

Міністерство освіти і науки України
Харківський національний університет імені В.Н. Каразіна
Фізичний факультет

Кафедра астрономії та космічної інформатики

«Допущено до захисту»
Зав. кафедри астрономії та
космічної інформатики
проф. Юрій ШКУРАТОВ

Оцінка « відмінно »

Голова ЕК
проф. Віктор ТИШКОВЕЦЬ

17.05.2023 р.



24.05.2024 р.



Сорока Кирило Олександрович

Ідентифікація та класифікація АЯГ на основі даних місій
Gaia та WISE

Кваліфікаційна робота на здобуття
освітньо-кваліфікаційного рівня
«Магістр» спеціальність
104 – «фізика та астрономія»
освітньо-професійна програма
«астрономія»

(підпис студента)

Науковий керівник – пров. науков.спів.,
канд.фіз-мат. наук
В. С. Ахметов

(підпис керівника)

Харків 2024

Анотація

Сорока К.О. Ідентифікація та класифікація АЯГ на основі даних місій Gaia та WISE. – Рукопис

Дипломна робота на здобуття освітньо-кваліфікаційного рівня “Магістр” за спеціальністю 104 – «фізика та астрономія». – Харківський національний університет імені В. Н. Каразіна, Харків, 2024. – 53 с. і – Іл. 18

Робота присвячена створенню класифікаційної моделі на основі методів машинного навчання з використанням даних сучасних фотометричних оглядів неба в оптичному (Gaia) та ближньому інфрачервоному діапазоні довжин хвиль (WISE). В якості навчальної вибірки був використаний каталог позагалактичних об'єктів, створений з використанням спектроскопічних даних каталогу SDSS DR17. Проведено детальний аналіз точності та повноти класифікаційної моделі. Середня точність класифікації для АЯГ становить 69%, для квазарів 98% та для зореутворюючих галактик 78%.

Ключові слова: АЯГ, квазари, галактики, методи машинного навчання, класифікація даних.

Abstract

Soroka K.O. Identification and classification of AGN based on data from the Gaia and WISE missions. – Manuscript

Diploma work to obtain a higher qualification level “Master” for specialty 104 – “Physics and astronomy”. – V. N. Karazin Kharkiv National University, Kharkiv, 2024. – 53 p. and – Img. 18

The work is devoted to the creation of a classification model based on machine learning methods using data from modern photometric surveys of the sky in the optical (Gaia) and near-infrared wavelength range (WISE). The catalogue of extragalactic objects created using spectroscopic data from the SDSS DR17 catalogue was used to determine the quality of the training sample. A detailed analysis of the accuracy and completeness of the classification model was performed. The average classification accuracy is 69% for AGNs, 98% for quasars, and 78% for star-forming galaxies.

Keywords: AGN, quasars, galaxies, SFG, machine learning methods, data classification.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ.....	5
ВСТУП.....	7
1 Огляд літератури	
1.1 Фізичні особливості АЯГ	9
1.2 Методи класифікації АЯГ	10
1.3 Методи машинного навчання для класифікації АЯГ	14
2 Огляд вхідних даних	
2.1 Вибір спектроскопічних даних - навчальна вибірка.....	17
2.2 Вибір фотометричних каталогів - досліджувальна вибірка.....	18
2.3 Попередній аналіз і обробка даних.....	19
3 Методика класифікації та регресії	
3.1 Моделі класифікації.....	22
3.1.1 НМ - нейрона мережа.....	22
3.1.2 Алгоритм випадковий ліс.....	22
3.2 Методи оцінки якості роботи моделі.....	23
3.2.1 Метрики ефективності класифікаційної моделі.....	24
3.2.2 Метрики якості прогнозування червоного зміщення.....	25
3.3 Інтерпретації роботи класифікаційних моделей.....	26
4 Результати	
4.1 Класифікація даних.....	30
4.2 Оцінки червоних зміщень.....	35
Висновки.....	38
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	39
ДОДАТКИ.....	46

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

ESA – (англ. European Space Agency) Європейське космічне агентство

SNR – (англ. Signal-to-noise ratio) відношення потужності корисного сигналу до потужності шуму.

QSO – (англ. Quasi-stellar object) Квасари

AGN – (англ. Active galactic nuclei) Активне ядро галактики (АЯГ)

SBG – (англ. starburst galaxy) Галактики зі спалахом зореутворення (СЗУГ).

SFG – (англ. starforming galaxy) Галактика з активним зореутворенням (ЗУГ).

GAL – (англ. galaxy) Звичайні галактики

НМЧД – (англ. Supermassive Black Hole, SMBH) Надмасивна чорна діра

ІЧ – (англ. Infrared, IR) інфрачервоний діапазон довжин хвиль

БІЧ – (англ. Near Infrared, NIR) ближній інфрачервоний діапазон

СІЧ – (англ. Mid Infrared, MIR) Середній інфрачервоний діапазон довжин хвиль

МН – (англ. Machine learning, ML) Машинне навчання

ВЛ – (англ. Random Forest, RF) випадковий ліс

SVM – (англ. Support Vector Machine) метод опорних векторів

k-NN – (англ. k-nearest neighbors) Метод k-найближчих сусідів

t-SNE – (англ. t-distributed stochastic neighbor embedding) Т-розподілене вкладення стохастичної близькості

SHAP – (англ. SHapley Additive exPlanations) теоретико-ігровий підхід для пояснення результатів роботи будь-якої моделі машинного навчання.

KFold – (англ. K-Fold) алгоритм кросвалідації

TP – (англ. True Positive) істинно позитивне

FP – (англ. False Positive) хибно позитивне

TN – (англ. True Negative) істинно негативне.

FN – (англ. False Negative) хибно негативне.

Precision – Влучність – частка релевантних зразків серед знайдених.

Recall – Повнота – частка загального числа позитивних зразків, яку було дійсно знайдено.

bAcc – (англ. Balanced Accuracy) Збалансована точність.

ROC – (англ. Receiver Operating Characteristic) робоча характеристика приймача.

AUC – (англ. Area Under Curve) площа обмежена кривою.

TPR – (англ. True Positive Rate)

FPR – (англ. False Positive Rate)

NMAD – (англ. Normalized Median Absolute Deviation) Нормоване медіанне абсолютне відхилення

ВСТУП

Великі фотометричні огляди неба дають можливість вивчати позагалактичну та галактичну астрофізику, наприклад скласти карту розподілу маси гало темної матерії галактик, ідентифікувати велику кількість молодих зоряних об'єктів, шукати квазари з великим червоним зміщенням та гравітаційно лінзовані квазари тощо.

Фотометрична інформація в апіорі є обмеженою в спектральному діапазоні довжин хвиль і, зазвичай, спектроскопічні дані надають більше можливостей для аналізу галактичних та позагалактичних джерел. З іншого боку, через технічні обмеження, сьогодні не вдається побудувати багатомільйонні каталоги спектрів, тому необхідність використання фотометричної інформації очевидна.

Актуальність роботи полягає у встановленні взаємозв'язку фотометричних даних позагалактичних об'єктів з їх спектроскопічними властивостями, що дозволить виконати пошук та класифікацію нових позагалактичних об'єктів у сучасних фотометричних оглядах неба.

Мета дослідження: класифікація об'єктів за фотометричними даними та визначення червоного зміщення.

Об'єкт дослідження: квазари, активні ядра галактик (АЯГ), галактики зі спалахом зореутворення (СЗУГ), з активним зореутворенням (ЗУГ), звичайні галактики та зорі.

Предмет дослідження: фотометричні та спектроскопічні дані, червоні зміщення та показники класифікації об'єктів.

Для досягнення поставленої цілі було виконано наступні завдання:

1. Проведено аналіз літератури по методам класифікації об'єктів з використанням спектроскопічних та фотометричних даних.
2. Створено навчальну вибірку на основі спектроскопічного каталогу SDSS DR17 та доповнено фотометричними даними з каталогів CatWISE та GaiaDR3.

3. З використанням методів машинного навчання створено класифікаційну модель, що визначає приналежність об'єкта до класу використовуючи його спостережувальні фотометричні дані.

4. Створено модель, що обчислює оцінку значень червоних зміщень за фотометричними даними.

5. Проведено аналіз та інтерпретація отриманих результатів класифікації та виконано порівняння оцінок червоних зміщень.

6. Зроблено висновки про застосування методів машинного навчання для класифікації об'єктів і сформульовано напрям подальших досліджень.

Наукова цінність отриманих результатів. Отримані результати мають велику наукову цінність. По-перше, створена класифікаційна модель дозволяє виконати класифікацію даних фотометричних та астрометричних даних з каталогів CatWISE та GaiaDR3. По-друге, отримані оцінки червоних зміщень дозволяють уточнити шкалу відстаней на різних космологічних масштабах, що особливо важливо для вдосконалення сучасних моделей будови та еволюції Всесвіту.

1 Огляд літератури

1.1 Фізичні особливості АЯГ

Активні ядра галактик — це астрофізичні джерела, що живляться за рахунок акреції матерії на надмасивні ($10^6 M_{\odot}$) чорні діри, які знаходяться в ядрах усіх правильних галактик, і представляють унікальні сигнатури спостереження, які охоплюють повний електромагнітний спектр. Спостереження в різних електромагнітних діапазонах, а також варіації в часі забезпечують повною інформацією щодо будови та еволюції АЯГ.

Для ідентифікації цих джерел використовуються різні методи в різних діапазонах довжин хвиль, але, що найважливіше, різні діапазони випромінювання відповідальні за різні фізичні явища які виникають в АЯГ. Зокрема, інфрачервоний діапазон здебільшого чутливий до пилу та матеріалу що затемнює, оптичний/ультрафіолетовий діапазон пов'язаний з випромінюванням акреційного диска, тоді як рентгенівський діапазон відстежує випромінювання (передбачуваної) корони. З іншого боку, γ - та радіо- випромінювання (з високою щільністю потоку) переважно виокремлюють АЯГ, що випромінює сильне нетеплове (пов'язане зі струменем або пелюстком) випромінювання (див. рис. 1.1). Розподіл поверхневої щільності АЯГ також сильно різниться по всьому електромагнітному спектру через вплив різних факторів.

АЯГ випромінює практично в усіх діапазонах і, отже, історично описувалося в термінах кількох класів об'єктів залежно від діапазону, в якому вони були виявлені. Класифікація АЯГ детально розглянута у [\[1\]](#).

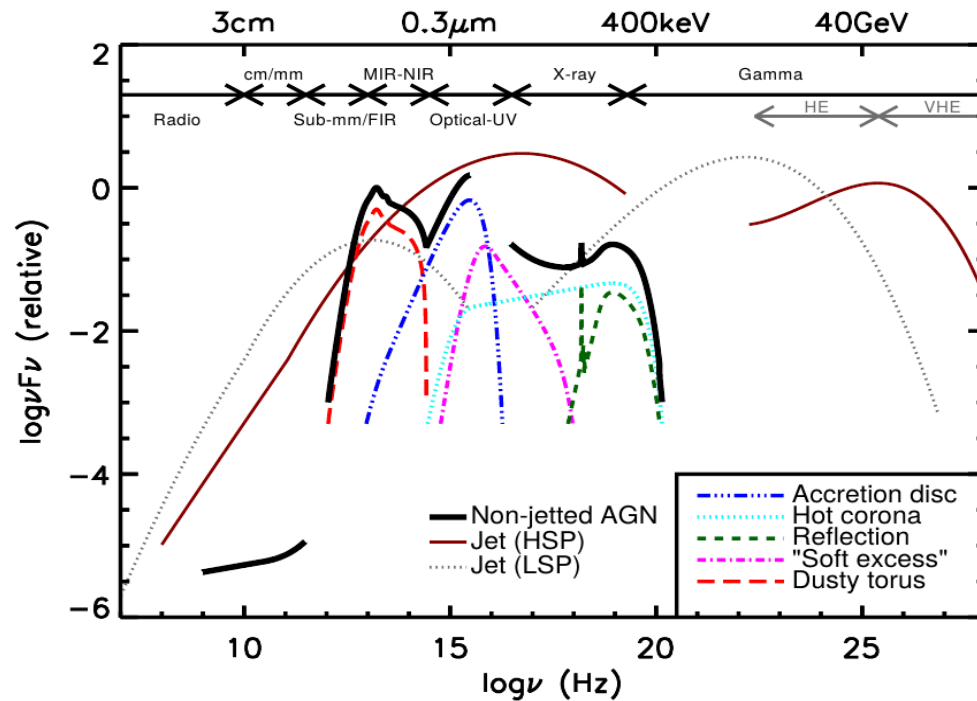


Рисунок 1.1 Схематичне представлення спектрального розподілу енергії АЯГ, частково засноване на спостереженнях не струминних квазарів. Чорна суцільна крива представляє загальне значення, а різнокольорові криві представляють окремі компоненти. Зображення взято з роботи С. М. Harrison [1].

1.2 Методи класифікації АЯГ

Протягом багатьох років було запропоновано ряд критеріїв виокремлення АЯГ у СІЧ. Перші з них уже були розроблені для місії IRAS [2], і їхня кількість надзвичайно зросла після появи в останнє десятиліття космічних телескопів Spitzer [3,4], AKARI [5] і WISE [6,7,8,9]. Ці критерії відбору, як правило, були відкалібровані відповідно до незалежних методів відбору АЯГ і покладаються головним чином на кольори, щоб відокремити АЯГ від зірок або галактик з неактивними ядрами, оскільки очікується, що АЯГ буде значно червонішим у короткохвильових діапазонах СІЧ [4]. Ці критерії не працюють, для випадків АЯГ з червоними зміщеннями $4 \lesssim z \lesssim 5$ для яких лінія випромінювання $H\alpha$ робить додатковий вклад у відповідні довжини хвиль [10].

Assef разом з командою [9], представили цікаве порівняння критеріїв відбору на основі WISE та Spitzer (див. Додаток А). Використовуючи зразок великого, багатохвильового, спектроскопічно ідентифікованих АЯГ, вони визначили, наскільки надійним і повним є кожен критерій.

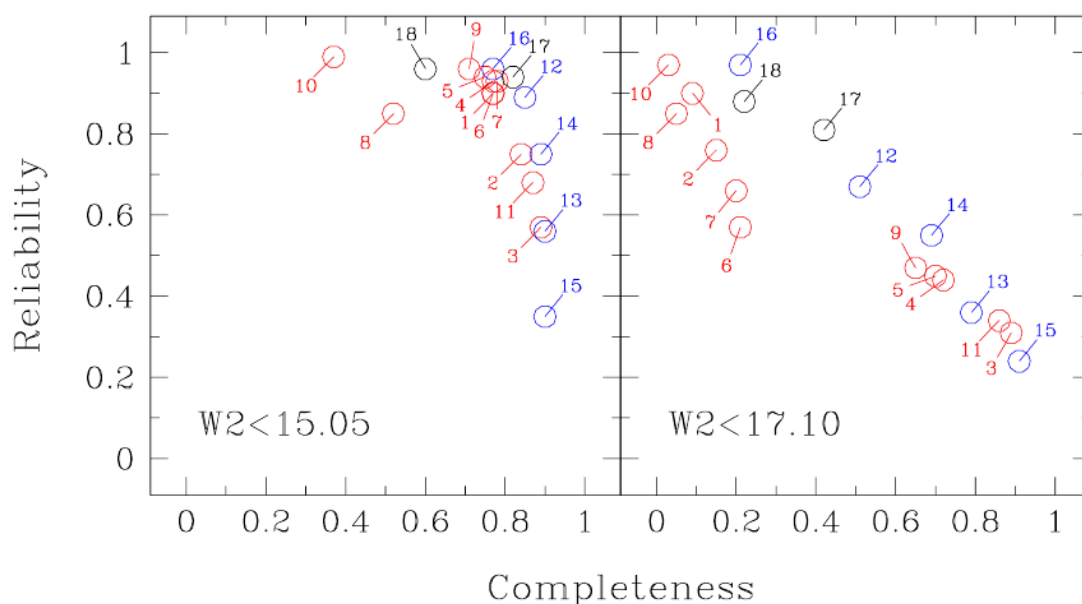


Рисунок 1.2 Порівняння повноти та надійності популярних фотометричних критеріїв вибору АЯГ в СІЧ, оцінених за допомогою тесту, розробленого Assef [9]. Ліва та права панелі показують випадки не глибоких та глибоких спостережень відповідно. Показані критерії відповідають посиланням, наведеним у Додатку А. Червоні точки показують лише критерії відбору WISE, сині точки показують лише критерії відбору Spitzer, а чорні точки показують критерії БІЧ+Spitzer [1].

На рисунку 1.2 показано результати для двох граничних величин $W2$ Vega, а саме $W2 < 15,05$ ліворуч, для найменших полів у місії WISE, і $W2 < 17,1$ праворуч, для даних на рівні 3σ . Видно, що малі та глибокі дослідження необхідно аналізувати окремо. Для не глибоких досліджень, більшість критеріїв відбору дозволяють отримати зразки з високою надійністю та повнотою. Однак для глибоких досліджень ситуація інша, більшість критеріїв вибору вишиковуються по діагоналі на діаграмі

повноти-надійності. Слід також мати на увазі, що через їх нижчу чутливість повнота критеріїв відбору, що спираються на діапазони W3 і W4, нижча, ніж повнота, отримана за допомогою критеріїв, отриманих на основі вибірки, обмеженої W2. Вибір, що спирається на ці довші довжини хвиль, може бути особливо корисним у полях WISE ближче до полюсів екліптики, де візерунок оглядового сканування щільніший, а коротші смуги хвиль досягають межі розмиття [1,11].

Baldwin разом з Phillips та Terlevich [12] вперше запропонував діагностичну діаграму $[N II]/H\alpha$ проти $[O III]/H\beta$ (нині відому як діаграма BPT (див. рис. 3)) для розділення нормальних областей H II, планетарних туманностей та об'єктів, фотоіонізованих сильнішим полем випромінювання. Жорстке поле випромінювання може бути створено або степеневим континуумом від АЯГ, або від ударного збудження. Veilleux на пару з Osterbrock [13] розширили та вдосконалили цю класифікаційну схему. Вони використали теоретичні моделі фотоіонізації, щоб визначити форму лінії класифікації між ЗУГ та АЯГ, і додали дві нові діагностичні діаграми, які використовують співвідношення ліній $[S II]/H\alpha$ і $[O I]/H\alpha$. Kewley [14] поєднала моделі синтезу зоряної популяції та фотоіонізації, щоб побудувати першу суто теоретичну схему класифікації для відділення АЯГ від галактик, що містять утворення зірок (див. рис. 4). [15]

Спектральна класифікація за межами $z > 0,8$ у минулому була важкою. Співвідношення $[N II]/H\alpha$ червоно зсувається в БЧ область при $z \sim 0,5$, тоді як при червоних зміщеннях $z \sim 1,5$ усі стандартні оптичні діагностичні співвідношення вже знаходяться в БЧ області. На щастя, багатооб'єктні спектрографи в БЧ областях на 8–10 метрових телескопах тепер дозволяють класифікувати статистично значущу кількість галактик відповідно до джерела енергії, яке в них домінує.

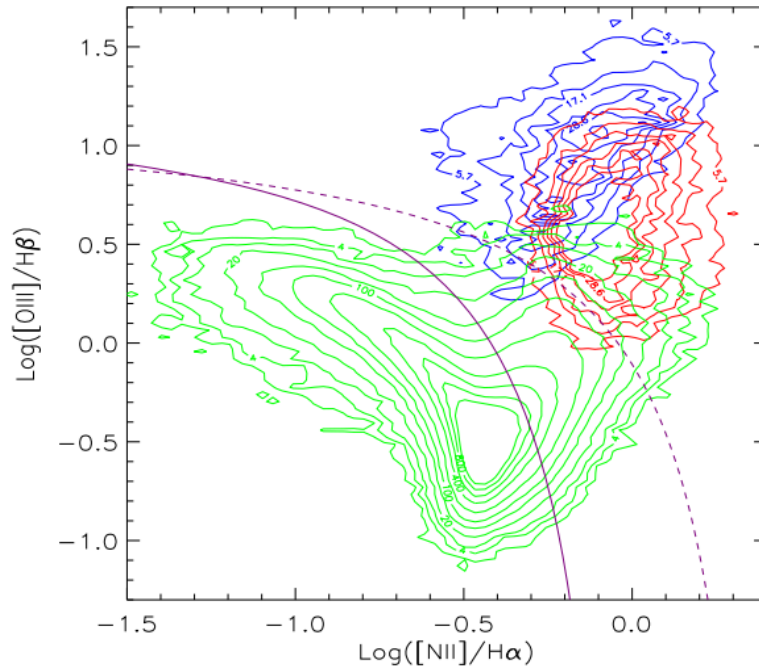


Рисунок 1.3 Діаграма BPT [12] для AGN 5437 Type-1 (контур синім) і для AGN 6587 Type-2 (контур червоним) з надійними вузькими лініями випромінювання. Суцільні та пунктирно-пунктирні фіолетові лінії представляють лінії розділу, які обговорювали Kauffmann [16] та Kewley [17]. Контур зеленого кольору показує результати на основі більш ніж 240 000 галактик з вузькими лініями випромінювання в SDSS DR12 [18].

Співвідношення $[\text{N II}]/\text{H}\alpha$ і $[\text{O III}]/\text{H}\beta$ зараз спостерігаються для невеликої кількості галактик із високим червоним зміщенням [19]. Більшість цих галактик демонструють більші співвідношення $[\text{N II}]/\text{H}\alpha$ і $[\text{O III}]/\text{H}\beta$, ніж у локальних ЗУГ галактиках. Brinchmann [20] використовував моделі синтезу зоряної популяції та фотоіонізації, щоб показати, що локальні галактики з великими співвідношеннями $[\text{N II}]/\text{H}\alpha$ та $[\text{O III}]/\text{H}\beta$ мають більший параметр іонізації, ніж загальна популяція ЗУГ. Вони припускають, що вища густина електронів і більша частка вихідних Н-іонізуючих фотонів можуть бути відповідальними за більший параметр іонізації [15].

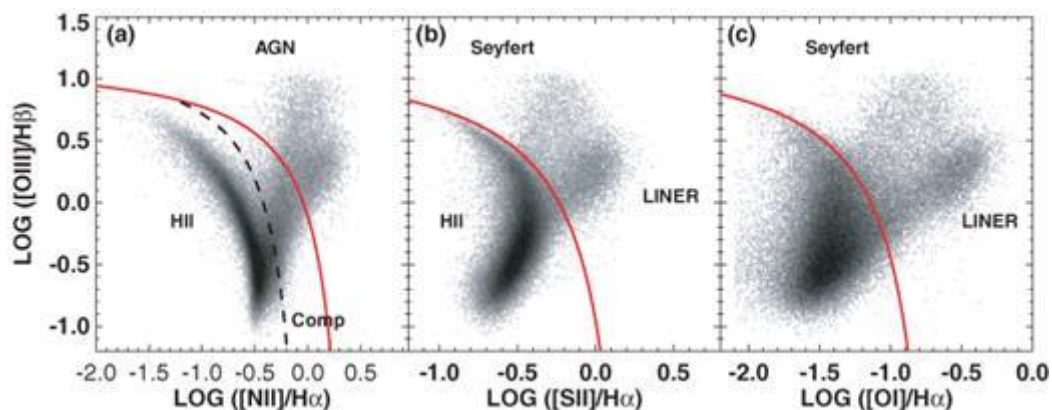


Рисунок 1.4 (а) Діагностична діаграма $[\text{N II}]/\text{H}\alpha$ проти $[\text{O III}]/\text{H}\beta$ для галактик SDSS із $\text{SNR} > 3$. Лінія Kewley яка відмежовує СЗУГ [14] та лінія класифікації яку запропонував Kauffmann [16] показані суцільною та пунктирною лініями відповідно. (b) Діагностична діаграма $[\text{S II}]/\text{H}\alpha$ проти $[\text{O III}]/\text{H}\beta$; (c) діагностична діаграма $[\text{O I}]/\text{H}\alpha$ проти $[\text{O III}]/\text{H}\beta$ [17].

Оцінка ефективності класифікації АЯГ складна, оскільки це потребує встановлення контрольованого експерименту з класифікації, але є задокументовані випадки фальшивої ідентифікації джерела, які були скасовані при більш детальному дослідженні [21]. З іншого боку, ефективність автоматичних підходів можна легко оцінити на невидимому тестовому наборі даних. З цих причин класифікація АЯГ для надзвичайно великих наборів даних, таких як SDSS, ймовірно, потребуватиме автоматизованого підходу [22].

1.3 Методи машинного навчання для класифікації АЯГ

Машинне навчання є практичним інструментом для автоматизації завдань класифікації, які включають кілька відомих класів і велику кількість ознак притаманних таким класам, за умови можливості відрізнити їх від інших [23,24]. Наприклад, Chang з командою у роботі [25] ідентифікували різні класи кандидатів на АЯГ, використовуючи каталоги оптичної та ближньої інфрачервоної фотометрії.

Ключем до ефективної класифікації методами МН є класифікаційна модель, яка створена на точних навчальних даних. Класифікація за спектроскопічними дослідженнями є ідеальними навчальними даними завдяки своїй високій надійності. Було виконано кілька робіт, які досліджують продуктивність класифікації Зірки/Галактики/Квазари [26]. У цих дослідженнях класифікатори МН були побудовані з фотометричними кольорами та спектроскопічними класами, і вони показують більш точні прогнози, ніж інші традиційні методи, такі як показники кольору [27]. Проте все ще є деякі місця в багатоколірному просторі, які не були досліджені класифікаторами через малий розмір спектроскопічних даних. Тому потрібен класифікатор з МН, побудований на основі великої спектроскопічної вибірки, щоб охопити більш повний багатоколірний простір і забезпечити точну класифікацію мільярдів джерел [28].

Дивно, що автоматична класифікація за допомогою МН оптичних спектрів АЯГ була застосована лише кілька разів на основі штучних нейронних мереж [29] та k-NN [30]. У всіх випадках акцент був зроблений на правильній автоматичній класифікації, а не на інтерпретації отриманої моделі. Це також стосується останнього та, наскільки нам відомо, найточнішого результату класифікації АЯГ, заснованого на керованій структурі методу МН, представлений Тао [31]. Вони навчили різні моделі машинного навчання, які можна розглядати як “чорний ящик”, на 10000 спектрах SDSS DR-14, досягнувши надзвичайно високої продуктивності класифікації, $\approx 93\%$ за показником F1 (див. пункт 3.2.1). Автори також використовують важливість ознак випадкового лісу [32], щоб отримати деяке уявлення про те, які основні компоненти простору ознак спектрів є більш інформативними, але не обговорюють їх фізичне значення. Попри високу продуктивність класифікації, сучасному автоматизованому класифікуванню АЯГ бракує інтерпретації [22].

Також у літературі був описаний метод навчання SVM [33] моделі для класифікації AGN за спектральними даними, який дозволив отримати досить

точні результати на тестовому наборі ($F1 \approx 94\%$). Спроба фізично інтерпретувати роботу моделі, виявила що області спектра, які найбільше впливають на прогноз класифікатора, подібні до тих, які використовує людина-експерт, а саме області навколо спектральних ліній [O II] 3727, He I 3889, H δ 4101, H γ 4340, H β 4861, [O III] 4959 і [O III] 5007. Таким чином, Peruzzi та команда [22], роблять висновок, принаймні для розглянутих ними спектрів, що їх класифікатор працює так само як і людина, просто автоматично та набагато швидше.

Однак існують додаткові відомі проблеми, з якими стикаються методи МН у більш загальному плані, а саме: екстраполяційне прогнозування [34]. Наприклад, моделям, заснованим на глибокому навчанні, не вистачає кількісного визначення невизначеності, що погіршує їхню здатність робити прогнози, окрім даних, які вони бачили [34]. Кількісна оцінка невизначеності та упереджені моделі є темами активних досліджень у глибокому навчанні, і вони мають вирішальне значення для забезпечення високоточної науки [35].

2 Огляд вхідних даних

Наявність фотометричних оглядів всього неба у комбінації з якісними спектроскопічними даними грає фундаментальне значення для створення класифікаційної моделі. Через велику чисельність АЯГ в оптичному діапазоні хвиль, ми можемо застосовувати до них різноманітні методи статистичної обробки даних та аналізу для подальшого створення класифікаційної моделі.

2.1 Вибір спектроскопічних даних - навчальна вибірка

В даній роботі був використаний 17-й випуск даних 4-го етапу місії SDSS [36]. Він включає дані SDSS, отримані до січня 2020 року, (e)BOSS [37] та SEGUE [38] і охоплює понад третину всієї небесної сфери.

Класифікація джерел яку пропонує SDSS [39] заснована на критерії діаграми BPT яка була детально розглянута в огляді літератури. Вона дозволяє розділити об'єкти на класи: Квасари, АЯГ, ЗУГ, СЗУГ, Галактики та Зорі, залежно від значення полів CLASS та SUBCLASS таблиці SpecObjAll (див. табл. 2.1).

До того ж ми відбираємо об'єкти з прапорцем ZWARNING=0, додатково вилучаємо джерела які мають помилку визначення червоного зміщення більше ніж 10% від значення та обмежуємо об'єкти класу Квazarів у червоному зміщенні по правилу $z < 4.0$, що відповідає вищій надійності класифікації та оцінки червоного зміщення. Ми гарантуємо, що між 6-ма наборами даних для класів немає спільних джерел.

Назва класу	CLASS	SUBCLASS	Кількість
Квасари	QSO	All without AGN, AGN BROADLINE	942881
АЯГ	QSO, GALAXY	AGN, AGN BROADLINE	36172
СЗУГ	GALAXY	STARBURST, STARBURST BROADLINE	99385

ЗУГ	GALAXY	STARFORMING, STARFORMING BROADLINE	287806
Галактики	GALAXY	NULL, BROADLINE	2444919
Зорі	STAR	All	443269

Таблиця 2.1 Значення полів CLASS та SUBCLASS таблиці SpecObjAll які відповідають об'єктам певного класу та кількість представників кожного з них.

2.2 Вибір фотометричних каталогів - досліджувальна вибірка

З метою узагальнення правила класифікації на якнайбільшу кількість джерел ми скористаємося даними фотометричних оглядів неба. Для цього ми візьмемо два найбільших на цей час каталоги з оптичного та середнього інфрачервоного діапазону хвиль відповідно. Цей вибір обумовлений численними дослідженнями [40,41,1] які вказують на суттєву відмінність яскравості об'єктів різних класів у цих фотометричних смугах [42].

Gaia DR3 [43] – це третій випуск даних космічного проєкту Gaia ESA. Цей каталог містить дані для більше ніж 1.8 млрд. об'єктів що отримані в результаті роботи перших 34 місяців космічного телескопа Gaia. Фотометричні величини G_{BP} та G_{RP} отримані для приблизно 1,542 млрд. та 1,555 млрд. джерел відповідно. Крім того, фотометричні величини G наведені для 1,806 млрд джерел. Також каталог містить точні астрометричні дані (положення, власні рухи та паралакси) для 1.56 млрд. об'єктів. В рамках третього релізу було отримано астрофізичні параметри для 470 млн. зір та виконано попередню класифікацію для 1.5 млрд. об'єктів серед яких є квазари, галактики.

Каталог CatWISE2020 [44] налічує близько 1,8 млрд. об'єктів з визначеною фотометрією у ближньому ІЧ діапазоні, вибраних з даних

спостережень WISE [45] і NEOWISE на смугах 3.4 і 4.6 мкм (W1 і W2 відповідно), які збиралися з 7 січня 2010 року по 13 грудня 2018 року. На відміну від AllWISE [46], який ми використовували раніше, список виявлених об'єктів для CatWISE2020 був створений за допомогою "краудсорсингового" програмного забезпечення [47] що дозволило більше ніж у два рази збільшити кількість розпізнаних об'єктів. Повнота каталогу на рівні 90% знаходиться приблизно на рівні значень $W1=17.7$ і $W2=17.5$. У порівнянні з Gaia, значення власних рухів об'єктів CatWISE2020 більш ніж у десятки разів точніші, ніж власні рухи AllWISE.

Координатне ототожнення каталогів GaiaDR3 та CatWISE2020, з радіусом вікна пошуку в 1 секунду дуги, було виконано за допомогою інструменту CDS X-match [48]. У результаті чого були отримані об'єкти з характеристиками GaiaDR3 та вимірами зоряної величини CatWISE2020 у діапазонах W1 та W2. Для створення класифікаційної моделі паралакси та власні рухи не були використані. Такий підхід дає набагато більший набір галактик, оскільки більшість галактик, що спостерігаються в Gaia, не мають значень паралаксів і власних рухів через погану відповідність астрометричної моделі яка зумовлена їх фізичною протяжністю. Зміна кількості значень на різних етапах обробки наведена у таблиці 2.2.

2.3 Попередній аналіз і обробка даних

Цей етап є не менш важливим за проєктування класифікаційної моделі, адже, відсутність впевненості у навчальному матеріалі може привести нас до не очікуваних наслідків при передбаченні. Залежно від того, який саме аспект попередньої підготовки даних був не врахований можна зрозуміти наскільки є універсальним те правило яке пропонує нам конкретна модель класифікації. У нашому випадку будемо намагатися врахувати якнайбільше факторів які впливають на точність розпізнавання класів.

Відомо, що повнота та гранична величина фотометрії Gaia демонструють складні просторові варіації, пов'язані насамперед із законом

сканування Gaia [49]. Через це ми обмежили вибірку на основі каталогу Gaia DR3 діапазоном значень смуги G з 14-ї до 22-ї зоряної величини [40]. Для CatWISE2020 цей показник складає 11 та 19, для нижньої та верхньої межі відповідно.

Ми залишаємо джерела які мають не нульові значення параметрів $nW1$ та $nW2$, що відповідають кількості зображень на яких об'єкт має $SNR \geq 3$ для кожної фотометричної смуги. Також ми виключаємо з вибірки об'єкти для яких відсутні важливі для нас значення фотометрії та їх помилки.

Через те, що ми маємо справу з об'єктами які знаходяться від нас на значній відстані, було б добре додатково врахувати втрати які зазнає світло під час подорожі до приймача, але узагальнення впливу поглинання частки випромінювання пилом коштує обмеження області значень, тобто, площі на небесній сфері, не є дієвим, оскільки розподіл пилу по небесній сфері нерівномірний і, взагалі-то, не лінійний. Для редагування значень зоряних величин у вибірці позагалактичних джерел випромінювання, ми користуємося мапою кольорового почервоніння SFD [50] з урахуванням коефіцієнта впливу до шуканого фотометричного фільтра [51]. Через те, що модель яка вираховує коефіцієнт обмежена значенням $E(B-V) \leq 0.5$, ми застосовуємо цей фільтр до всіх об'єктів навчальної вибірки. Значення кольорового почервоніння для конкретних координат ми отримуємо за допомогою спеціально розробленої бібліотеки Python [52].

Також, ми врахуємо викиди. Викидом у статистиці вважається значення, яке суттєво виділяється з поміж загальної вибірки. Це може відбутися за різних причин, наприклад, помилки у вимірюванні, відсутності даних або екзотичної природи джерела. Оскільки у нашому випадку ми маємо справу з багатовимірним простором ознак, то буде вигідно звести задачу з розпізнавання викидів до одновимірної. Розрахуємо для кожного об'єкта окремого класу відстань Махаланобіса [53], яка у порівнянні з Евклідовою враховує кореляцію між змінними та інваріантна до масштабу, і задається формулою

$$d(x_i) = \sqrt{(x_i - T)'S^{-1}(x_i - T)}$$

де x_i - значення ознак у векторному вигляді, T - середнє арифметичне всіх значень x_i , а S - класична коваріаційна матриця.

Припустимо, що ці значення мають нормальний розподіл та використовуючи правило 3-х сигм виконаємо очищення даних від статистичних викидів.

Назва класу	Origin	CatWISE	Gaia DR3	Врахування відсутніх значень	Врахування почервоніння	Врахування діапазону	Врахування викидів
Квазари	942881	564740	391433	387432	387300	364545	358452
АЯГ	36172	26916	16837	16805	16798	16326	16056
СЗУГ	99385	55345	23596	23374	23374	22910	22557
ЗУГ	287806	176891	86451	86260	86256	85266	84026
Галактики	2444919	1718425	393987	388762	388577	381642	376558
Зірки	443269	274727	270424	243039	240389	235973	232680
Разом	4.25 млн.	2.8 млн.	1.2 млн	1.1 млн	1.1 млн	1.1 млн	1090359

Таблиця 2.2 Кількість об'єктів кожного класу під час попередньої обробки даних.

У результаті було створено, на базі даних спектроскопічного каталогу SDSS DR17, навчальну вибірку для різних класів: Квазари, АЯГ, СЗУГ, ЗУГ, галактики та зорі та виконано доповнення їх фотометричними даними з каталогів GaiaDR3 та CatWISE. Блок-схема створення навчальної вибірки для всіх класів об'єктів наведена в Додатку В.

3 Методика класифікації та регресії

Для ідентифікації об'єктів визначених раніше класів та розрахунку червоних зміщень було використано методи машинного навчання з учителем. Через те, що ми маємо справу з нелінійними залежностями між компонентами яскравості світил, доцільно буде скористатися методами, за допомогою яких можна побудувати гнучку модель класифікації.

Загальний алгоритм навчання класифікаційної моделі наведений у Додатку D.

3.1 Моделі класифікації

3.1.1 НМ - нейрона мережа

Вже існує багато робіт з використанням концепції глибокого навчання нейронної мережі для визначення приналежності об'єктів до позагалактичного класу [54, 25]. Саме тому, ми використовуємо модель прямого поширення на базі Keras [55], який є інтерфейсом прикладного програмування для глибокого навчання, написаним мовою Python, що працює на базі модуля TensorFlow. Така модель складається з поступово з'єднаних шарів нейронів, де кожен шар має одну вхідну функцію активації. Структура моделі вказана у вигляді блок-схеми в Додатку C. З огляду на швидкість обчислень і ефективність функції оптимізації Adam [56], ми використовуємо всього 200 епох для тренування моделі. Встановлення 6 нейронів в останньому шарі перетворює всю послідовну модель на нормалізований класифікатор.

3.1.2 Алгоритм випадковий ліс

Для розрахунку червоного зміщення ми використовуємо алгоритм випадкового лісу – це широко використовуваний алгоритм керованого машинного навчання, який генерує надійні моделі та ефективно працює з великими наборами даних.

Алгоритм ВЛ спирається на ансамбль дерев рішень для прогнозування як для задач класифікації, так і для задач регресії [32]. Дерева рішень будуються незалежно на основі особливостей (тобто властивостей джерела в нашому випадку) вхідних наборів даних, які є різними підвибірками початкової навчальної вибірки.

Дерево рішень будується за принципом "зверху вниз". У кожному вузлі дерева набір даних розбивається на дві підмножини відповідно до ознаки серед випадково вибраної підмножини всіх ознак, яка дає найбільший інформаційний виграш. Вузли вирощуються рекурсивно, поки не буде виконано критерій зупинки. У задачі класифікації кожне дерево обчислює ймовірність (1 або 0) приналежності вхідного об'єкта до певного класу, і повертає середню ймовірність класу для всіх дерев. У задачі регресії кожне дерево робить прогноз щодо невідомої величини, яка нас цікавить (у нашому випадку – фотометричне червоне зміщення), а середнє значення з усіх дерев використовується як остаточна оцінка.

Алгоритм ВЛ був успішно застосований до різноманітних задач в астрономії [57], включаючи класифікацію АЯГ та фотометричну оцінку червоного зміщення [40, 58].

3.2 Методи оцінки якості роботи моделі

Для визначення якості моделі ми покладаємося на принцип KFold [59] який дозволяє випадковим чином поділити початкову вибірку на певну кількість ідентичних частин (у нашій роботі 4) з метою сліпого тестування вже навченої моделі на підвибірках основної вибірки, які не брали участь у навчанні. За цю певну кількість циклів ми встигаємо навчити та протестувати модель на всіх даних. Для прикладу, надалі ми будемо використовувати або результати з 2-го розподілу KFold, або усереднене значення усіх 4-х.

Але для рівноцінного застосування моделі до двох різних наборів даних обов'язковою є умова єдності діапазону значень їх ознак в одному й тому

самому просторі ознак. Тому для цього ми за допомогою однокласового методу SVM [60] виділяємо об'єм який займає сукупність значень ознак вибірки позагалактичних джерел випромінювання та застосовуємо це правило до каталогу який збираємося класифікувати. Його програмне втілення є у відкритому доступі у складі окремої бібліотеки Python [61].

3.2.1 Метрики ефективності класифікаційної моделі

Ефективність класифікації оцінюється на наборі даних, який не використовувався на етапі навчання, тобто на тестовому наборі даних. Ми будемо оцінювати ефективність за допомогою таких метрик, як влучність (precision), повнота (recall), показник F1 та збалансована точність (bacc). Влучність, також відома як точність, є відношенням кількості істинних позитивних передбачень (TP) до всієї кількості передбачених об'єктів у цільовому класі. До того ж влучність може розглядатися як міра забруднення, що відображає частоту хибнопозитивних спрацьовувань (FP). Чим вище значення влучності, тим менше забруднення.

$$precision = \frac{TP}{TP + FP}$$

Повнота, також відома як відгук або чутливість, є відношенням кількості TP до загальної суми правильних прогнозів та істинних невиявлень (хибно негативних (далі FN)) для цільового класу. Досконала модель має влучність та повноту, що дорівнюють 1.

$$recall = \frac{TP}{TP + FN}$$

Оцінка F1 обчислюється як середнє гармонійне значення повноти та влучності моделі:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Досконала модель має оцінку F1, що дорівнює 1.

Оскільки ми маємо справу з нерівномірним розподілом кількості об'єктів для кожного з класів, буде доречно скористатися збалансованою точністю, що

дозволяє уникнути такого роду спотворення під час оцінки якості моделі [62].

Формула за якою обчислюється значення $bacc$ має такий вигляд:

$$bacc = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

Але для більш загальної оцінки ми скористаємося кривою робочої характеристики приймача (ROC-крива) [63], та кривою залежності Влучності та Повноти від порогу ймовірності. Перша має за основу співвідношення ймовірності правильного детектування об'єктів цільового класу (TPR) до ймовірності помилкового детектування об'єктів іншого класу (FPR). Кількісну інтерпретацію ROC дає показник AUC, який є площею обмеженою ROC-кривою. Досконала модель має ROC криву у вигляді точки з координатами (0,1), з показником AUC рівним 1.

Оцінка повноти та влучності класифікаційної моделі може дещо погіршуватися для малочисельних класів порівняно з великою кількістю зірок, які спостерігаються в Gaia DR3. [41]

3.2.2 Метрики якості прогнозування червоного зміщення

Тут ми вводимо дві міри якості фотометричної оцінки червоного зсуву. Точність фотометричного червоного зміщення (σ) зазвичай визначається в літературі як нормоване середнє абсолютне відхилення [64]:

$$\sigma_{NMAD} = 1.48 \times \text{медіана} \left(\frac{|z_{phot} - z_{spec}|}{1 + z_{spec}} \right)$$

де z_{phot} – це червоне зміщення розраховане за фотометричними даними, а z_{spec} – червоне зміщення розраховане за спектроскопічними даними.

Використання σ_{NMAD} є кращим за стандартне відхилення, оскільки медіана менш чутлива до екстремальних значень, а масштабний коефіцієнт 1.48 введено для того, щоб можна було інтерпретувати σ_{NMAD} як стандартне відхилення нормально розподілених даних.

Другою мірою ми будемо вважати кількість катастрофічних викидів, адже для низки джерел фотометричне червоне зміщення обов'язково є

хибним. Такими ми будемо вважати ті значення для яких виконується рівняння що наведене в роботі [65]:

$$|z_{phot} - z_{spec}| > 0.15 \cdot (1 + z_{spec})$$

Модель передбачення вважається хорошою, якщо цей показник не перевищує 5%.

3.3 Інтерпретації роботи класифікаційних моделей

Для розуміння роботи класифікаційної моделі, у випадку простої моделі, трудностей не виникає при аналізі самої моделі. А для складних моделей, таких як ансамблеві методи або глибокі мережі, ми не можемо використовувати оригінальну модель як пояснення, тому що її нелегко зрозуміти. Замість цього ми повинні використовувати простішу модель пояснення, яку ми визначаємо як будь-яке інтерпретоване наближення оригінальної моделі [66].

Метою SHAP [67] є пояснення прогнозу для вектору ознак шляхом обчислення внеску кожної ознаки в прогноз. Метод пояснення SHAP обчислює значення Шеплі з теорії коаліційних ігор. Значення ознак екземпляра даних діють як гравці в коаліції. Значення Шеплі показують нам, як справедливо розподілити "виплату" (= прогноз) між інвесторами. Для табличних даних, гравцем може бути окреме значення ознаки. Гравцем також може бути група значень ознак. Наприклад, щоб пояснити зображення, пікселі можна згрупувати в суперпікселі, а прогноз розподілити між ними. Одне з нововведень SHAP полягає в тому, що пояснення значень Шеплі представлено у вигляді адитивного методу атрибуції ознак, лінійної моделі. Цей погляд пов'язує значення LIME і Шеплі. SHAP визначає пояснення як

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j$$

де g – модель пояснення, $z' \in \{0, 1\}^M$ вектор коаліції, M – максимальний розмір коаліції та $\phi_j \in \mathbb{R}$ атрибуція ознаки для ознаки j , значення Шеплі [66].

З недоліків варто відзначити складність обчислення, яка напряду впливає на швидкість та майже унеможлиблює використання цього методу для великої кількості даних.

Ми використовуємо реалізацію у вигляді пакета для Python [68].

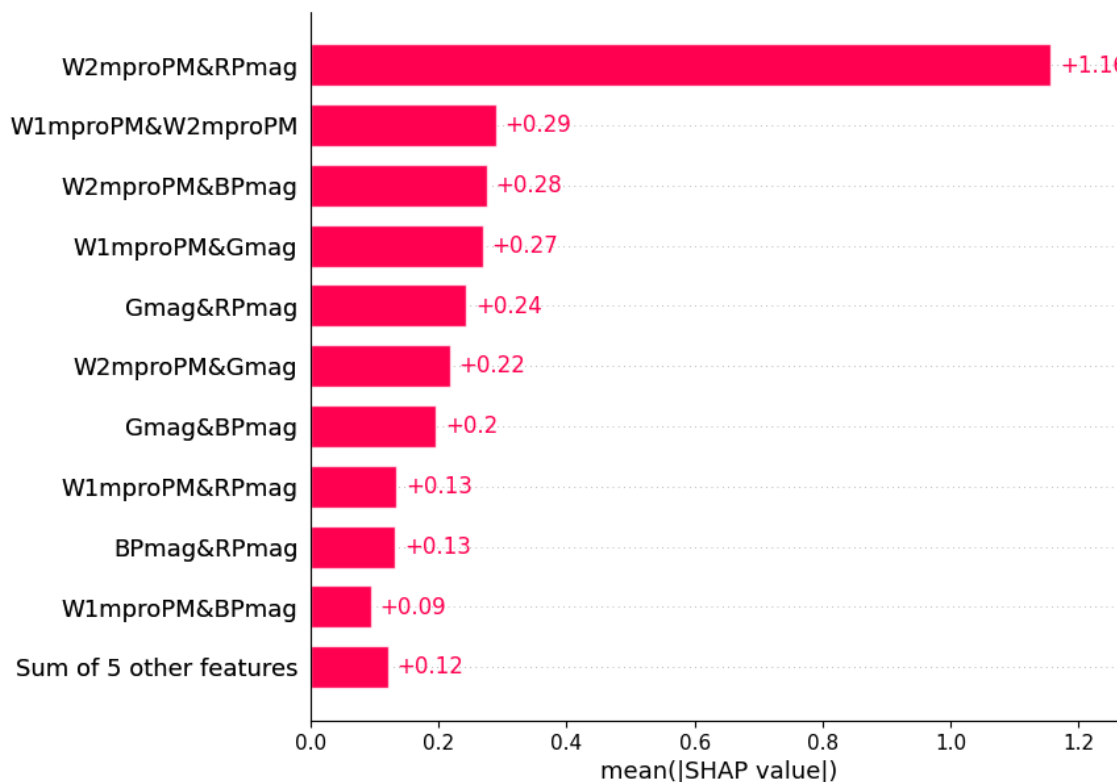


Рисунок 3.1 Один з можливих варіантів застосування методу SHAP. На малюнку вказане середнє абсолютне значення SHAP для кожної ознаки. Для обчислення ми використали 2000 випадково обраних об'єктів з навчальної вибірки. Загальний час виконання розрахунків зайняв ~40 хв.

t-SNE [69] - це алгоритм неконтрольованого зменшення розмірності, який використовується для візуалізації та дослідження даних. Метою зменшення розмірності є відображення даних високої розмірності у простір нижчої розмірності (у нашому випадку площини) зі збереженням попарних відстаней між точками. Це неможливо зробити строго, тому що високорозмірний простір не може бути вкладений у площину, але t-SNE досягає цього приблизно, надаючи пріоритет відстаням точок, які знаходяться

близько одна від одної, тому короткі відстані спотворюються найменше, тоді як великомасштабна структура набору даних здебільшого втрачається.

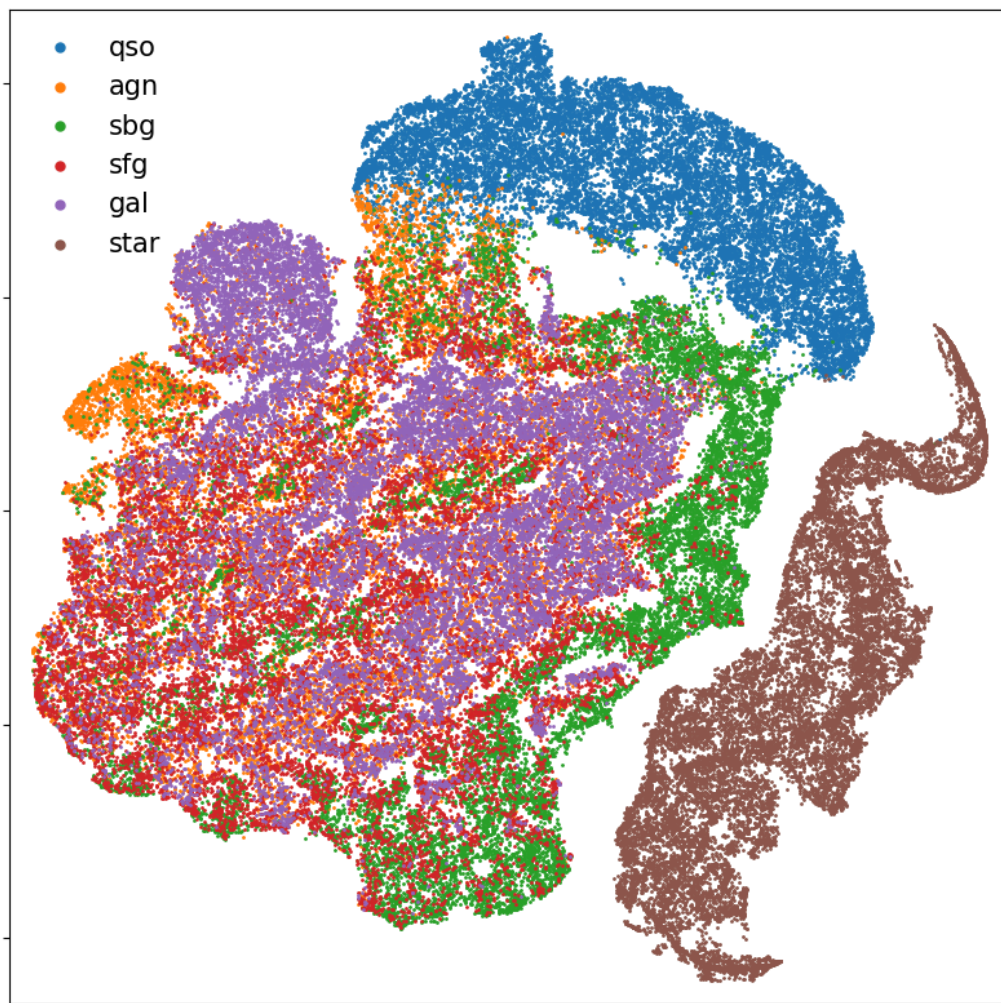


Рисунок 3.2 Приклад застосування t-SNE для створення площини з розподілом класів навчальної вибірки. Для розрахунку ми створили вибірку з рівномірним розподілом кількості об’єктів по класам та встановили такі гіпер-параметри моделі: perplexity = 100, learning_rate = 200.0, min_grad_norm = 1e-7, metric = “euclidean”, init = “pca”, method = “barnes_hut”, angle = 0.5.

Це досягається шляхом мінімізації втрат

$$L = - \sum_{i \neq j} p_{ij} \log \frac{q_{ij}}{p_{ij}}$$

де p_{ij} – міра подібності між точками i та j у вихідному високорозмірному просторі, а q_{ij} – (інша) міра подібності у низькорозмірному просторі. У той

час як p_{ij} розпадається як гаусівський розподіл з відстанню між точками i та j , q_{ij} розпадається як t-розподіл Стюдента з одним ступенем свободи, звідси й назва алгоритму. З рівняння втрат видно, що точки, які знаходяться далеко одна від одної у високовимірному просторі, не вносять значного внеску у втрати, оскільки їхні p_{ij} прямують до нуля експоненціально з квадратом відстані. Результат t-SNE залежить від гіпер-параметру perplexity, який визначає стандартне відхилення гаусівського показника, що використовується для визначення p_{ij} , і може бути вільно інтерпретований як типовий розмір підгруп, що очікується в даному наборі даних [22].

Ми використовуємо t-SNE у реалізації scikit-learn [70] для Python.

4 Результати

4.1 Класифікація даних

Для тренування нашої моделі ми вирішили застосувати похідні значення від яскравості об'єктів у різних фотометричних фільтрах, а саме їх різницю, так звані кольори. Поєднання цих самих величин ми провели між всіма наявними значеннями та отримали 10 кольорів. Також, ми додали до них показники які можуть надати нам додаткові знання про мінливість джерела [40]. Значення FW1 ми обчислюємо по такій формулі:

$$FW1 = \sqrt{nW1} \times \frac{e_{FW1pm}}{FW1pm}$$

де $nW1$ – кількість спостережень за час проведення місії зі значенням $SNR > 3$, $FW1pm$ – це кількісне значення потоку частинок від джерела які потрапили на приймач, e_{FW1pm} – фонове значення потоку.

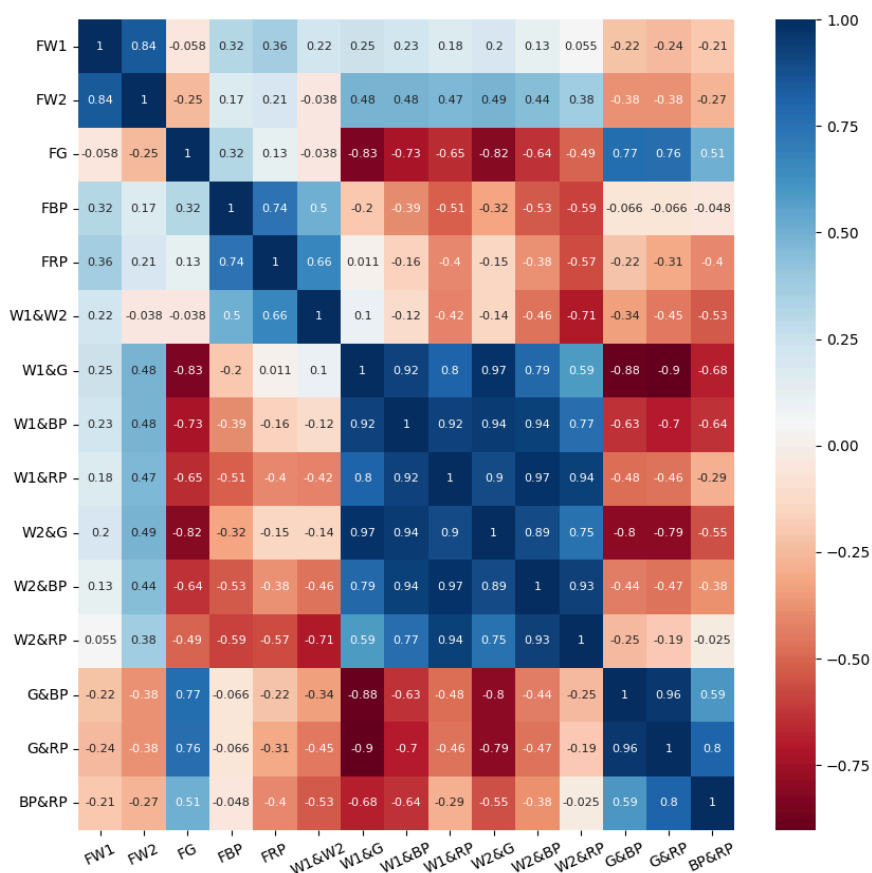


Рисунок 4.1 Кореляційна матриця вхідних даних. (умовні позначення у скороченому вигляді)

Вид кореляційної матриці (див. рис. 4.1) вказує на явний лінійний зв'язок між кольорами які мають спільну фотометричну базу, але ми не виключаємо їх у надії що це допоможе краще зрозуміти природу класифікації.

Для самої класифікації в моделі були виставлені відповідні гіпер-параметри:

batch size	epoch	optimizer	loss	validation split
1024	200	Adam	categorical_crossentropy	0.2

Також, додатково, ми скористалися модулем передчасного завершення навчання, на випадок, якщо модель почне “перенавчання”, що погано позначиться на якості самої класифікації. Ми виставили параметри що відповідає зупинці через 100 епох у випадку відсутності зміни максимального значення повноти для валідаційної вибірки впродовж поточного навчання та збереження тих значень ваги нейронів моделі для яких значення повноти було максимальним.

monitor	patience	mode	restore_best_weights
val_recall	100	max	True

Оскільки ми маємо справу з нерівномірним розподілом кількості джерел по класах, буде доречним розрахувати для кожного з них його частку в загальній вибірці та додати до моделі в якості гіпер-параметра `class_weights`. У нашому випадку цей параметр приймає такі значення:

qso	agn	sbg	sfg	gal	star
0.5070	11.3183	8.0563	2.1627	0.4826	0.7810

Хід навчання можна прослідкувати на графіку зміни втрат залежно від значення епохи (див. рис. 4.2). Різка зміна якості передбачення протягом перших 5 епох, є особливістю оптимізації метода Adam. Пунктирною лінією позначені оцінки для валідаційної вибірки, відносно велика амплітуда їх значень зумовлена незбалансованістю класів.

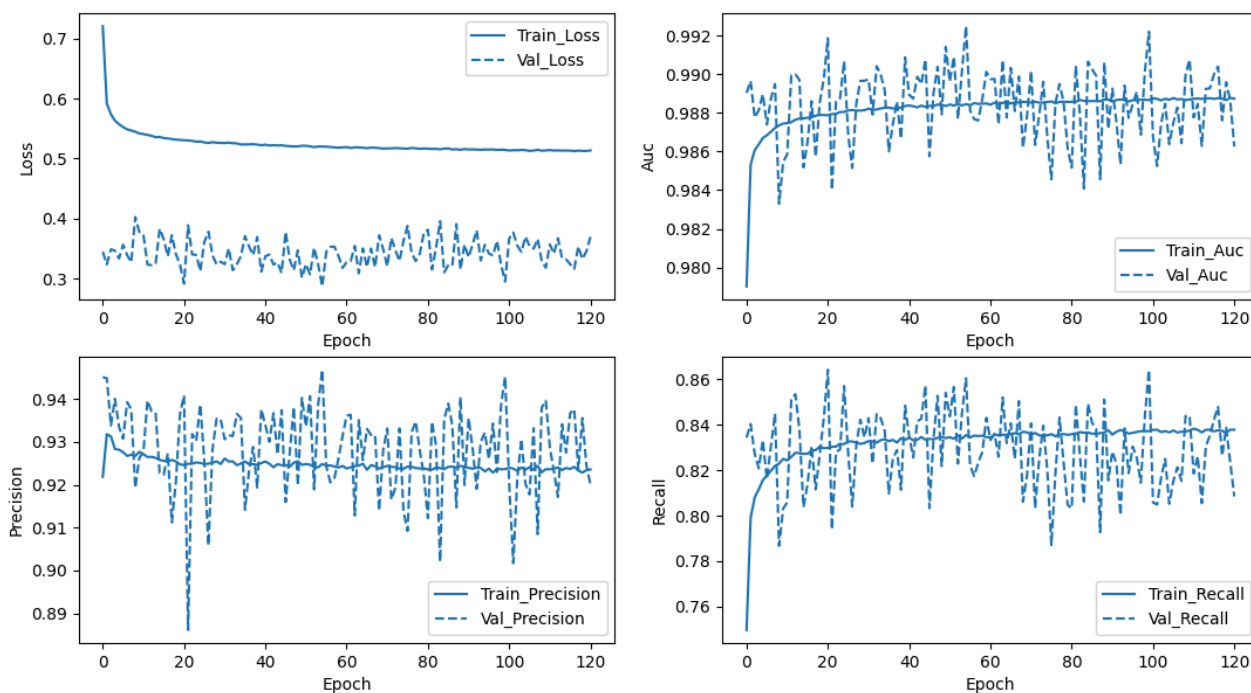


Рисунок 4.2 Зміна втрат (loss) та метрик під час навчання на всій вибірці з включенням валідаційної вибірки. Значення для підвибірок KFold міститься в Додатку Е.

У результаті ми отримали розподіл передбачення ймовірності приналежності об'єкта до певного класу залежно від порогового значення вірогідності. Для всіх наступних обчислень ми визначаємо приналежність до певного класу на основі найбільшого значення ймовірності приналежності об'єкта до класу (див. рис. 4.3).

На їх основі ми розраховували значення метрик для кожного з класів (див. Додаток F). На рисунку 4.4 представлена їх візуалізація, з якої випливає факт про великі труднощі в розрізненні АЯГ та ЗУГ у порівнянні з іншими класами.

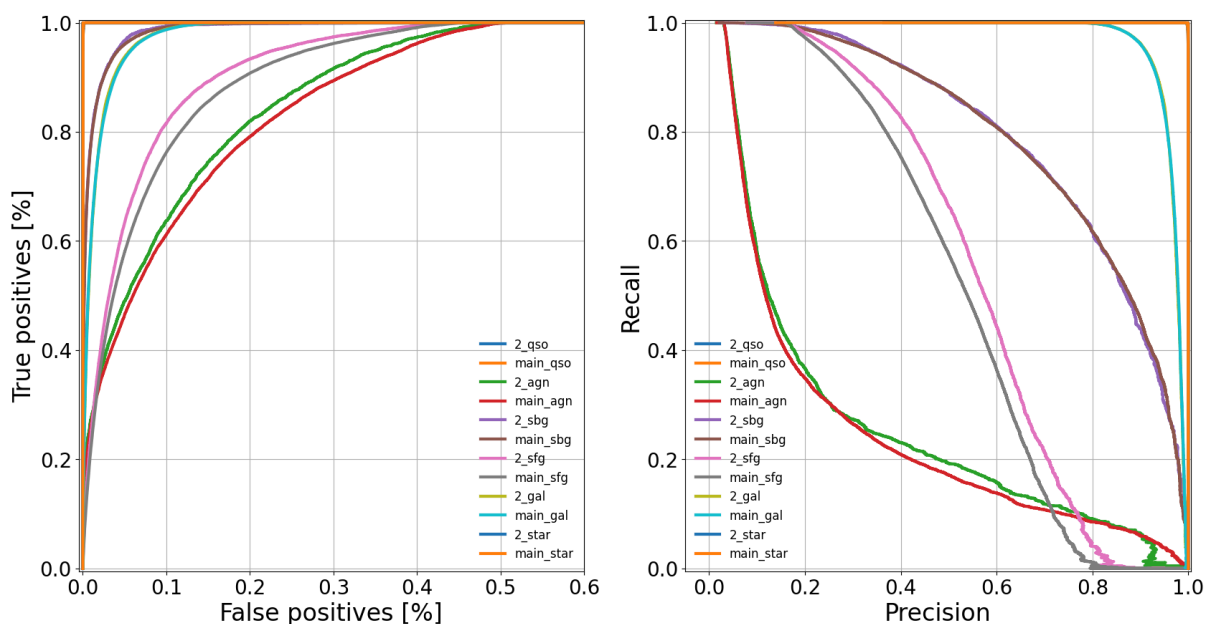


Рисунок 4.3 Криві ROC (з лівого боку) та залежності Повноти від Влучності (з правого боку).

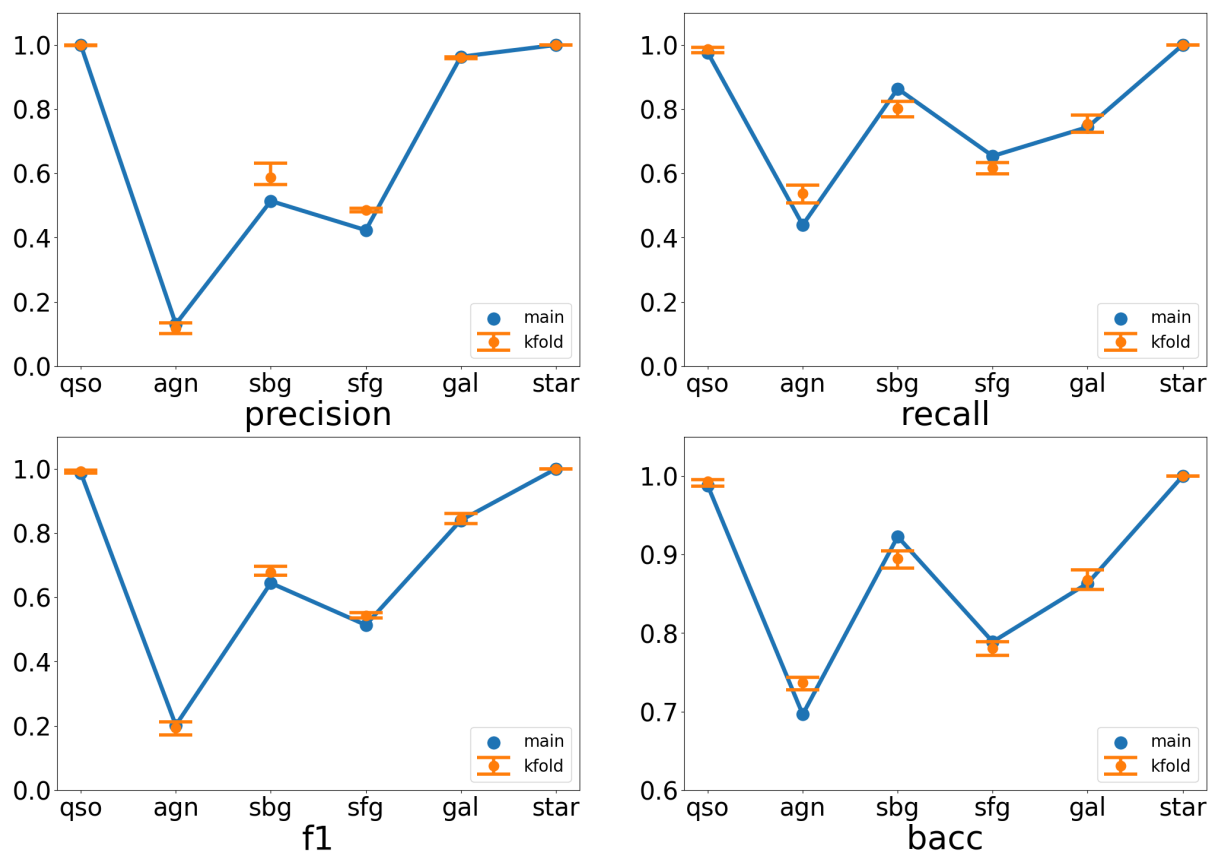


Рисунок 4.4. Візуалізація значень метрик для кожного з класів. Точка kfold є арифметичним середнім усіх значень, натомість верхня та нижня риски відповідають їх максимальному та мінімальному значенням.

Для розуміння природи таких низьких показників ми скористаємося матрицею невідповідностей (див. рис. 4.5). Вона може надати інформацію про те з якими саме класами модель “плутає” інші класи. На ній видно абсолютну перевагу Квazarів та Зірок у показниках якості розділення. Серед галактик непогано проявляються СЗУГ, наступні за ними ЗУГ, а для АЯГ більшість об’єктів цього класу класифікаційна модель віднесла до інших класів, що при високій забрудненості АЯГ іншими класами загалом каже нам про те, що проблема є у самих вхідних даних.

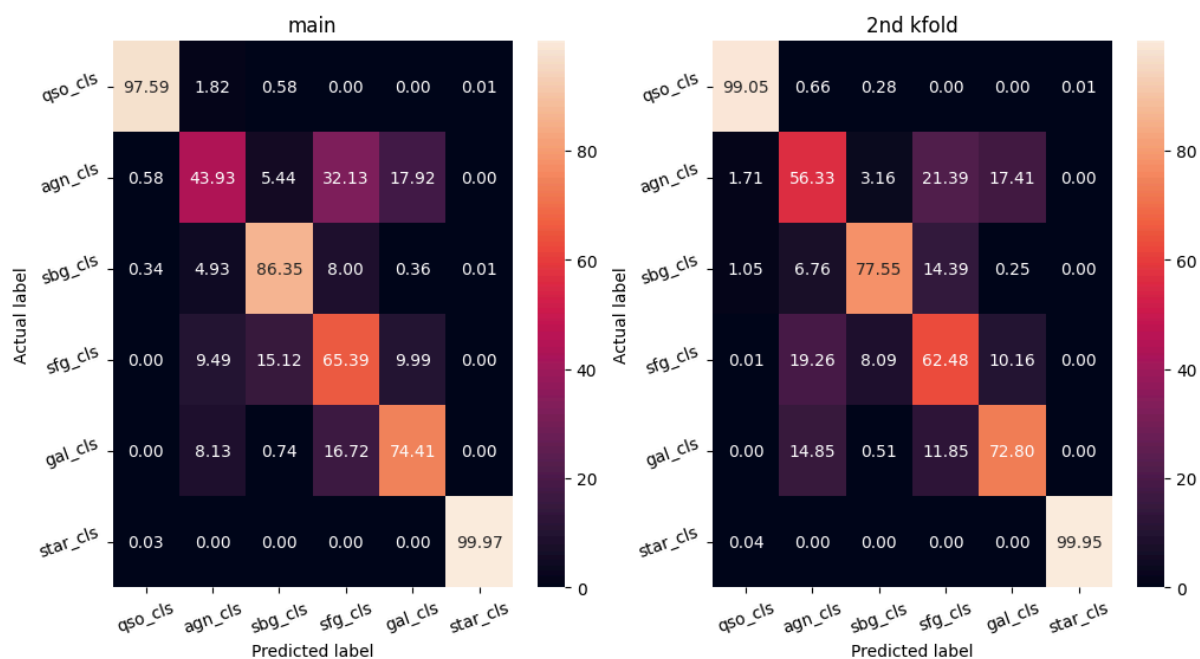


Рис. 4.5 Матриця невідповідностей класифікованих об’єктів.

Метрики вказують на дуже низьке значення влучності для АЯГ, що пов’язане з хибним віднесенням до їх числа частини об’єктів галактик, а також, повноти, що вказує на малу кількість повторно знайдених вже відомих АЯГ, що може вказувати на те, що вони є змішаними з іншими класами галактик. Це відкриває питання, щодо присутності АЯГ компоненти у цих галактиках, або, навпаки, відсутності у тих, які ми першочергово вважали АЯГ.

Якщо подивитися на переріз побудований на основі двох кольорів які найбільше вплинули на результати класифікації (див. рис. 4.6), можна зрозуміти, що задача класифікації трьох основних типів джерел на відміну від

класифікації підмножин галактик, не складає труднощів. Знаходження всіх підкласів в одному геометричному місці точок ускладнює процес класифікації, а в цьому конкретному випадку, унеможлиблює побудову моделі яка б з великою точністю могла їх розрізняти.

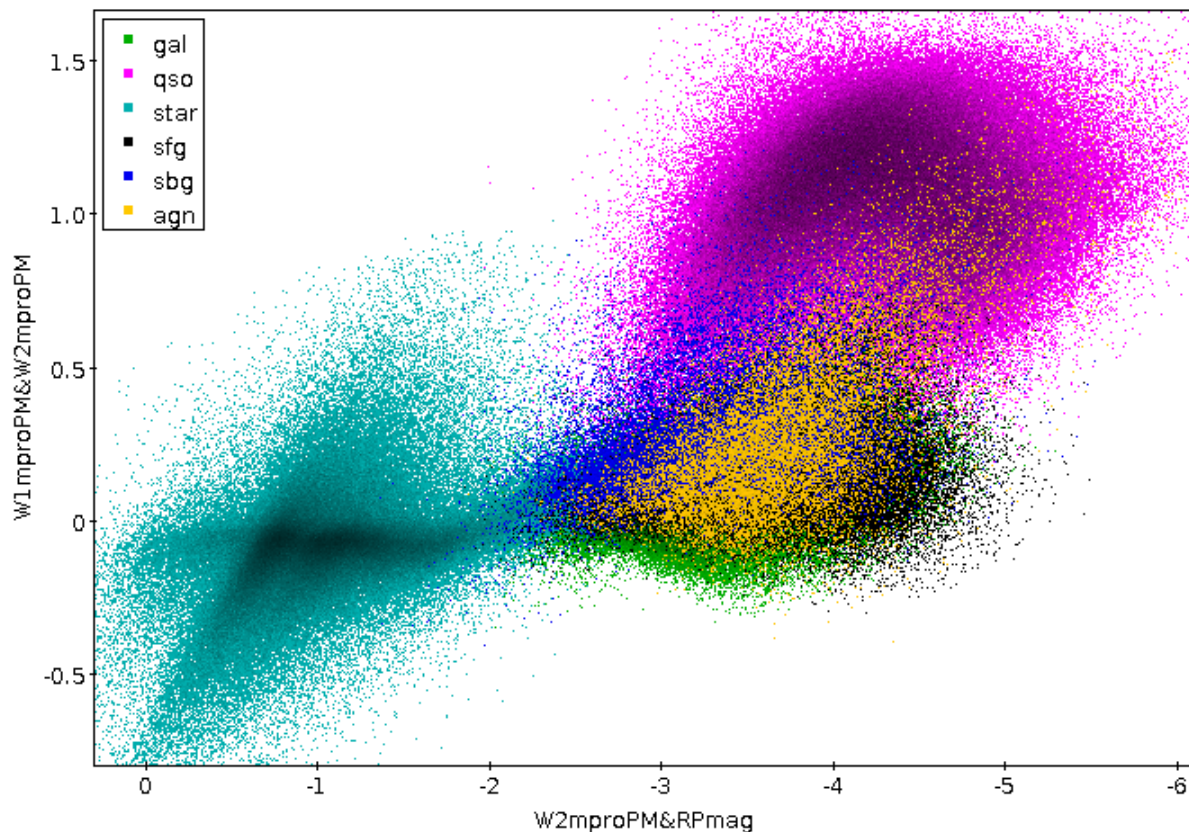


Рисунок 4.6 Переріз навчальної вибірки для двох найбільш впливових (див. пункт 3.1) значень кольорів.

4.2 Оцінки червоних зміщень

Знаючи, що кожен з класів займає різні діапазони значень червоного зміщення (див. рис. 4.7) ми не стали створювати для них усіх однієї спільної моделі для визначення червоних зміщень. Натомість ми тренуємо регресивну модель для кожного класу окремо, а потім застосовуємо її залежно від попереднього віднесення класифікатором об'єкта до певного класу.

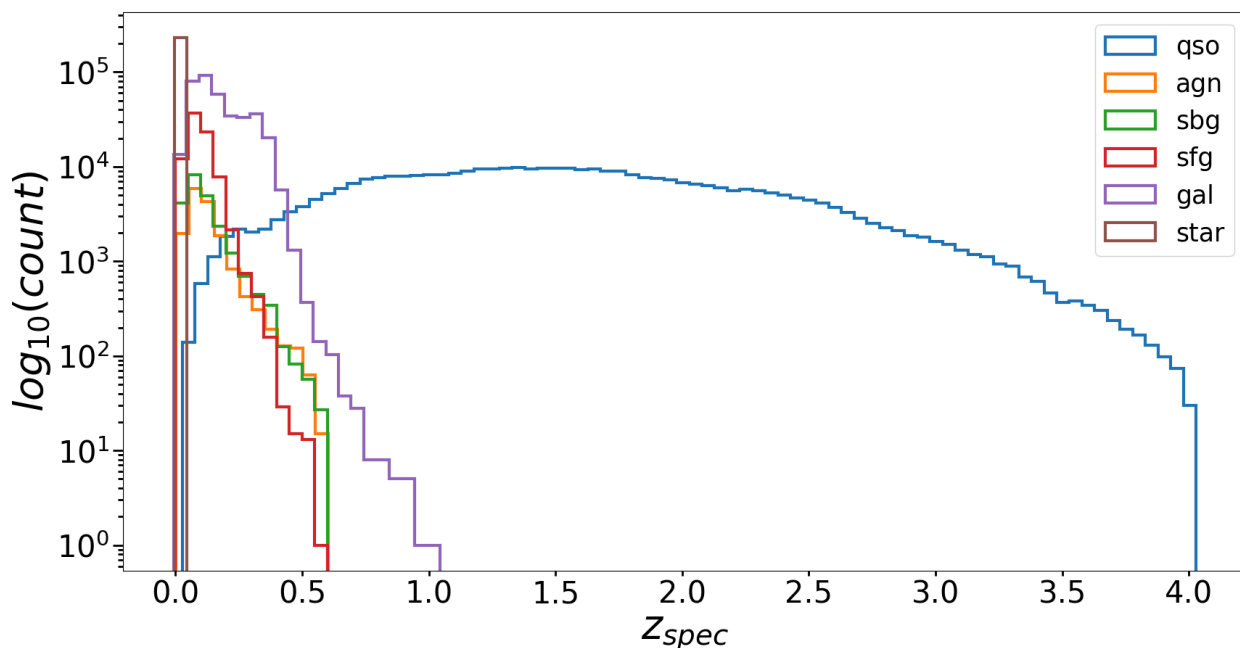


Рисунок 4.7 Розподіл значень червоного зміщення для кожного з класів вибірки.

Для моделі регресії були застосовані відповідні гіпер-параметри, запозичені з [40] як найкращі:

n_estimators	max_features	criterion	max_depth	bootstrap
100	3	squared_error	50	True

На рисунку 4.8 можна побачити результати роботи моделі передбачення червоних зміщень для класу Квазарів. Даний рисунок наведений спеціально як ілюстративний, бо методи оцінювання еволюційно створювалися саме для цього класу об'єктів. Відповідно, для кожного іншого класу об'єктів було створено окрему модель прогнозу червоного зміщення.

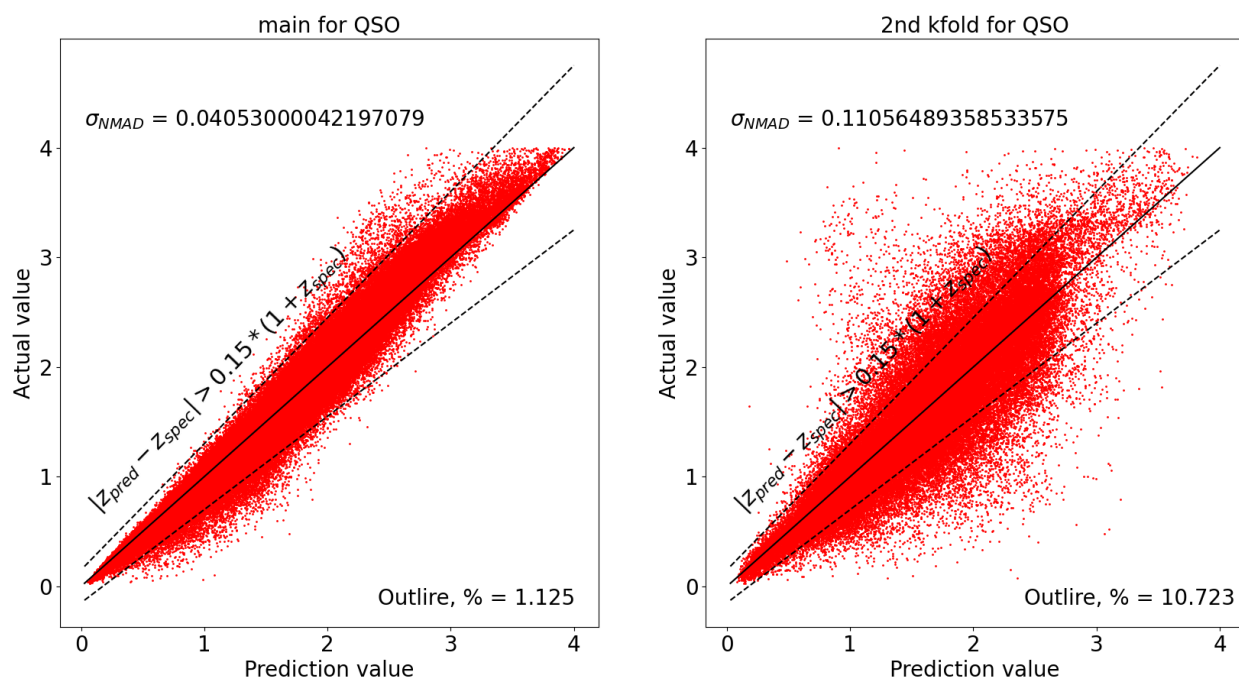


Рисунок 4.8 Діаграма передбачення червоного зміщення для головної вибірки Квазарів та з 2-го набору KFold.

Висновки

У результаті виконання дипломної роботи були отримані наступні наукові результати:

1. Проаналізовано основні сучасні дослідження з тематики класифікації;
2. Створено навчальну вибірку на основі спектроскопічної інформації з каталогу SDSS DR17 та фотометричних даних з каталогів CatWISE та GaiaDR3;
3. Запропоновано методику попередньої підготовки навчальної вибірки;
4. Створено модель, що передбачає значення червоних зміщень за фотометричними даними з $\sigma_{NMAD} < 0.05$;
5. Проведено аналіз результатів роботи класифікаційної моделі.

Виходячи з результатів аналізу, стає зрозуміло, що поточного набору ознак недостатньо для якісної класифікації галактик на підкласи виокремлені SDSS. Не останню роль у цьому зіграла загальна схожість розподілу енергії вздовж спектру для кожного з підкласів трьох головних класів: Квазарів, Галактик та Зір [39]. Натомість точність їх розпізнавання тою ж моделлю перевищує 97%, що свідчить про її спроможність до якісної класифікації.

У майбутніх дослідженнях ми сподіваємося отримати кращі результати залучивши до параметрів навчання візуальне зображення об'єктів, а також спектри низької роздільної здатності від Gaia.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Padovani, P., Alexander, D.M., Assef, R.J. et al. Active galactic nuclei: what's in a name?. *Astron Astrophys Rev* 25, 2 (2017).
2. de Grijp M. H. K., Miley G. K., Lub J., de Jong T. Infrared Seyferts - A new population of active galaxies? *Nature*. 1985. 314:240–242.
3. Werner M. W., Roellig T. L., Low F. J., et al. The Spitzer Space Telescope Mission. *ApJS*. 2004. 154:1–9.
4. Stern D., Eisenhardt P., Gorjian V., Kochanek C. S., Caldwell N., Eisenstein D., Brodwin M., Brown M. J. I., et al. Mid-Infrared Selection of Active Galaxies. *ApJ*. 2005. 631: 163-168
5. Murakami H., Baba H., Barthel P., et al. The Infrared Astronomical Mission AKARI. *PASJ*. 2007. 59:S369.
6. Wright E. L., Eisenhardt P. R. M., Mainzer A. K. et al. The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance *AJ*. 2010. 140 1868
7. Mateos S., Alonso-Herrero A., Carrera F. J. et al. Using the Bright Ultrahard XMM–Newton survey to define an IR selection of luminous AGN based on WISE colours. *MNRAS*. 2012. 426 3271
8. Stern D., Assef R. J., Benford D. J. et al. Mid-infrared Selection of Active Galactic Nuclei with the Wide-field Infrared Survey Explorer. I. Characterizing WISE-selected Active Galactic Nuclei in COSMOS. *ApJ*. 2012. 753 30
9. Assef R. J., Stern D., Kochanek C. S. et al. Mid-infrared Selection of Active Galactic Nuclei with the Wide-field Infrared Survey Explorer. II. Properties of WISE-selected Active Galactic Nuclei in the NDWFS Boötes Field. *ApJ*. 2013. 772 26

10. Assef R. J., Kochanek C. S., Brodwin M. et al. Low-Resolution Spectral Templates for Active Galactic Nuclei and Galaxies from 0.03 to 30 μm . *ApJ*. 2010. 713 970
11. Jarrett T. H., Cohen M., Masci F., et al. The Spitzer-WISE Survey of the Ecliptic Poles. *ApJ*. 2011 735:112.
12. Baldwin J. A., Phillips M., Terlevich R. Classification parameters for the emission-line spectra of extragalactic objects. *PASP*. 1981. 93, 5
13. Veilleux S. and Osterbrock D. E. Spectral Classification of Emission-Line Galaxies. *ApJS*. 1987. 63 295
14. Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A. and Trevena J. Theoretical Modeling of Starburst Galaxies. *ApJ*. 2001. 556 121
15. Lisa J. Kewley et al. The Cosmic BPT Diagram: Confronting Theory with Observations. *ApJL*. 2013. 774 L10
16. Kauffmann G. et al. The host galaxies of active galactic nuclei. *MNRAS*. 2003. 346, 1055
17. Kewley L. J., Groves B., Kauffmann G., Heckman T. The host galaxies and classification of active galactic nuclei. *MNRAS*. 2006. 372, 961
18. Xue-Guang Zhang, Evidence for obscured broad [O III] components in Type-2 AGN, *MNRAS*. 2021. 502, 2, 2508–2512
19. Erb, D. K., Shapley, A. E., Pettini, M., et al. The Mass-Metallicity Relation at $z \sim 2$. *ApJ*. 2006. 644, 813
20. Brinchmann, J., Pettini, M., & Charlot, S. New insights into the stellar content and physical conditions of star-forming galaxies at $z = 2\text{--}3$ from spectral modelling. *MNRAS*. 2008. 385, 769

21. Järvelä, E., Berton, M., Ciroi, S., et al. SDSS J211852.96–073227.5: The first non-local, interacting, late-type intermediate Seyfert galaxy with relativistic jets. *A&A*. 2020. 636, L12
22. T. Peruzzi, M. Pasquato, S. Ciroi, M. Berton, P. Marziani and E. Nardini. Interpreting automatic AGN classifiers with saliency maps. *A&A*. 2021. 652 A19
23. Debosscher, J., Sarro, L. M., Aerts, C., et al. Automated supervised classification of variable stars. *A&A*. 2007. 475, 1159
24. Rimoldini L., Holl B., Garvas P. et al. Gaia Data Release 3: All-sky classification of 12.4 million variable sources into 25 classes. *A&A*. 2023. 674, A14
25. Chang, Y.-Y., Hsieh, B.-C., Wang, W.-H., et al. Identifying AGN Host Galaxies by Machine Learning with HSC+WISE. *ApJ*. 2021. 920, 68
26. Krakowski T., Małek K., Bilicki M. et al. Machine-learning identification of galaxies in the WISE $\times$ SuperCOSMOS all-sky catalogue. *A&A*. 2016. 596 A39
27. Weir N., Fayyad U. M. and Djorgovski S. Automated Star/Galaxy Classification for Digitized POSS-II. 1995. *AJ*. 109 2401
28. Yu Bai, JiFeng Liu, Song Wang, Fan Yang. Machine Learning Applied to Star–Galaxy–QSO Classification and Stellar Effective Temperature Regression. *AJ*. 2019. 157 9
29. Rawson, D. M., Bailey, J., & Francis, P. J. Neural networks and the classification of active galactic nucleus spectra. *PASA*. 1996. 13, 207
30. Zhao, M.-F., Wu, C., Luo, A.-L., Wu, F.-C., & Zhao, Y.-H. Automated Spectral Classification of Broad-line and Narrow-line Active Galactic Nuclei Based on the K-nearest Neighbor Method. *Chinese A&A*. 2007, 31, 352

31. Tao, Y., Zhang, Y., Cui, C., & Zhang, G. Automated Spectral Classification of Galaxies using Machine Learning Approach on Alibaba Cloud AI platform (PAI). 2018. ArXiv:1801.04839v2
32. Breiman, L. Random Forests. Machine Learning. 2001. 45, 5–32
33. Vapnik V., Cortes C. Support-vector networks. Machine Learning. 1995. 20, 273-297
34. Lucie-Smith, L., Peiris, H.V., Pontzen, A., Nord, B., Thiyaalingam, J. Deep learning insights into cosmological structure formation. 2020. arXiv:2011.10577.
35. Priyamvada Natarajan et al. QUOTAS: A new research platform for the data-driven investigation of black holes.2021. arXiv:2103.13932v1
36. Abdurro'uf et al. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. ApJS. 2022. 259 35
37. Dawson, K., Kneib, J.-P., Percival, W., et al. The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. AJ. 2016 151, 44
38. Brian Yanny et al. SEGUE: A Spectroscopic Survey of 240,000 Stars with $g = 14-20$. AJ. 2009. 137 4377
39.
<https://www.sdss4.org/dr17/spectro/catalogs/#Objectinformation:~:text=Back%20to%20Top-,Object%20information,-Each%20spectroscopic%20object>
40. Yiping Shu, Sergey E Koposov, et. al. Catalogues of active galactic nuclei from *Gaia* and unWISE data. MNRAS. 2019. 489, 4, 4741–4759
41. Arvind C. N. Hughes, Coryn A. L. Bailer-Jones, Sara Jamal. Quasar and galaxy classification using Gaia EDR3 and CatWise2020. A&A. 2022. 668, 99, 20 pp.

42. Sanders, D. B., Phinney, E. S., Neugebauer, G., Soifer, B. T., & Matthews, K. Continuum Energy Distributions of Quasars: Shapes and Origins. *ApJ*. 1989. 347, 29
43. Gaia Collaboration: Vallenari A., et. al. Gaia Data Release 3: Summary of the contents and survey properties. *A&A*. 2022. 674, A1, 22
44. Peter R. M. Eisenhardt et al. The CatWISE Preliminary Catalog: Motions from WISE and NEOWISE Data. *ApJS*. 2020. 247 69, URL:
<https://irsa.ipac.caltech.edu/cgi-bin/Gator/nph-scan?mission=irsa&submit=Select&projshort=WISE>
45. Cutri, R. M. et al. VizieR Online Data Catalog: WISE All-Sky Data Release. 2012. URL: <http://wise2.ipac.caltech.edu/docs/release/allsky/expsup>
46. Cutri, R. M. et al. VizieR Online Data Catalog: AllWISE Data Release. 2013. URL: <https://wise2.ipac.caltech.edu/docs/release/allwise/expsup/index.html>
47. Edward F. Schlafly et al. The unWISE Catalog: Two Billion Infrared Sources from Five Years of WISE Imaging. *ApJS*. 2019. 240 30
48. URL: <http://cdsxmatch.u-strasbg.fr/>
49. Arenou et al. *Gaia* Data Release 2: Catalogue validation. 2018. *A&A*. 616, A17
50. Schlegel D. J., Finkbeiner D. P., Davis M. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *ApJ*. 1998. 500, 2
51. Ruoyi Zhang, Haibo Yuan. Empirical Temperature- and Extinction-dependent Extinction Coefficients for the GALEX, Pan-STARRS 1, Gaia, SDSS, 2MASS, and WISE Passbands. *ApJS*. 2023. 264 14

52. URL: https://github.com/vnohhf/extinction_coefficient/
53. Mahalanobis P.C. On the Generalised Distance in Statistics. *Sankhya A*. 1936. 80, 1-7e
54. Chen B.H. et. al. An active galactic nucleus recognition model based on deep neural network. *MNRAS*. 2021. 501, 3, 3951–3961
55. URL: <https://keras.io/>
56. Diederik P. Kingma et. al. Adam: A Method for Stochastic Optimization. 2017. arXiv:1412.6980v9
57. Chen et. al. Three-dimensional interstellar dust reddening maps of the Galactic plane. *MNRAS*. 2019. 483, 4, 4277–4289
58. Nakoneczny et al. Catalog of quasars from the Kilo-Degree Survey Data Release 3. *A&A*. 2019. 624, A13
59. URL:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
60. Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Computation*. 2001. 13(7), pp.1443-1471
61. URL: <https://github.com/yzhao062/pyod>
62. Mosley L. A balanced approach to the multi-class imbalance problem. *IJCV*. 2010.
63. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006. 27(8), 861-874.

64. Ilbert, O. et al. Cosmos Photometric Redshifts with 30-Bands for 2-deg². ApJ. 2009. 690, 1236
65. Fotopoulou S., Paltani S., CPz: Classification-aided photometric-redshift estimation. A&A. 2018. 619, A14
66. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). 2022. URL: <https://christophm.github.io/interpretable-ml-book/shap.html>
67. Lundberg S. Lee S.-I. A Unified Approach to Interpreting Model Predictions. 2017. arXiv:1705.07874v2
68. URL: <https://github.com/slundberg/shap>
69. van der Maaten L., Hinton G. Viualizing data using t-SNE. JMLR. 2008. 9, 2579-2605
70. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. JMLR. 2011. 12, 2825-2830

ДОДАТКИ

Додаток А

ID	Reference	Criteria
WISE only criteria		
(1)	Assef et al. (2013) – R90 [†]	$W1 - W2 > 0.662 \exp\{0.232 (W2 - 13.97)^2\}$
(2)	Assef et al. (2013) – R75 [†]	$W1 - W2 > 0.530 \exp\{0.183 (W2 - 13.76)^2\}$
(3)	Assef et al. (2013) – C90 [†]	$W1 - W2 > 0.50$
(4)	Assef et al. (2013) – C75 [†]	$W1 - W2 > 0.77$
(5)	Stern et al. (2012) [‡]	$W1 - W2 > 0.80$
(6)	Jarrett et al. (2011)*	$W2 - W3 > 2.2 \wedge W2 - W3 < 4.2 \wedge$ $W1 - W2 > (0.1 (W2 - W3) + 0.38 \wedge$ $W1 - W2 < 1.7, \text{ except}$ (i) $W1 < 10.5 \wedge W2 - W3 < 1.5 \wedge$ $W1 - W2 < 0.4 \text{ or}$ (ii) $W3 - W4 < 1.2$
(7)	Mateos et al. (2012)* – 3-band	$W1 - W2 > -3.172 (W2 - W3) + 7.624 \wedge$ $W1 - W2 < 0.315 (W2 - W3) + 0.796 \wedge$ $W1 - W2 > 0.315 (W2 - W3) - 0.222$
(8)	Mateos et al. (2012)* – 4-band	$W1 - W2 > -2.00 (W3 - W4) + 4.33 \wedge$ $W1 - W2 < 0.50 (W3 - W4) + 0.979 \wedge$ $W1 - W2 > 0.50 (W3 - W4) - 0.405$
(9)	Assef et al. (2010) – 2-band	$W1 - W2 > 0.85$
(10)	Assef et al. (2010) – 4-band	$W3 - W4 > 2.1 \wedge W1 - W2 > 0.85 \wedge$ $W1 - W2 > 1.67 (W3 - W4) - 3.41$
(11)	Wu et al. (2012b)	$W1 - W2 > 0.57$
Spitzer only criteria		
(12)	Stern et al. (2005)	$[5.8] - [8.0] > 0.6 \wedge$ $[3.6] - [4.5] > 0.2 ([5.8] - [8.0]) + 0.18 \wedge$ $[3.6] - [4.5] > 2.5 ([5.8] - [8.0]) - 3.5$
(13)	Lacy et al. (2004)	$\log f_{8.0}/f_{4.5} > -0.2 \wedge \log f_{5.8}/f_{3.6} > -0.2 \wedge$ $\log f_{8.0}/f_{4.5} < 0.8 \log f_{5.8}/f_{3.6} + 0.5$
(14)	Lacy et al. (2007)	$\log f_{8.0}/f_{4.5} > -0.2 \wedge \log f_{5.8}/f_{3.6} > -0.1 \wedge$ $\log f_{8.0}/f_{4.5} < 0.8 \log f_{5.8}/f_{3.6} + 0.5$
(15)	Lacy et al. (2013)	$\log f_{8.0}/f_{4.5} > -0.3 \wedge \log f_{5.8}/f_{3.6} > -0.3 \wedge$ $\log f_{8.0}/f_{4.5} < 0.8 \log f_{5.8}/f_{3.6} + 0.5$
(16)	Donley et al. (2012)	$\log f_{5.8}/f_{3.6} \geq 0.08 \wedge \log f_{8.0}/f_{4.5} \geq 0.15 \wedge$ $\log f_{8.0}/f_{4.5} \geq 1.21 \log f_{5.8}/f_{3.6} - 0.27 \wedge$ $\log f_{8.0}/f_{4.5} \leq 1.21 \log f_{5.8}/f_{3.6} + 0.27 \wedge$ $f_{4.5} > f_{3.6} \wedge f_{5.8} > f_{4.5} \wedge f_{8.0} > f_{5.8}$
NIR + Spitzer criteria		
(17)	Messias et al. (2012) – KI	$K_s - [4.5] > 1.42 \wedge [4.5] - [8.0] > 1.14$
(18)	Messias et al. (2012) – KIM	$K_s - [4.5] > 1.42 \wedge [8.0] - [24] > 2.87 \wedge$ $[8.0] - [24] > -2.9 ([4.5] - [8.0]) + 8.47$

[†]Calibrated for galaxies with $W1 < 18.50$ and $W2 < 17.11$.

[‡]Calibrated for galaxies with $W2 < 15.05$.

*Calibrated for galaxies with $W1 < 18.1$, $W2 < 17.2$, $W3 < 13.4$.

*Calibrated for galaxies detected with $S/N > 5$ in all WISE bands used.

Таблиця А-1. Таблиця критеріїв відбору АЯГ, де фотометричні смуги представлені в системі Vega, тоді як f_x представляє щільність потоку смуги Х. Для емпірично відкаліброваних критеріїв відбору WISE ми вказуємо діапазон величин калібрувального зразка, якщо він застосовувався, оскільки порівняння на рис. 2 може бути екстрапольоване на менші величини в деяких випадках [1].

Додаток В

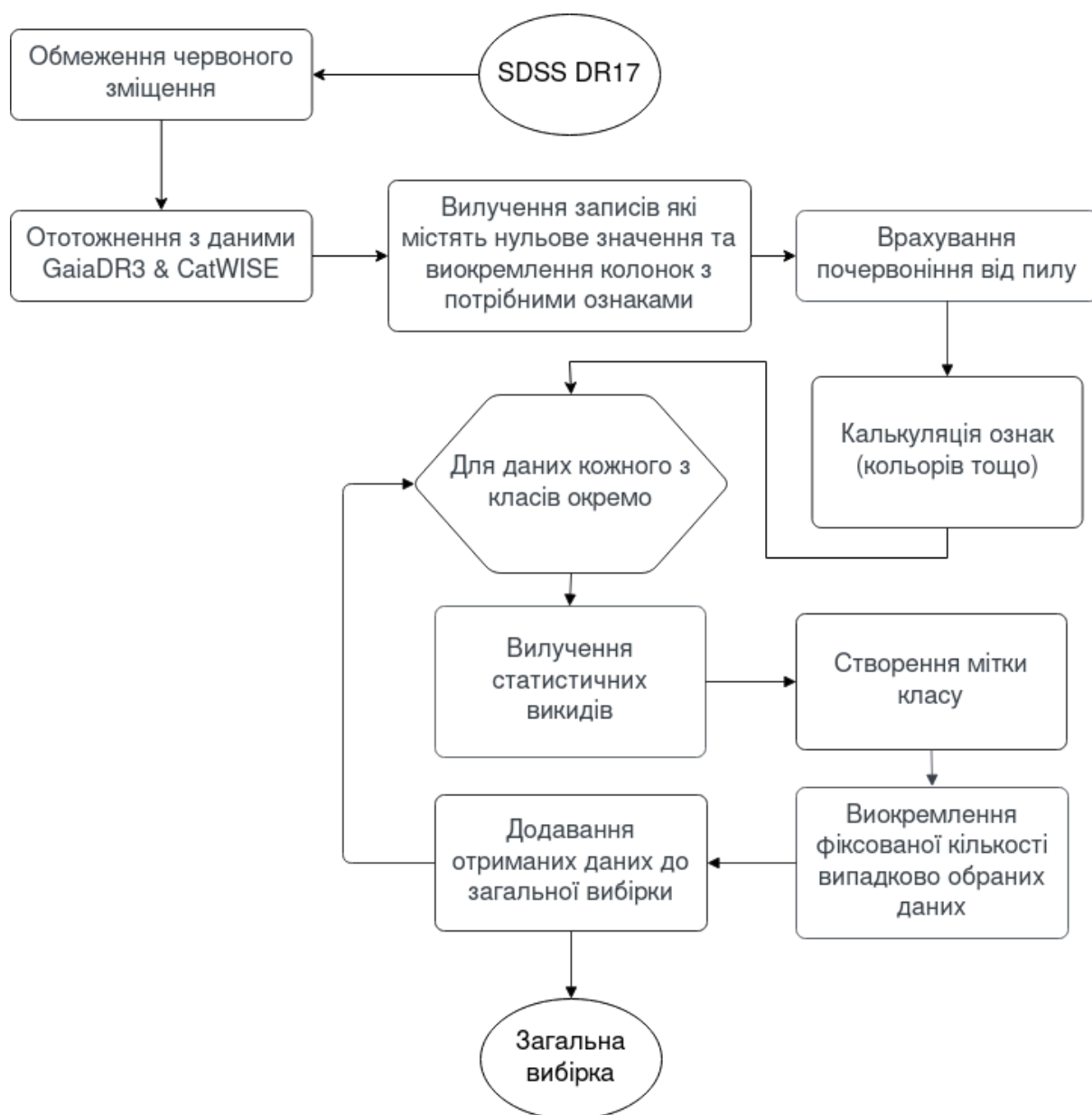


Рисунок А-1. Алгоритм отримання навчальної вибірки у вигляді блок-схеми.

Додаток С

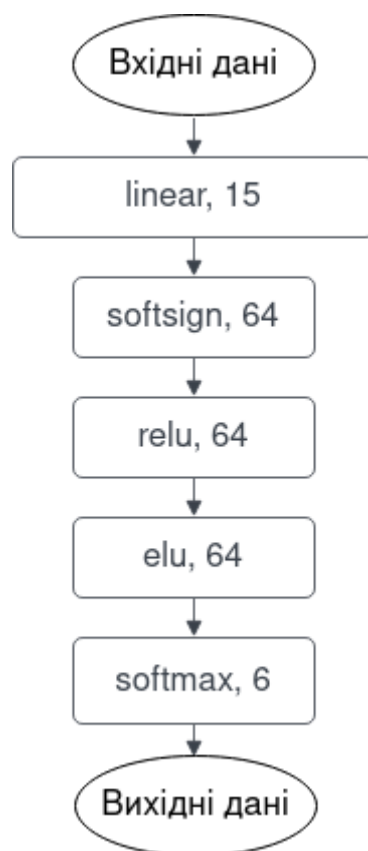


Рисунок А-2. Структура нейронної мережі у вигляді блок-схеми.

Додаток D

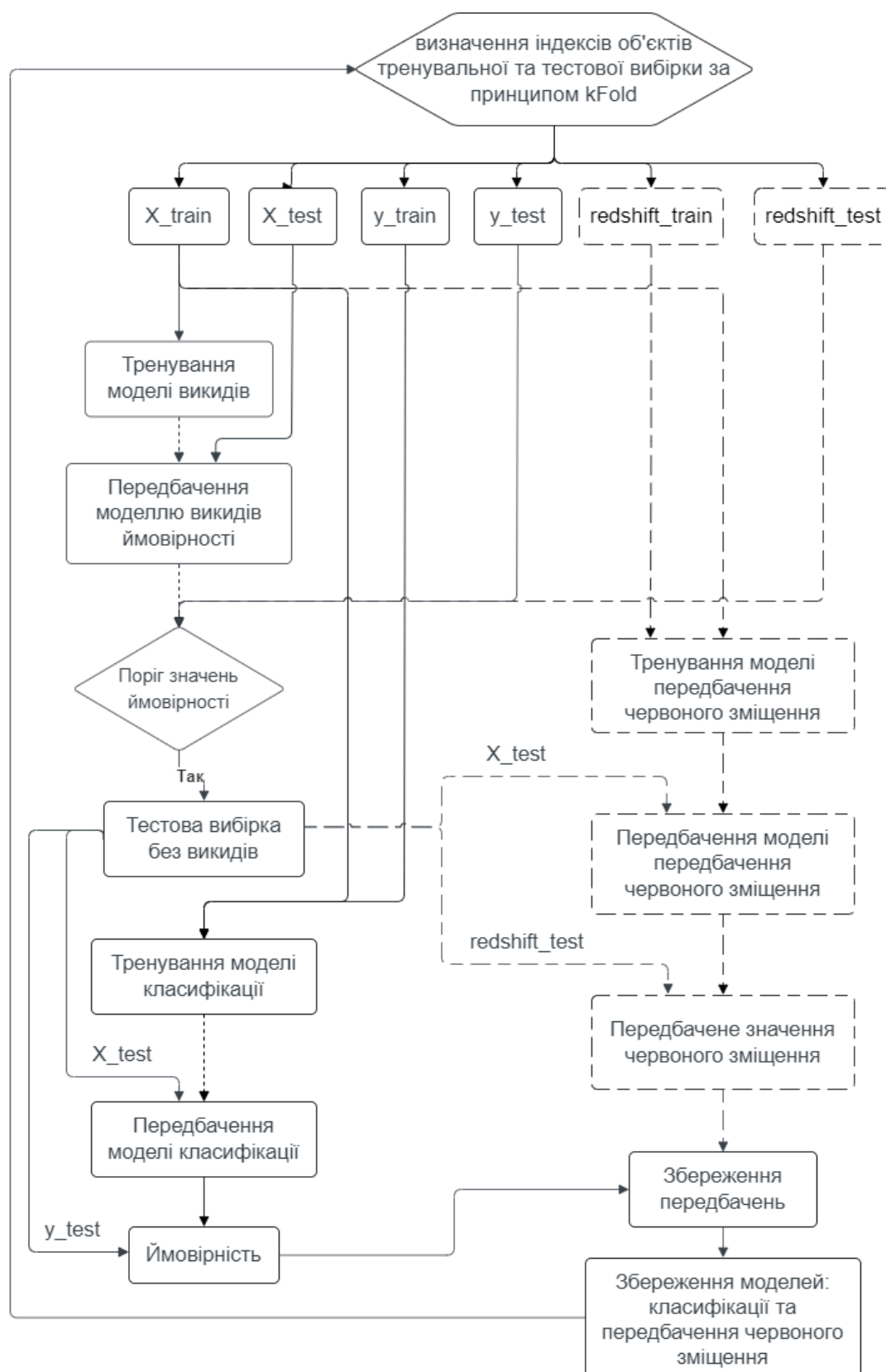


Рисунок А-3. Блок-схема алгоритму проведення навчання

Додаток Е

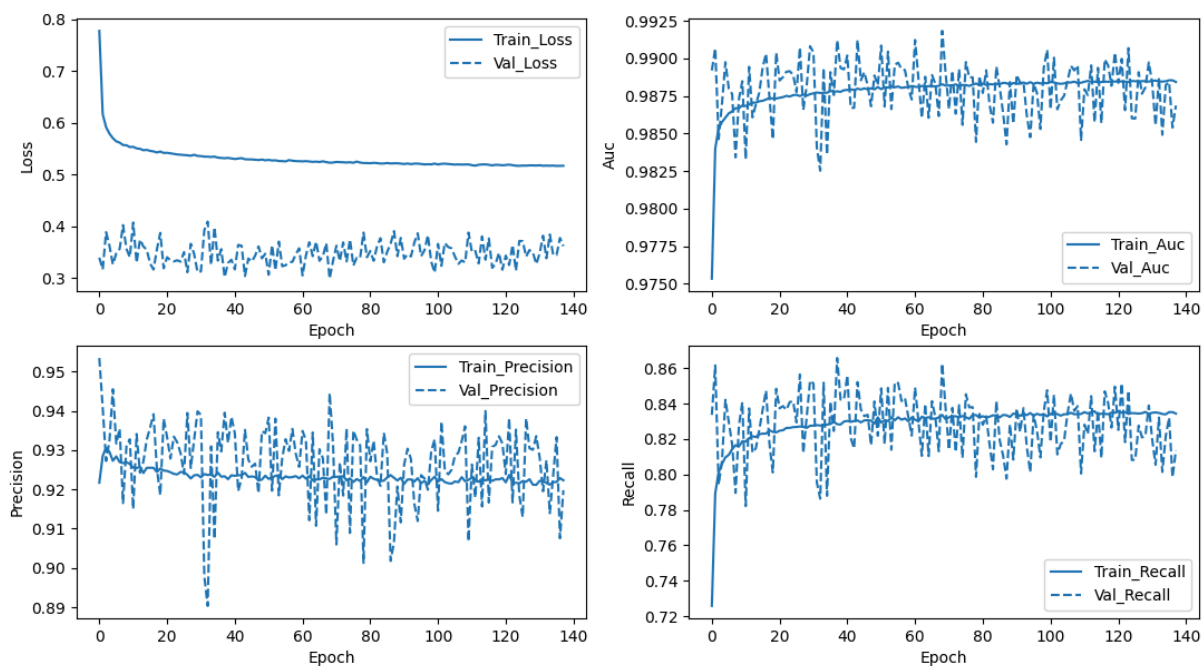


Рисунок А-4. Зміна втрат (loss) та метрик під час навчання на другій вибірці kfold з включенням валідаційної вибірки.

Додаток F

Назва класу	Precision, %	Recall, %	F1, %	bAcc, %
Квазари	99.94	97.59	98.75	98.78
АЯГ	13.07	43.93	20.15	69.63
СЗУГ	51.41	86.35	64.45	92.27
ЗУГ	42.32	65.39	51.38	78.88
Галактики	96.32	74.41	83.96	86.33
Зорі	99.97	99.97	99.97	99.98

Таблиця А-2. Значення метрик для головної вибірки

Назва класу	Precision, %	Recall, %	F1, %	bAcc, %
Квазари	99.88	98.65	99.26	99.29
АЯГ	11.84	53.77	19.35	73.66
СЗУГ	58.70	80.27	67.74	89.51
ЗУГ	48.58	61.71	54.36	78.08
Галактики	96.02	75.44	84.48	86.77
Зорі	99.97	99.98	99.97	99.99

Таблиця А-3. Середні значення передбачень моделі на підмножинах визначених за допомогою KFold.