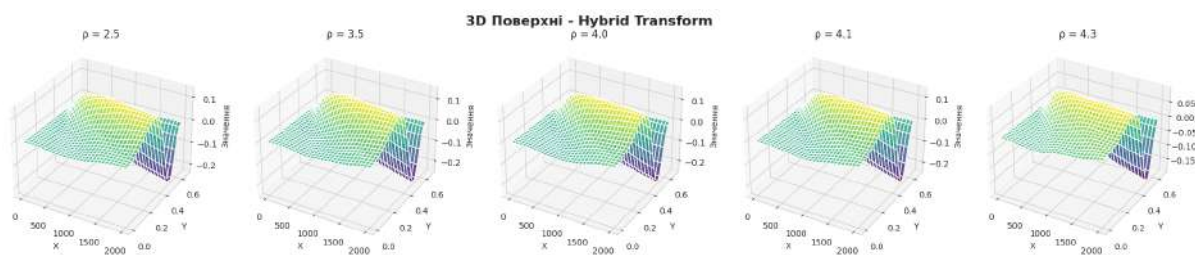


Рис. 11: 3D поверхні гібридного перетворення для різних значень ρ .

На рис. 11 показано, що 3D-поверхні мають чітко виражені структурні особливості, які відрізняють класи один від одного. Зокрема, видно, що при зростанні ρ спостерігається стабілізація амплітуди й зміщення

максимумів дескрипторів, що свідчить про залежність від топологічної складності траєкторій.

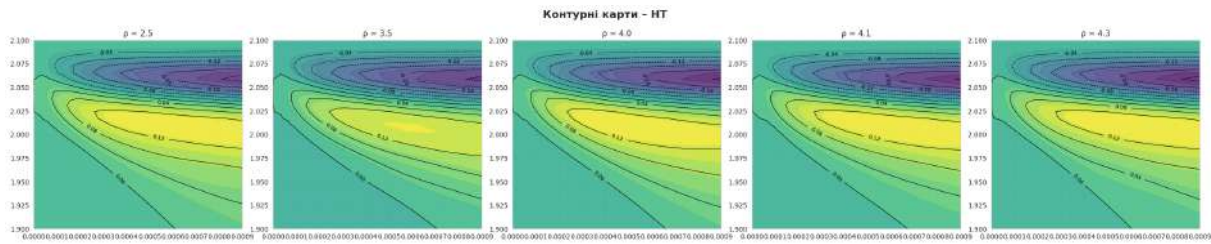


Рис. 12: Контурні карти гібридного перетворення для кожного класу.

Контурні карти на рис. 12 дозволяють легко виявляти області з однаковими значеннями дескрипторів, що може бути корисним при виявленні топологічно схожих сегментів даних. Чіткі контури підтверджують стабільність гібридного перетворення для кожного класу.

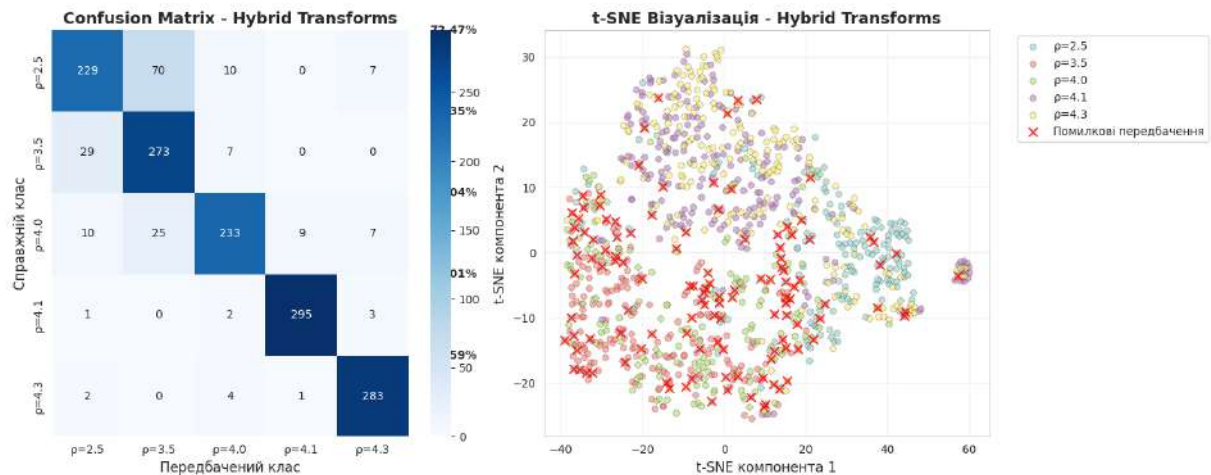


Рис. 13: Матриця помилок і t-SNE візуалізація для НТ.

Рис. 13 демонструє результати класифікації. Зліва — матриця неточностей, з якої видно, що найбільше помилок припадає на класи $\rho = 2.5$ та $\rho = 3.5$, які є найближчими топологічно. Праворуч — t-SNE візуалізація, що демонструє доволі чіткі кластери, незважаючи на часткове перекриття між класами.

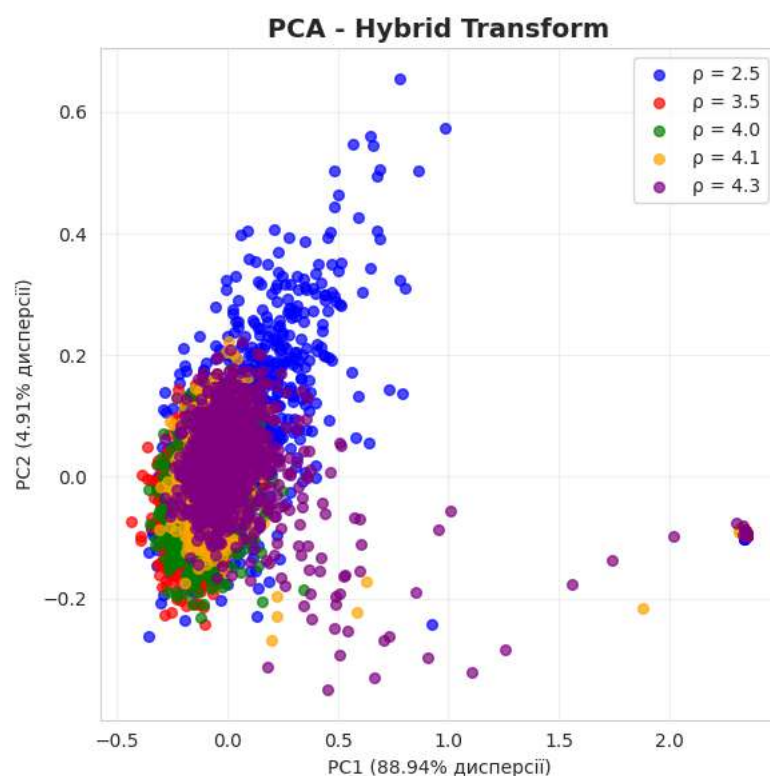


Рис. 14: PCA візуалізація НТ-дескрипторів.

РСА-проекція на рис. 14 підтверджує наявність основного напрямку варіації, який пояснює понад 88% дисперсії. Хоча кластери частково перекриваються, більшість класів мають окремі компактні області в головному просторі.

Табл. 3: Детальний звіт класифікації для дескриптора НТ.

Клас	Precision	Recall	F1-score	Кількість
$\rho = 2.5$	0.85	0.72	0.78	316
$\rho = 3.5$	0.74	0.88	0.81	309
$\rho = 4.0$	0.91	0.82	0.86	284
$\rho = 4.1$	0.97	0.98	0.97	301
$\rho = 4.3$	0.94	0.98	0.96	290
Середнє (macro)	0.88	0.88	0.88	—
Зважене (weighted)	0.88	0.88	0.87	1500

Загальна точність класифікації склала **87.53%**, що є одним із найкращих результатів серед усіх досліджених методів. Гібридні дескриптори забезпечують хороший баланс між стійкістю до шуму та здатністю розрізняти складні класи.

6.5 Порівняння результатів

У цьому розділі наведено порівняльний аналіз ефективності трьох підходів до побудови топологічних дескрипторів: профілів Ойлерової характеристики, перетворень Радона та гібридних трансформацій. Для оцінювання використовувалась точність класифікації на основі моделі XGBoost, а також аналіз поведінки класифікаторів при зміні параметрів.

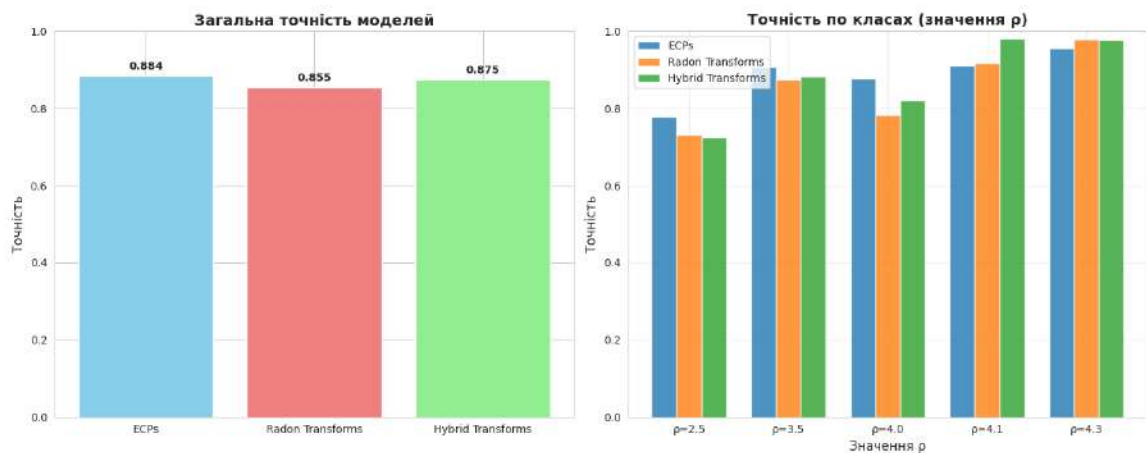


Рис. 15: Порівняння загальної точності та точності по класах

Як видно з рисунка 15, найвищу середню точність показав дескриптор ECP (88.4%), тоді як гібридне перетворення НТ досягло 87.5%, а RT — 85.5%. Попри загальну ефективність усіх підходів, ECP стабільно демонструє кращі результати в усіх класах параметра ρ .

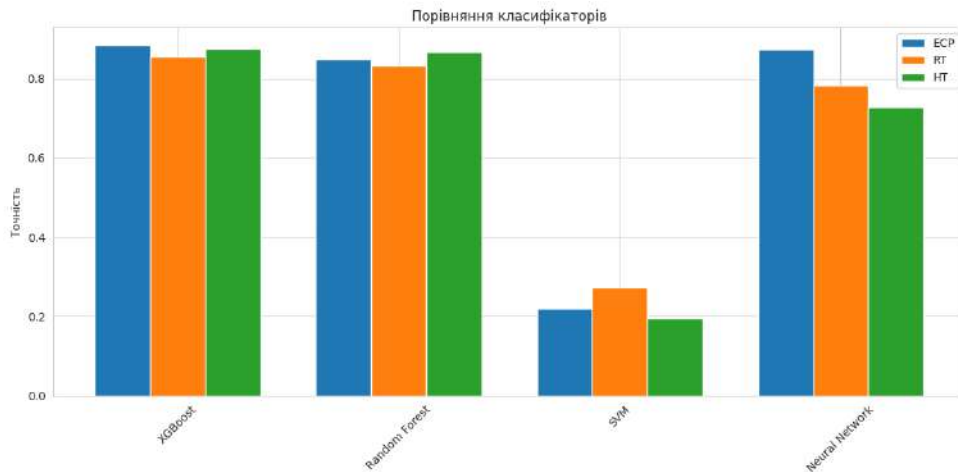


Рис. 16: Точність різних класифікаторів для кожного типу дескриптора

На рисунку 16 представлено порівняння точності чотирьох популярних моделей: XGBoost, Random Forest, SVM та нейронної мережі. Найвищу ефективність в усіх випадках показали моделі XGBoost та Random Forest, тоді як SVM значно поступається, особливо для ЕСР та НТ.

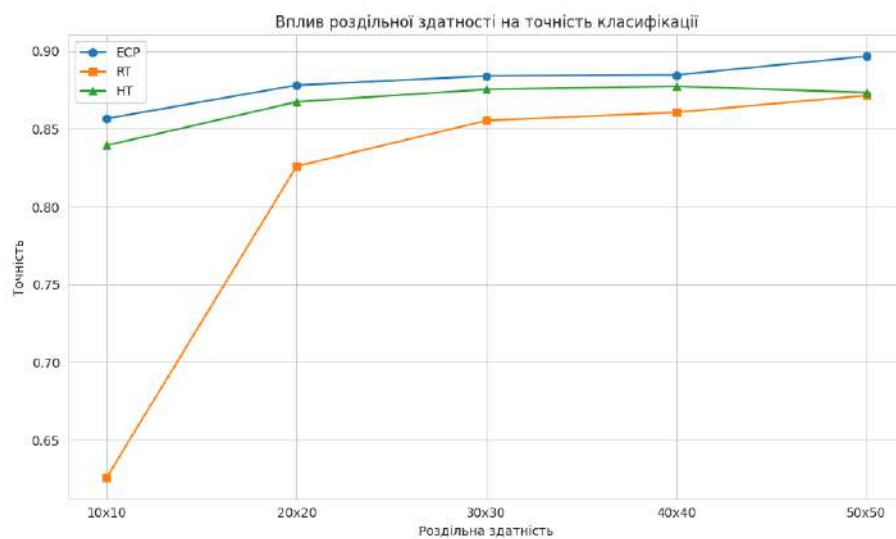


Рис. 17: Вплив роздільної здатності дескрипторів на точність класифікації

Оцінка впливу роздільної здатності, представлена на рисунку 17, показує, що збільшення роздільної здатності покращує точність для всіх трьох методів. Особливо чутливим до цього є РТ, який демонструє значне зростання точності при переході з 10×10 до 20×20 .

Таким чином, експериментальні результати свідчать про високу ефективність профілів Ойлерової характеристики, особливо при використанні моделей бустингу, а також демонструють практичну доцільність гібридних перетворень для задач класифікації.

7 Висновки

У цій роботі ми дослідили топологічний аналіз даних як інструмент для створення інформативних і стійких ознак у задачах класифікації. Теоретичною основою даного аналізу є алгебраїчна топологія, зокрема гомологічні інваріанти, які дозволяють досліджувати структурні особливості простору. Особливу увагу ми приділили характеристиці Ойлера та її похідним (кривій Ойлерової характеристики, профілю Ойлерової характеристики та гібридним згортковим перетворенням). Завдяки теоретичному аналізу ми з'ясували, що крива Ойлерової характеристики є стійким дескриптором, адже її поведінка залишається незмінною навіть при деформації вхідних даних.

У практичній частині роботи було реалізовано та протестовано три методи побудови дескрипторів: профіль Ойлерової характеристики, перетворення Радона та гібридне перетворення. Усі методи було проведено на синтетичному наборі траєкторій ORBIT5K. Для кожного з методів було побудовано тривимірні поверхні та контурні карти дескрипторів, проведено аналіз роздільної здатності ознак за допомогою методів PCA і t-SNE, а також виконано класифікацію із застосуванням моделей XGBoost, Random Forest, SVM і нейронних мереж. Всі класифікації було виконано з оптимізацією гіперпараметрів.

Результати нашого експерименту показали, що дескриптор профілю Ойлерової характеристики є найточнішим для задач класифікації — 88.4%, демонструючи найкращу здатність до розмежування класів. Гібридне перетворення показало дуже близький, втім все ж дещо нижчий результат — 87.53%. Незважаючи на це, цей дескриптор виявив найвищу стійкість до шуму. Перетворення Радона також досягло досить гарної точності — 85.53%, і продемонструвало чітку кластеризацію у візуалізаціях ознак.

Крім того, було досліджено вплив параметра ρ на точність класифікації. Встановлено, що найбільше помилок спостерігається для класів з меншим значенням ρ . Вірогідно це відбувається через їхню менш ви-

ражену топологічну структуру. Також було досліджено вплив роздільної здатності на дескриптори. Ми виявили, що збільшення роздільної здатності покращує результати лише до певного порогу, після чого їхня точність стабілізується.

Як результат, топологічні дескриптори продемонстрували дуже високу ефективність у задачах класифікації складних структур. Застосування топологічного аналізу даних дозволяє будувати стійкі та інформативні ознаки, які здатні вловлювати як локальні, так і глобальні топологічні особливості даних. У майбутньому існує перспектива досліджень щодо розширення до багатопараметричних фільтрацій, побудови інтегрованих гібридних моделей та їхнє поєднання з методами глибокого навчання.

Література

- [1] Richeson D. S. *Euler's Gem: The Polyhedron Formula and the Birth of Topology*. Princeton University Press, 2008.
- [2] https://en.wikipedia.org/wiki/Euler_characteristic
- [3] Albin, Pierre (2019). *History of algebraic topology*. Archived at Ghostarchive and the Wayback Machine. YouTube.
- [4] *Invariant Definition (Illustrated Mathematics Dictionary)*.
- [5] Lundell, A. T.; Weingram, S. *The topology of CW complexes*. Van Nostrand University Series in Higher Mathematics, 1970.
- [6] Barile, Margherita and Weisstein, Eric. *Betti number*. From MathWorld—A Wolfram Web Resource.
- [7] Hacquard, O.; Lebovici, V. *Euler Characteristic Tools for Topological Data Analysis*. Journal of Machine Learning Research, 25 (2024), 1–39.
- [8] Ghrist, Robert W. *Elementary applied topology*. (1st ed.), 2014. ISBN 9781502880857.
- [9] Edelsbrunner H., Harer J. *Computational Topology: An Introduction*. American Mathematical Society, 2010.
- [10] Lecha Sánchez, M. *Persistent Homology: Functional Summaries of Persistence Diagrams for Time Series Analysis*. Treball final de grau, Universitat de Barcelona, Facultat de Matemàtiques i Informàtica, 2021.
- [11] Munkres, J. R. *Elements of Algebraic Topology*. Addison-Wesley, Redwood City, California, 1984.
- [12] Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. *Stability of Persistence Diagrams*. Discrete and Computational Geometry, 37(1):103–120, 2007.

- [13] Fasy, B. T.; Lecci, F.; Rinaldo, A.; Wasserman, L.; Balakrishnan, S.; Singh, A. *Confidence Sets for Persistence Diagrams*. The Annals of Statistics, 42(6):2331–2369, 2014. DOI: 10.1214/14-AOS1252.
- [14] Kusano, G.; Fukumizu, K.; Hiraoka, Y. *Kernel Method for Persistence Diagrams via Kernel Embedding and Weight Factor*. Journal of Machine Learning Research, 18(189):1–41, 2018.
- [15] Dłotko, P.; Gurnari, D. *Euler Characteristic Curves and Profiles: A Stable Shape Invariant for Big Data Problems*. arXiv preprint arXiv:2308.06296, 2023.
- [16] Liu, A. *Introduction to Topological Data Analysis*. Unpublished manuscript, 2021.
- [17] Fasy, B. T.; Lecci, F.; Rinaldo, A.; Wasserman, L.; Balakrishnan, S.; Singh, A. *Confidence Sets for Persistence Diagrams*. The Annals of Statistics, 42(6):2301–2339, 2014. DOI: 10.1214/14-AOS1252.
- [18] Vin de Silva, Primoz Skraba, Mikael Vejdemo-Johansson. *Topological Analysis of Recurrent Systems*. Department of Mathematics, Pomona College; Jozef Stefan Institute; University of St Andrews.

Додаток

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 from mpl_toolkits.mplot3d import Axes3D
4 import seaborn as sns
5 from sklearn.model_selection import train_test_split
6 from sklearn.decomposition import PCA
7 from sklearn.manifold import TSNE
8 from sklearn.metrics import confusion_matrix,
   classification_report
9 from sklearn.ensemble import RandomForestClassifier
10 from sklearn.svm import SVC
11 from sklearn.neural_network import MLPClassifier
12 from xgboost import XGBClassifier
13 from ripser import ripser
14 from persim import plot_diagrams
15 import time
16 from tqdm import tqdm
17
18 from eulearning.datasets import gen_orbit5K
19 from eulearning.transforms import EulerCharacteristicProfile,
   RadonTransform, HybridTransform
20 from eulearning.filtrations import codensity
21 from eulearning.utils import vectorize_st
22 import gudhi as gd
23
24 rhos = [2.5, 3.5, 4.0, 4.1, 4.3]
25 n_pt_clds_per_class = 1000
26 n_classes = len(rhos)
27 n_pt_clds = n_classes * n_pt_clds_per_class
28
29
30 X, y = gen_orbit5K(rhos=rhos, size_each=n_pt_clds_per_class)
31
32
33 vec_sts = []
```

```

34 for i in tqdm(range(n_pt_clds), desc="Vectorizing point
    clouds"):
35     ac = gd.AlphaComplex(X[i])
36     st = ac.create_simplex_tree()
37     X_ = np.array([ac.get_point(j) for j in range(st.
        num_vertices())])
38     codensity_filt = codensity(X_)
39     vec_st = vectorize_st(st, filtrations=[codensity_filt])
40     vec_sts.append(vec_st)
41
42
43 euler_profile = EulerCharacteristicProfile(resolution=(30,30)
    , val_ranges=[(0, 0.0009), (1.9, 2.1)], pt_cld=True)
44 ecps = euler_profile.fit_transform(vec_sts)
45
46 radon_transform = RadonTransform(resolution=(30,30),
    val_ranges=[(0, 5000), (0, 0.53)], pt_cld=True)
47 rdns = radon_transform.fit_transform(vec_sts)
48
49 hybrid_transform = HybridTransform(resolution=(30,30),
    val_ranges=[(0, 2000), (0, 0.65)], kernel_name='wavelet_8'
    , pt_cld=True)
50 hts = hybrid_transform.fit_transform(vec_sts)
51
52
53 from ecps_visualizations import run_all_ecp_visualizations
54 run_all_ecp_visualizations(ecps, rhos, euler_profile)
55
56
57 random_state = 42
58
59 ecps_train, ecps_test, y_train_ecps, y_test_ecps =
    train_test_split(ecps, y, test_size=0.3, random_state=
        random_state)
60 rdns_train, rdns_test, y_train_rdns, y_test_rdns =
    train_test_split(rdns, y, test_size=0.3, random_state=
        random_state)

```

```

61 hts_train, hts_test, y_train_hts, y_test_hts =
    train_test_split(hts, y, test_size=0.3, random_state=
        random_state)
62
63
64 clf_ecps = XGBClassifier(random_state=random_state).fit(
    ecps_train, y_train_ecps)
65 clf_rdns = XGBClassifier(random_state=random_state).fit(
    rdns_train, y_train_rdns)
66 clf_hts = XGBClassifier(random_state=random_state).fit(
    hts_train, y_train_hts)
67
68
69 ecps_pred = clf_ecps.predict(ecps_test)
70 rdns_pred = clf_rdns.predict(rdns_test)
71 hts_pred = clf_hts.predict(hts_test)
72
73
74 ecps_score = clf_ecps.score(ecps_test, y_test_ecps)
75 rdns_score = clf_rdns.score(rdns_test, y_test_rdns)
76 hts_score = clf_hts.score(hts_test, y_test_hts)
77
78 print("\n===
    ===")
79 print(f"ECPs:    {ecps_score:.3f} ({ecps_score*100:.1f}%)")
80 print(f"RTs:    {rdns_score:.3f} ({rdns_score*100:.1f}%)")
81 print(f"HTs:    {hts_score:.3f} ({hts_score*100:.1f}%)")
82
83 def compare_classifiers(ecps, rdns, hts, y):
84     classifiers = {
85         'XGBoost': XGBClassifier(n_estimators=100,
            random_state=42),
86         'Random Forest': RandomForestClassifier(n_estimators
            =100, random_state=42),
87         'SVM': SVC(gamma='auto', random_state=42),
88         'Neural Network': MLPClassifier(hidden_layer_sizes
            =(100,), max_iter=1000, random_state=42)
89     }

```

```

90
91     results = {'ECP': {}, 'RT': {}, 'HT': {}}
92
93     ecps_train, ecps_test, y_train, y_test = train_test_split
94         (ecps, y, test_size=0.3, random_state=42)
95     rdns_train, rdns_test, y_train_r, y_test_r =
96         train_test_split(rdns, y, test_size=0.3, random_state
97             =42)
98     hts_train, hts_test, y_train_h, y_test_h =
99         train_test_split(hts, y, test_size=0.3, random_state
100             =42)
101
102     for name, clf in tqdm(classifiers.items(), desc="Testing
103         classifiers"):
104         clf.fit(ecps_train, y_train)
105         results['ECP'][name] = {'score': clf.score(ecps_test,
106             y_test)}
107
108         clf.fit(rdns_train, y_train_r)
109         results['RT'][name] = {'score': clf.score(rdns_test,
110             y_test_r)}
111
112         clf.fit(hts_train, y_train_h)
113         results['HT'][name] = {'score': clf.score(hts_test,
114             y_test_h)}
115
116     return results
117
118 classifier_results = compare_classifiers(ecps, rdns, hts, y)
119
120 def parameter_sensitivity_analysis(vec_sts, y):
121     resolutions = [(10, 10), (20, 20), (30, 30), (40, 40),
122         (50, 50)]
123     results = {'ECP': [], 'RT': [], 'HT': []}
124
125     for res in tqdm(resolutions, desc="Resolution testing"):

```

```

117     ecp = EulerCharacteristicProfile(resolution=res,
118         val_ranges=[(0, 0.0009), (1.9, 2.1)], pt_cld=True)
119     ecps = ecp.fit_transform(vec_sts)
120     X_train, X_test, y_train, y_test = train_test_split(
121         ecps, y, test_size=0.3, random_state=42)
122     clf = XGBClassifier().fit(X_train, y_train)
123     results['ECP'].append(clf.score(X_test, y_test))
124
125     rt = RadonTransform(resolution=res, val_ranges=[(0,
126         5000), (0, 0.53)], pt_cld=True)
127     rdns = rt.fit_transform(vec_sts)
128     X_train, X_test, y_train, y_test = train_test_split(
129         rdns, y, test_size=0.3, random_state=42)
130     clf = XGBClassifier().fit(X_train, y_train)
131     results['RT'].append(clf.score(X_test, y_test))
132
133     ht = HybridTransform(resolution=res, val_ranges=[(0,
134         2000), (0, 0.65)], kernel_name='wavelet_8', pt_cld
135         =True)
136     hts = ht.fit_transform(vec_sts)
137     X_train, X_test, y_train, y_test = train_test_split(
138         hts, y, test_size=0.3, random_state=42)
139     clf = XGBClassifier().fit(X_train, y_train)
140     results['HT'].append(clf.score(X_test, y_test))
141
142     plt.figure(figsize=(10,6))
143     labels = [f"{r[0]}x{r[1]}" for r in resolutions]
144     for key, val in results.items():
145         plt.plot(labels, val, marker='o', label=key)
146
147     plt.title('
148
149
150     plt.xlabel('
151
152     plt.ylabel('
153
154     plt.grid(True)
155     plt.legend()
156     plt.tight_layout()
157     plt.show()

```

```
147
148     return results
149
150 resolution_results = parameter_sensitivity_analysis(vec_sts,
151     y)
152 print("\n
```

,

.

")

48