NARODOWE CENTRUM BADAŃ JĄDROWYCH

ROZPRAWA DOKTORSKA

Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii

---

Autor:
Artem POLISZCZUK

Promotor:
Agnieszka POLLO
Promotor pomocniczy:
Aleksandra Solarz

Rozprawa doktorska przygotowana w

Zakładzie Astrofizyki
Departament Badań Podstawowych

NATIONAL CENTRE FOR NUCLEAR RESEARCH ŚWIERK

3 maja 2022

Oświadczenie o autorstwie

Ja, Artem POLISZCZUK, oświadczam, że niniejsza rozprawa doktorska zatytułowana *Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii* i wyniki w niej przedstawione są mojego autorstwa. Potwierdzam, że:

- Niniejsza praca została wykonana w całości lub w przeważającej części podczas ubiegania się o stopień naukowy w Narodowym Centrum Badań Jądrowych.
- Jeżeli jakakolwiek część tej pracy została wcześniej przedłożona w celu uzyskania stopnia naukowego lub innych kwalifikacji w Narodowym Centrum Badań Jądrowych lub innej instytucji, zostało to wyraźnie zaznaczone.
- Jeżeli korzystałem z opublikowanych prac innych osób, zawsze jest to wyraźnie zaznaczone.
- W przypadku cytowania z prac innych osób, zawsze podawane jest źródło. Z wyjątkiem takich cytatów, niniejsza praca jest w całości moim dziełem.
- Wymieniłem wszystkie główne źródła pomocy.
- Jeśli rozprawa opiera się na pracy wykonanej przeze mnie wspólnie z innymi, wyraźnie zaznaczam, co zostało wykonane przez innych, a co jest moim własnym wkładem.

Podpis: _____

Data: _____

NARODOWE CENTRUM BADAŃ JĄDROWYCH

Streszczenie

Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii

Artem POLISZCZUK

Celem przedstawionej rozprawy doktorskiej było opracowanie nowych technik selekcji aktywnych jąder galaktyk opartych na algorytmach uczenia maszynowego, które pozwoliłyby na efektywne przeszukiwanie dużych zbiorów danych fotometrycznych ze skutecznością niedostępną dla tradycyjnych metod selekcji. Zaproponowane metody pozwolą stworzyć wysokiej jakości katalogi aktywnych jąder galaktyk do zastosowań w astronomii pozagalaktycznej i kosmologii obserwacyjnej. Badania przedstawione w rozprawie pokazują, że możliwe jest stworzenie modelu opartego na algorytmach uczenia maszynowego, który jest w stanie naśladować selekcję aktywnych jąder galaktyk w zakresie średniej podczerwieni, używając jedynie danych fotometrycznych z zakresu optycznego i bliskiej podczerwieni. Taka metoda zapewnia wysoką efektywność selekcji charakterystyczną dla technik stosowanych w średniej podczerwieni. Jednocześnie nowa metoda pozwala uniknąć znacznego zmniejszenia rozmiaru katalogu wynikającego z wymogu pomiaru w średniej podczerwieni.

W pracy wykorzystano dane z głębokiego przeglądu nieba w polu AKARI NEP-Wide. Wprowadzono szereg rozwiązań niestosowanych dotąd w astronomii. Mechanizm naśladowania przez model selekcji opartej na średniej podczerwieni został uzyskany poprzez dwa podstawowe mechanizmy. Pierwszym z nich było pośrednie dostarczenie informacji na temat technik selekcji w średniej podczerwieni zawartych w konstrukcji próbki treningowej. Drugim było ograniczenie ryzyka ekstrapolacji w danych spoza próbki treningowej poprzez zastosowanie algorytmu Najmniejszego Wyznacznika Kowariancji. Zasosowana technika pozwoliła efektywnie ograniczyć obszar w wielowymiarowej przestrzeni cech do regionu pokrywanego przez dane treningowe.

Ponadto przeprowadzono badania nad efektywnością zastosowania różnych technik logiki rozmytej w selekcji aktywnych jąder galaktyk na podstawie różnych nadzorowanych algorytmów klasyfikacyjnych. Następnie zbadano efektywność zastosowania metod wyszukiwania anomalii oraz technik niskowymiarowej wizualizacji danych do znajdowania zanieczyszczeń katalogu wynikowego. Pozwoliło to zidentyfikować przypadki niepoprawnej fotometrycznej estymacji przesunięcia ku czerwieni, a także potencjalne grupy nieprawidłowo sklasyfikowanych obiektów.

Metody wprowadzone w pracy pozwalają na ominięcie trudności wynikających z ograniczeń instrumentów pomiarowych, a także umożliwiają precyzyjną kontrolę nad jakością katalogu wynikowego oraz rozpoznanie potencjanych źródeł zanieczyszczeń. Pozwala to na dopasowanie katalogu wynikowego do potrzeb konkretnych zastosowań i tworzy efektywny zestaw narzędzi dla współczesnej i przyszłej astrofizyki.

NARODOWE CENTRUM BADAŃ JĄDROWYCH

Abstract

Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii

Artem POLISZCZUK

The aim of this work was to develop new machine learning (ML) techniques for the automatic selection of active galactic nuclei (AGN), which would allow one to mine big photometric catalogs with effectiveness unreachable for traditional methods. The ML-based approach can then be used to create high-quality catalogs for astrophysical and observational cosmology purposes. This work shows it is possible to create a machine learning model which will be able to mimic mid-IR based photometric AGN selection using only optical and near-IR broadband photometry. The described model can preserve efficiency similar to mid-IR techniques. However, it allows one to obtain much larger catalogs due to the lack of mid-IR detection conditions.

Studies are performed on the data from the deep sky survey in the AKARI NEP-Wide field. This work introduces several methods which were not been used in astronomy before. The technique of mimicking the mid-IR selection by the ML model was based on two crucial mechanisms. The first one was connected to a specific construction of the training sample. It allows one to indirectly impose information about the mid-IR selection into the structure of the ML model. The second mechanism was based on the avoidance of extrapolation risk. It was achieved by limiting the shape of the generalization sample to the shape of training sample via the Minimum Covariance Determinant estimator algorithm. This way, a better control of the model performance and a higher quality of the AGN candidates catalog was achieved.

Additionally, this work presents an in-depth study on the effectiveness of various fuzzy logic strategies for an AGN selection. For this purpose, a large set of supervised classification algorithms was used. Finally, the reader will find a study on the effectiveness of outlier detection methods combined with low-dimensional embedding visualization techniques to detect and remove various contamination sources from the catalog. This way, cases of wrong photometric redshift estimation and misclassified groups of sources were identified.

Methods developed in this work overcome detector limitations and allow one to precisely control the quality of the final source catalog. Moreover, a user of this method can identify different sources of catalog contamination. Presented techniques allow one to match catalog properties to specific scientific needs, making them an effective tool for modern astrophysics.

Podziękowania

Chciałbym podziękować moim promotorom, prof. Agnieszce Pollo i dr Aleksandrze Solarz, które pomogły mi rozwinąć się naukowo i udzieliły mi wsparcia w trakcie studiów doktoranckich. Chciałbym również podziękować moim kolegom dr Katarzynie Małek, prof. Matthew Malkanowi, prof. Tomotsugu Goto i dr. Seong Jin Kimowi oraz członkom zespołu North Ecliptic Pole za wnikliwe komentarze i cenne dyskusje naukowe.

Spis treści

Spis rysunków

Spis tablic

mojemu dziadkowi

1

Wstęp

Gwałtowny rozwój astrofizyki i kosmologii obserwacyjnej w XXI wieku doprowadził do istotnego rozwoju nowych metod analizy statystycznej w tych dziedzinach. Znaczna część tej rewolucji metodologicznej jest związana z szybko rozwijającymi się dziedzinami *big data* i uczenia maszynowego, ang. *machine learning* (ML). Obecnie algorytmy uczenia maszynowego są z dużym powodzeniem stosowane do różnych zadań w astrofizyce pozagalaktycznej i kosmologii. Są to m.in. fundamentalne problemy klasyfikacyjne związane z tworzeniem katalogów (Clarke, A. O. i in., 2020), problemy estymacji parametrów astrofizycznych i kosmologicznych (D'Isanto i Polsterer, 2018; Henghes i in., 2021; Pan i in., 2020) albo niskopoziomowe potoki przetwarzania danych z teleskopów oparte na ML, służące do rekonstrukcji i klasyfikacji sygnałów (Narayan i in., 2018). Szybkie pojawianie się nowych algorytmów i wykładniczo rosnąca ilość danych astronomicznych pozwalają uznać, że metody ML stają się nieodłączną częścią współczesnej astronomii (Sen i in., 2022).

Podstawowym celem przedstawionej rozprawy było stworzenie opartej na ML metody fotometrycznej selekcji Aktywnych Jąder Galaktyk, ang. *Active Galactic Nuclei* (AGN), która naśladowałaby szerokopasmową fotometryczną metodę selekcji w średniej podczerwieni, wykorzystując jedynie szerokopasmową fotometrię optyczną i w bliskiej podczerwieni. Znaczenie takiej techniki wynika z kompromisu, jaki wiąże się z selekcją AGN-ów w zakresie średniej podczerwieni. Zakres średniej podczerwieni widma elektromagnetycznego zawiera istotną część informacji o emisji AGN-ów, pozwalając na uzyskanie katalogów charakteryzujących się zarówno wysoką czystością, jak i kompletnością (Padovani i in., 2017). Co więcej, selekcja oparta na średniej podczerwieni jest wrażliwa na określony etap wydajności akrecji AGN-ów, co wiąże te obiekty z procesami ewolucyjnymi galaktyk-gospodarzy (ang. *host galaxies*) oraz z ich umiejscowieniem w strukturze wielkoskalowej. Te cenne cechy widma pozwalające na selekcję AGN-ów w średniej podczerwieni mają jednak istotne ograniczenia. Głównym z nich jest zdecydowanie mniejszy rozmiar katalogów opartych na średniej podczerwieni, w porównaniu z odpowiednikami stworzonymi w paśmie optycznym i bliskiej podczerwieni. Ta właściwość wynika z dwóch przyczyn. Pierwszą z nich jest niska rozdzielczość detektorów średniej podczerwieni. Drugą jest sama natura obserwacji w średniej podczerwieni. Ta część widma elektromagnetycznego jest blokowana przez ziemską atmosferę. Dlatego wykonanie obserwacji w średniej podczerwieni jest możliwe tylko przy wykorzstaniu teleskopów kosmicznych. Taki teleskop satelitarny musi być dodatkowo sztucznie chłodzony, aby dane z detektora średniej podczerwieni nie były zaszumione emisją termiczną pochodzącą z elektroniki teleskopu. Kriogeniczna faza obserwacji teleskopu jest ograniczona w czasie

(patrz np. Murakami i in., 2007). W związku z powyższymi problemami, połączenie niskiej rozdzielczości instrumentu i ograniczonego czasu pracy sprawia, że uzyskanie dużych katalogów w zakresie średniej podczerwieni jest zadaniem bardzo trudnym.

Podstawowym pomysłem, który pozwolił przezwyciężyć trudności napotykane w selekcji opartej na średniej podczerwieni, było pośrednie wprowadzenie informacji o właściwościach obiektów w zakresie średniej podczerwieni do struktury modelu klasyfikacyjnego. Tak skonstruowany model może skutecznie poszukiwać kandydatów na AGN o charakterystycznych właściwościach w średniej podczerwieni, wykorzystując jedynie informacje z danych optycznych i bliskiej podczerwieni. Takie podejście pozwala przezwyciężyć warunek detekcji w średniej podczerwieni i uzyskać katalog kandydatów na AGN-y o podobnych własnościach, ale zawierający znacznie więcej obiektów. Autorska metoda przedstawiona w tej pracy opiera się na badaniach opublikowanych w dwóch pracach (Poliszczuk i in., 2019; Poliszczuk i in., 2021). Publikacje te zawierają badania dotyczące różnych technik ML zastosowanych do danych podczerwonych zebranych przez kosmiczny teleskop AKARI w rejonie północnego bieguna ekliptycznego, ang. *north ecliptic pole* (NEP). Pierwsza praca (Poliszczuk i in., 2019) była wstępnym studium metod opartych na ML dla połączonej selekcji AGN-ów w bliskiej i średniej podczerwieni. W tym celu użyto specyficznego algorytmu ML o nazwie *maszyna wektorów nośnych* do testowania różnych problemów klasyfikacji. Jednym z nich było zastosowanie logiki rozmytej w strukturze algorytmu klasyfikacyjnego. Pozwoliło to na zróżnicowanie wpływu różnych obiektów ze zbioru treningowego na klasyfikację na podstawie ich specyficznych właściwości, takich jak precyzja pomiaru. Według naszej najlepszej wiedzy, tego typu fizycznie umotywowana modyfikacja modelu za pomocą logiki rozmytej nie była nigdy wcześniej stosowana w astronomii. Kolejnym zadaniem było sprawdzenie, jak ekstrapolacja poza obszar wyznaczony przez dane treningowe wpływa na działanie klasyfikatora. Wyniki te zostały potraktowane jako badania wstępne. Pozwoliły one na stworzenie nowatorskiej metody naśladowania selekcji średniej podczerwieni przy użyciu danych optycznych i bliskiej podczerwieni. Podejście to zostało opisane w drugiej publikacji (Poliszczuk i in., 2021), która również jest podstawą niniejszej rozprawy.

W porównaniu do publikacji (Poliszczuk i in., 2021), metody i wyniki przedstawione w niniejszej rozprawie zostały zmodyfikowane i wzbogacone. Po pierwsze, rozprawa zawiera dogłębne omówienie różnych strategii logiki rozmytej oraz porównanie ich wpływu na klasyfikację z różnymi typami algorytmów klasyfikacji nadzorowanej. Ta część w rozprawie jest znacząco rozwinięta w porównaniu z oryginalną pracą (Poliszczuk i in., 2021). Pozwala ona lepiej zrozumieć, w jaki sposób nowe podejście może modyfikować wydajność modelu i jak wpływa na właściwości katalogu wynikowego. Drugą istotną modyfikacją są dodatkowe badania nad nienadzorowanymi technikami wykrywania obserwacji odstających (ang. *outliers*) Metoody te zostały wykorzystane do lepszej kontroli własności katalogu AGN-ów oraz do usunięcia szczególnie problematycznych źródeł zanieczyszczeń. Dzięki temu otrzymany katalog AGN-ów może odpowiadać potrzebom różnych zastosowań. Ta część opracowanego potoku uczenia maszynowego, która koncentruje się na metodach detekcji obserwacji odstających, będzie głównym tematem przygotowywanej obecnie przez autora publikacji (Poliszczuk et al., *in prep*).

Rozprawa jest zorganizowana w następujący sposób. Rozdział 2, zawiera opis Zunifikowanego Modelu AGN oraz procesów fizycznych leżących u podstaw rozkładu energii widmowej AGN-ów. Ponadto omawia wpływ tych procesów na różne metody selekcji AGN-ów oraz niektóre aspekty związku AGN-ów z ewolucją galaktyk i kosmologią obserwacyjną. Rozdział 3 zawiera opis danych. W pierwszej części

czytelnik znajdzie ogólny opis danych panchromatycznych z Północnego Bieguna Ekliptycznego. Druga część zawiera omówienie wielozakresowego katalogu AKARI NEP-Wide, użytego do trenowania modeli ML i uzyskania ostatecznego katalogu kandydatów na AGN. Wreszcie, rozdział ten zawiera informacje o przygotowaniu i własnościach próbek treningowych i generalizacyjnych używanych w tej pracy. W szczególności w rozdziale 3.3.2 opisano, jak zmniejszyć ryzyko ekstrapolacji podczas przewidywania na nieoznakowanych danych przy użyciu algorytmu najmniejszego wyznacznika kowariancji, ang. *Minimum Covariance Determinant Algorithm* (MCD). W rozdziale 4 omówiono metody uczenia maszynowego zastosowane w tej pracy wraz z metrykami oceny wydajności, które zostały wykorzystane do trenowania klasyfikatorów, jak również do porównania ich z selekcją w średniej podczerwieni. Rozdział 5 zawiera dyskusję uzyskanych wyników. W szczególności omówiono skuteczność różnych algorytmów ML oraz wpływ na klasyfikację różnych strategii ważenia opartych na klasach i logice rozmytej. Ponadto czytelnik znajdzie tam dyskusję na temat własności otrzymanego katalogu AGN-ów i porównanie go z katalogiem utworzonym metodami selekcji w zakresie średniej podczerwieni. Ponadto przeprowadzono eksperyment w celu sprawdzenia możliwości dodatkowego zwiększenia skuteczności klasyfikacji, aby przezwyciężyć problemy występujące zarówno w selekcji opartej na średniej podczerwieni, jak i w metodzie opartej na ML. W ostatniej części pracy przedstawiono różne metody wykrywania obiektów odstających. Metody te są wykorzystywane do wyszukiwania obiektów z katastrofalnymi błędami estymacji fotometrycznych przesunięć ku czerwieni oraz zanieczyszczeniami katalogowymi z klasy próbek galaktyk. W rozdziale 6 przedstawiono podsumowanie uzyskanych wyników. Dodatek A zawiera informacje o oprogramowaniu użytym w niniejszej pracy. W dodatku B czytelnik znajdzie dodatkowe dane opisujące ocenę wydajności zastosowanych metod.

2

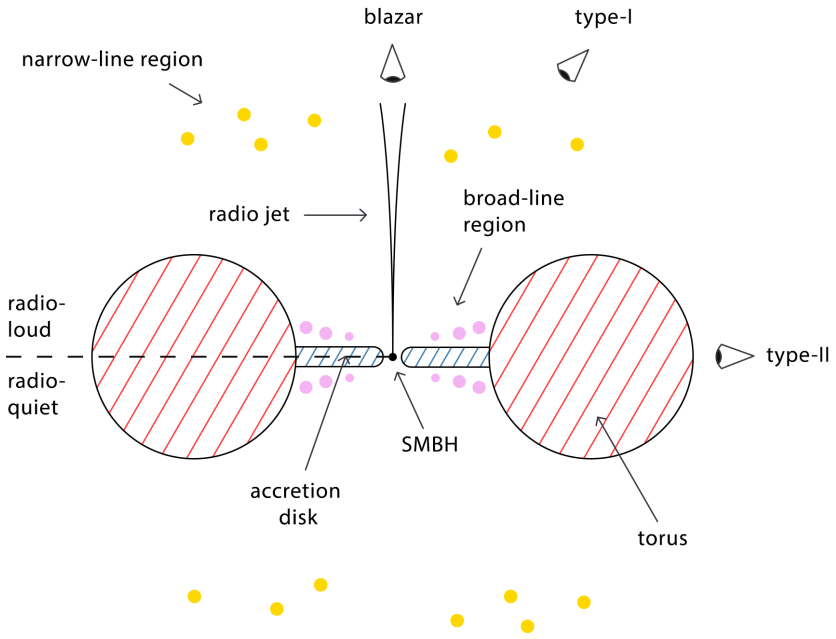
Aktywne jądra galaktyk

2.1 Zunifikowany model aktywnych jąder galaktyk

Galaktyki aktywne definiuje się jako galaktyki w okresie intensywnej akrecji materii na ich centralną supermasywną czarną dziurę. W centrach większości galaktyk znajduje sie supermasywna czarna dziura, ang. *supermassive black hole* (SMBH). W galaktykach aktywnych materia napływająca na SMBH wypromieniowuje dużą ilość energii w zakresie promieniowania rentgenowskiego i optycznego/ultrafioletowego (UV) widma elektromagnetycznego, która jest następnie reemitowana na dłuższych falach w bliskim sąsiedztwie SMBH. Ten świecący region w pobliżu SMBH nazywany jest aktywnym jądrem galaktyki (AGN). W tej pracy będziemy używać terminów galaktyka aktywna i AGN zamiennie, a galaktykę, w której centrum znajduje się aktywne jądro, będziemy określać mianem "gospodarza"(ang. *host galaxy*).

W dotychczasowych badaniach zidentyfikowano dużą liczbę klas AGN, które różnią się względną siłą emisji w różnych częściach widma elektromagnetycznego, jak również własnościami spektroskopowymi czy obecnością silnych relatywistycznych dżetów radiowych (obszerny przegląd poszczególnych klas można znaleźć w Padovani i in., 2017). Wiele z tych własności można wyjaśnić w kategoriach *zunifikowanego modelu AGN*. W swojej podstawowej formie, opisanej w pracy Antonucci (1993), odmienne właściwości klas AGN-ów zostały wyjaśnione za pomocą zmienności trzech parametrów: kąta nachylenia AGN względem linii widzenia (ang. *line of sight*, LOS), jasności AGN oraz współczynnika zakrycia AGN. Wizualizacja podstawowego zunifikowanego modelu jest pokazana na rys. 2.1.

Struktura AGN jest osiowosymetryczna i dzieli się na kilka głównych regionów. Centralna czarna dziura otoczona jest dyskiem akrecyjnym o promieniu poniżej 1 pc, składającym się z całkowicie zjonizowanej, wolnej od pyłu materii. Obłoki o dużej gęstości, składające się ze zjonizowanego, pozbawionego pyłu gazu, tworzą obszar szerokich linii emisyjnych (z ang. *broad line region*, BLR) w odległości od $\simeq$1 pc do 10^{3-5} promieni grawitacyjnych od centralnej czarnej dziury. *Torus* umieszczony poza BLR jest mieszaniną gazu i pyłu o częściowo zbitej strukturze. Wewnętrzny promień torusa jest w przybliżeniu wyznaczony przez *promień sublimacji*. Promień ten jest odległością, w której temperatura spada do $\sim$2000 K, czyli do temperatury, powyżej której ziarna pyłu zaczynają wyparowywać (Netzer, 2015; Jones, Lambourne i Serjeant, 2015). Region prostopadły do płaszczyzny torusa tworzy stożek jonizacyjny (ang. *ionization cone*), w którym lekko zjonizowane obłoki gazu znajdujące się w odległości setek parseków od płaszczyzny torusa tworzą obszar wąskich linii emisyjnych (ang. *narrow-line region*, NLR). Dodatkowo, wzdłuż osi centralnej, prostopadłej do



RYSUNEK 2.1: Poglądowy schemat przedstawiający zunifikowany model aktywnych jąder galaktyk zaproponowany w Antonucci (1993).

płaszczyzny torusa, może występować radiowy dżet relatywistyczny. Obecność lub brak dżetu leży u podstaw rozróżnienia AGN-ów na radiowo-głośne i radiowo-ciche (ang. *radio loud* i *radio quiet*). Prowadzi to do uzupełniającej klasyfikacji radiowej AGN-ów, która wykracza poza zakres tej pracy. Zachęcamy czytelnika do sięgnięcia po inne publikacje przeglądowe, takie jak Urry i Padovani (1995) w celu uzyskania dodatkowych informacji na temat klasyfikacji radiowej AGN-ów.

Nachylenie torusa w stosunku do linii obserwacji daje podstawowy podział na dwie klasy AGN. AGN typu pierwszego (ang. *type-I AGN*) charakteryzuje się tym, że jego stożek jonizacji skierowany jest w stronę obserwatora i nie jest przesłonięty przez zapylony torus. W tym przypadku widmo w zakresie od ultrafioletu do bliskiej podczerwieni pokazuje szerokie dozwolone i pół-wzbronione linie emisyjne z typowymi prędkościami gazu rzędu 1000-20 000 km s^{-1} pochodzącego z BLR oraz z jasnego, niegwiazdowego składnika centralnego. Ponadto w widmach AGN-ów typu I (z wyjątkiem niektórych obiektów o dużej jasności) występują wzbronione wąskie linie emisyjne [1], które powstały w NLR przy typowych prędkościach gazu rzędu 300-1000 km s^{-1}. Mimo że linie te określane są jako "wąskie", są one nadal szerokie w porównaniu z liniami emisyjnymi występującymi w widmach galaktyk. Na podstawie jasności źródła, obiekty typu I można podzielić na *galaktyki Seyferta typu I* (charakteryzujące się mniejszą jasnością) i *kwazary* (z ang. *quasi-stellar object*, QSO). Dodatkowo, oddzielna klasa AGN-ów z relatywistycznym dżetem skierowanym w stronę obserwatora jest nazywana *blazarami*. AGN typu drugiego (ang. *type-II AGN*) występuje, gdy pyłowy torus znajduje się na linii między obserwatorem i BLR. Taka sytuacja sprawia, że emisja z BLR staje się niewidoczna dla obserwatora. W

[1] Widmowe linie emisyjne w astronomii można podzielić ze względu na gęstość gazu w miejscu ich powstawania. W przeciwieństwie do linii dozwolonych, linie wzbronione powstają w obszarach o bardzo małej gęstości gazu. Linie te nie mogą powstawać w gęstszych środowiskach, ponieważ są związane z długo żyjącymi stanami wzbudzonymi, które w gęstym środowisku ulegają deekscytacji w wyniku zderzeń (Jones, Lambourne i Serjeant, 2015).

tym przypadku widma w zakresie od ultrafioletu do bliskiej podczerwieni pokazują jedynie wąskie linie emisyjne o niskiej jonizacji, które pochodzą z NLR. Szerokie linie emisyjne są nieobecne w widmach AGN-ów typu II. Typ drugi można dalej podzielić na dwie podgrupy. Pierwsza z podgrup jest nazywana *ukrytym typem I* (ang. *hidden type-I AGN*). W tym przypadku obecność szerokich linii emisyjnych przesłoniętych przez torus pyłowy jest widoczna w świetle spolaryzowanym. Podobnie jak typ I, ukryty typ I jest dzielony ze względu na jasność obiektu na *galaktyki Seyferta typu II* oraz *kwazary typu II*. Druga podgrupa jest nazywana *prawdziwym typem II* (ang. *true type-II AGN*). AGN-y prawdziwego typu II nie wykazują żadnych śladów obecności szerokich linii widmowych. Obiekty te, o typowo niższej jasności w porównaniu z ukrytym typem I, stanowią około 30% wszystkich AGN-ów typu II w lokalnym Wszechświecie (Brightman i Nandra, 2011; Merloni i in., 2014a).

Oprócz powyżej wymienionych głównych klas istnieją obiekty o mieszanych właściwościach normalnych i aktywnych galaktyk. Zalicza się do nich AGN-y zdominowane przez galaktykę-gospodarza (ang. *host-dominated AGN*, Kauffmann i in., 2003b), słabo-zjonizowane centralne obszary linii emisyjnych, (ang. *low-ionization nuclear emission-line region* lub LINER, Ho, 2008), AGN-y o niskiej jasności optycznej (ang. *low-luminosity optically dull AGNs*, Trump i in., 2009) lub kwazary o słabych liniach emisyjnych (ang. *weak line quasars*, Meusinger i Balafkan, 2014). Obiekty te są często interpretowane jako reprezentacje pośrednich stadiów ewolucyjnych pomiędzy galaktykami aktywnymi i normalnymi lub jako obiekty z niewystarczającym przepływem akrecyjnym w pobliżu SMBH.

Pełne wyjaśnienie mechanizmów fizycznych odpowiedzialnych za zachowanie AGN-ów o słabych liniach emisyjnych wciąż pozostaje kwestią otwartą (patrz dyskusja w Trump i in., 2009 i odnośniki tamże). Inną zagadkową klasą obiektów stanowią niedawno odkryte AGN-y o zmiennym wyglądzie (ang. *changing look AGN* lub CL AGN, LaMassa i in., 2015a; Charlton i in., 2019), które mogą wykazywać własności AGN-ów zarówno typu I jak i typu II. Dyskusje na temat możliwych wyjaśnień własności CL AGN można znaleźć w LaMassa i in. (2015b), Stern i in. (2018) oraz Dodd i in. (2021). Dalsza dyskusja na temat właściwości tych obiektów wykracza poza zakres niniejszej pracy, dlatego zachęcamy czytelników do sięgnięcia do cytowanych źródeł. Problemy związane z modelem zunifikowanym, takie jak trudności z wyjaśnieniem własności WLQ i CL AGN (Dodd i in., 2021), niesatysfakcjonujące wyniki przewidywania dotyczące ewolucji SMBH i modelowania łączących się gospodarzy, obserwacyjne przesłanki wskazujące na znacznie bardziej skomplikowaną strukturę torusa, a także złożony związek pomiędzy oddziaływaniem torusa z galaktyką-gospodarzem świadczą o niekompletności podstawowego schematu unifikacji AGN. Dyskusja na temat problemów związanych z modelem zunifikowanym oraz możliwych sposobów ich przezwyciężenia jest szczegółowo przedstawiona w artykułach Netzer (2015) oraz Heckman i Best (2014).

Poza zunifikowanym modelem AGN istnieje jeszcze jeden sposób interpretacji własności AGN i ich klasyfikacji na podstawie mechanizmu, który dominuje w transporcie energii w pobliżu centralnej czarnej dziury. Aby lepiej zrozumieć tę dodatkową klasyfikację, musimy wprowadzić pojęcie *granicy Eddingtona* (lub *jasności Eddingtona*). Granica Eddingtona definiuje maksymalną jasność, jaką może osiągnąć obiekt w stanie równowagi hydrostatycznej, tzn. gdy siła promieniowania działająca na zewnątrz jest równoważona przez siłę grawitacji działającą do wewnątrz (Peterson, 1997). Zatem, gdy ciśnienie promieniowania przewyższa siłę grawitacji na dowolnych odległościach od SMBH, gaz otaczający źródło zostanie zdmuchnięty przez występujące wiatry skierowane na zewnątrz. Dla centralnej czarnej dziury o masie M_{BH}, granica Eddingtona L_{Edd} jest dana przez Shapiro i Teukolsky (1983):

$$L_{Edd} = 1.3 \times \frac{M_{BH}}{M_\odot} \text{ erg s}^{-1}. \tag{2.1}$$

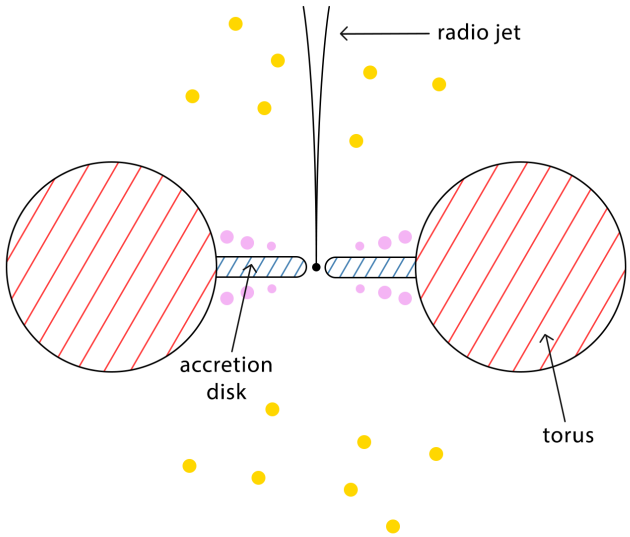
Inną istotną własnością obiektu ściśle związaną z granicą Eddingtona jest *współczynnik Eddingtona*, czyli stosunek jasności bolometrycznej obiektu do jego granicy Eddingtona: L/L_{Edd}.

Klasyfikacja oparta na dominującym mechanizmie wypływu energii w pobliżu SMBH pozwala wyróżnić dwie główne populacje lub tryby AGN-ów. Pierwsza klasa obiektów określana jest jako tryb radiacyjny (ang. *radiative mode*), często nazywany też trybem kwazarowym lub wiatrowym. Tryb ten występuje w AGN o dużej jasności, w których SMBH jest otoczona przez geometrycznie cienki i optycznie gruby dysk akrecyjny. Obiekty te charakteryzują się wydajnym przepływem akrecyjnym i wysokimi współczynnikami Eddingtona ($L/L_{Edd} > 0.01$). Druga klasa jest określana jako tryb kinetyczny (ang. *kinetic mode*), zwany też trybem dżetowym. Obiekty te charakteryzują się nieefektywną akrecją, której tempo nie przekracza 1% granicy Eddingtona ($L/L_{Edd} < 0.01$). W przypadku AGN-ów działających w trybie kinetycznym, geometrycznie cienki dysk akrecyjny jest obcięty w regionach wewnętrznych, a większość energii jest przekazywana w postaci kinetycznej przez duże radiowe dżety (należy zauważyć, że AGN-y pracujące w trybie radiacyjnym również mogą niekiedy posiadać struktury dżetowe). Wizualne porównanie ogólnej struktury obu trybów pokazane jest na Rys. 2.2. Niniejsza praca skupia się na AGN-ach pracujących w trybie radiacyjnym. Dalsza, dogłębna dyskusja na temat natury obu trybów i wpływu ich obecności na własności galaktyk-gospodarzy jest opisana w wielu publikacjach przeglądowych, takich jak Kormendy i Ho (2013) oraz Yuan i Narayan (2014). Klasyfikacja AGN na podstawie trybów ma kluczowe znaczenie dla zrozumienia związku między aktywnością AGN a ewolucją galaktyk oraz roli, jaką AGN-y odgrywają w kosmologii obserwacyjnej. Zagadnienia te są szerzej omówione w rozdziale 2.3.

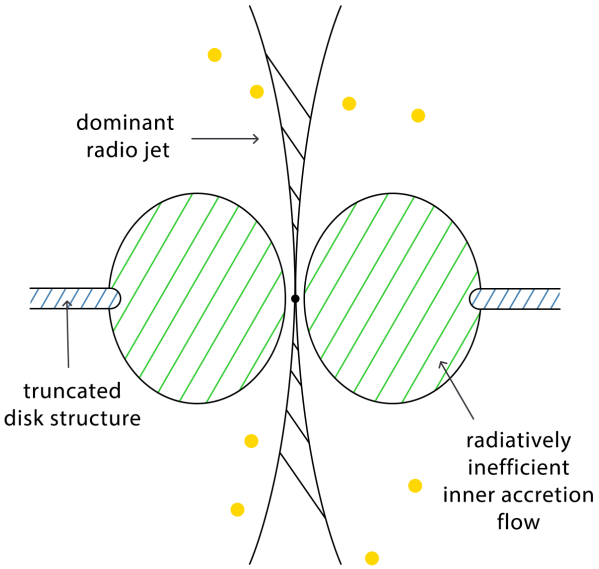
2.2 Wielozakresowa emisja aktywnych jąder galaktyk i odpowiadające jej metody selekcji

Obserwacje AGN w różnych częściach widma elektromagnetycznego uwypuklają emisję z określonych części struktury AGN i są ściśle związane z właściwościami selekcji AGN-ów w katalogach fotometrycznych. W tym rozdziale dokonano przeglądu najważniejszych mechanizmów wpływających na kształt widmowego rozkładu energii AGN-ów (ang. *spectral energy distribution*, SED) w zakresie od promieniowania rentgenowskiego do podczerwieni oraz opisano, jak są one wykorzystywane do odzyskiwania próbek AGN z katalogów przeglądów nieba. W rozdziale tym nie omówiono własności AGN-ów w zakresie promieniowania gamma i radiowego, ponieważ są to szerokie i złożone zagadnienia, niezwiązane z problemem naukowym tej pracy. Duża część opisu metod poszukiwania AGN-ów i związanych z nimi efektami selekcji w różnych częściach widma została zainspirowana przez Padovani i in. (2017) oraz Donley i in. (2012), gdzie czytelnik może znaleźć bardziej dogłębną analizę tych problemów.

Pasmo promieniowania rentgenowskiego zdefiniowane jako zakres energii 0.2–200 keV pozwala przeprowadzić efektywną selekcję AGN-ów o wysokiej kompletności i czystości. Główny wkład do emisji AGN w zakresie rentgenowskim pochodzi z odwrotnego rozpraszania Comptona fotonów dysku akrecyjnego w obszarze korony rentgenowskiej w pobliżu SMBH (Gilfanov i Merloni, 2014). Dodatkową składową promieniowania rentgenowskiego AGN stanowi emisja termiczna wewnątrz dysku



(A)



(B)

RYSUNEK 2.2: Klasyfikacja oparta na dominującym mechanizmie wypływu energii w pobliżu SMBH. *Panel a:* Tryb radiacyjny. SMBH jest otoczona przez geometrycznie cienki i optycznie gruby dysk akrecyjny. AGN-y pracujące w trybie radiacyjnym charakteryzują się wydajnym przepływem akrecyjnym i wysokimi współczynnikami Eddingtona. *Panel b:* Tryb kinetyczny. Obiekty te charakteryzują się dominującym dżetem radiowym i niewystarczającą akrecją materii do SMBH oraz niskimi wartościami współczynnika Eddingtona. Grifiki inspirowane pracą Heckman i Best (2014).

akrecyjnego (Sobolewska, Siemiginowska i Zycki, 2004) jak również promieniowanie z obszarów relatywistycznych dżetów (Harris i Krawczynski, 2006).

Pomimo szeregu mechanizmów odpowiadających za emisję w paśmie rentgenowskim AGN-ów, promieniowanie w danym zakresie jest silnie skorelowane z promieniowaniem dysku akrecyjnego i wykazuje izotropowe właściwości w całej populacji AGN-ów (Lusso i Risaliti, 2016). Uniwersalne własności emisji rentgenowskiej AGN w połączeniu z umiarkowaną odpornością emisji w tym zakresie na przesłanianie przez pył (zwłaszcza w paśmie twardego promieniowania rentgenowskiego) i niewielkim zanieczyszczeniem ze strony galaktyki-gospodarza sprawiają, że obserwacje rentgenowskie dobrze nadają się do selekcji AGN-ów. Podstawowe podejście do selekcji AGN-ów w paśmie rentgenowskim polega na nałożeniu dolnego ograniczenia jasności w zakresie rentgenowskim (L_X) na katalog zwartych, punktowych źródeł. W tym przypadku można wyróżnić kwazary jako obiekty o jasności $L_X > 10^{44}$ erg s^{-1}, galaktyki Seyferta znajdujące się w przedziale $L_X = 10^{42} - 10^{44}$ erg s^{-1} oraz AGN-y o małej jasności $L_X < 10^{42}$ erg s^{-1} (Padovani i in., 2017). Potencjalne zanieczyszczenie katalogu kandydatów na AGN może pochodzić z gromad galaktyk o dużym przesunięciu ku czerwieni oraz bardzo zwartych grup galaktyk. Ma to miejsce w szczególności wtedy, gdy obiekty są badane w pasmach rentgenowskich o niższych energiach, które są czułe na rentgenowskie promieniowanie termiczne (Bulbul i in., 2021). Dodatkowe zanieczyszczenie może wystąpić w przypadku źródeł o niskiej jasności, gdzie emisja rentgenowska z rentgenowskich układów podwójnych znajdujących sie w galaktyce-gospodarzu może mieć znaczący wkład do całkowitego strumienia promieniowania rentgenowskiego (Fabbiano, 2006).

W rentgenowskich metodach poszukiwania AGN-ów występuje kilka źródeł statystycznego obciążenia selekcji (ang. *selection bias*), w których główną rolę odgrywa zależna od energii absorpcja promieniowania rentgenowskiego. Emisja miękkiego promieniowania rentgenowskiego ma wyższą wydajność absorpcji niż emisja twardego promieniowania rentgenowskiego (Wilms, Allen i McCray, 2000). Zjawisko to przekłada się na statystyczne obciążenie selekcji związane z przesunięciem ku czerwieni (ang. *redshift bias*), gdzie źródła o dużym przesunięciu ku czerwieni są badane przy wyższej energii w układzie spoczynkowym, a zatem wykazują mniejszy efekt zaabsorbowanego strumienia promieniowania rentgenowskiego. Dodatkowy problem stanowi klasa AGN-ów znana jako comptonowsko-nieprzezroczyste (ang. *Compton-thick AGNs*, CT AGN). Te obiekty są zdefiniowane jako wysoce przesłonięte źródła, w których gęstość słupa materii przekracza wartość przekroju czynnego odwrotnego rozpraszania Thomsona. Są one szczególnie trudne do wykrycia w paśmie rentgenowskim i często wymykają się metodom selekcji opartym wyłącznie na danym zakresie promieniowania elektromagnetycznego (Comastri, 2004; Comastri i Fiore, 2004). Odzyskiwanie takich obiektów najczęściej wymaga dodatkowych wielozakresowych obserwacji. Istnieje dobra zgodność pomiędzy absorpcją w pasmach optycznym i ultrafioletowym a absorpcją w paśmie rentgenowskim: zdecydowana większość AGN-ów typu I nie wykazuje przesłonięcia w zakresie rentgenowskim, podczas gdy obiekty typu II są identyfikowane w paśmie rentgenowskim głównie jako CT AGN-y (Merloni i in., 2014b; Padovani i in., 2017).

Emisja z dysku akrecyjnego w układzie spoczynkowym cechuje się charakterystycznym wykładniczym kształtem kontinuum. Wraz z szerokimi i wąskimi liniami emisyjnymi pochodzącymi odpowiednio z BLR i NLR dostarcza ona wielu informacji o strukturze i kinematyce materii w pobliżu SMBH. Ta informacja jest uzyskiwana zarówno z badań spektroskopowych (Elvis i in., 1994; Vanden Berk i in., 2001; Netzer, 2015) jak i rewerberacyjnych. Mapowanie rewerberacyjne AGN-ów (ang. *reverberation mapping*) pozwala zastąpić rozdzielczość przestrzenną rozdzielczością czasową. Analiza opóźnienia czasowego pomiędzy zmiennością widma w zakresie optycznym i ultrafioletowym a zmiennością odpowiedzi szerokich linii emisyjnych pozwala

na rekonstrukcję procesów zachodzących w sąsiedztwie SMBH. W tym rozdziale omówione zostaną jedynie właściwości fotometrycznej selekcji AGN-ów w paśmie optycznym. W celu zaznajomienia się z metodą mapowania rewerberacyjnego czytelnik jest odsyłany do obszernych analiz zawartych w Peterson (1993) oraz Cackett, Bentz i Kara (2021).

Fotometryczna identyfikacja AGN-ów w pasmie optycznym pozwala na uzyskanie katalogów o dużych objętościach kosztem znacznego statystycznego obciążenia selekcji oraz dużego zanieczyszczenia. Optycznie wyselekcjonowane katalogi AGN-ów próbkują głównie populację typu I o wysokim współczynniku Eddingtona ($L/L_{Edd} > 0.01$) (Vestergaard i in., 2008; Trakhtenbrot i Netzer, 2012). Wygląd AGN-ów w szerokich pasmach optycznych jest silnie zależny od wykładniczego kształtu kontinuum oraz obecności szerokich linii emisyjnych. Te właściwości, obserwowane w układzie spoczynkowym, nadają AGN-om ściśle określone położenie w przestrzeni kolorów. Jednak w pewnych przedziałach przesunięć ku czerwieni kolory obiektu stają się podobne do gwiazdowych, co bardzo utrudnia separację obu klas. Jest to szczególnie problematyczne w przypadku kwazarów, które podobnie jak gwiazdy są obiektami punktowymi. Zjawisko to wprowadza duże zanieczyszczenie gwiazdowe, zwłaszcza w pobliżu płaszczyzny Galaktyki. Wspomniane wyżej problematyczne zakresy przesunięć ku czerwieni są obecne w obszarze wokół $z \sim 2.6$ oraz $z \sim 3.5$ (Richards i in., 2002; Richards i in., 2006; Padovani i in., 2017). Inny zauważalny spadek kompletności optycznych katalogów AGN-ów wynika z niskiej czułości na obiekty typu II (Zakamska i in., 2003).

Przeanalizujmy teraz dokładniej własności AGN-ów w podczerwieni (IR), ponieważ są one kluczowe dla tej pracy. Podczerwoną część widma można podzielić na trzy główne zakresy: bliską podczerwień (ang. *near-infrared*, NIR: 1–5 μm), średnią podczerwień (ang. *mid-infrared*, MIR: 5–50 μm) [2] i daleką podczerwień (ang. *far-infrared*, FIR: 50–500 μm). Podczerwone promieniowanie AGN-ów występuje głównie w zakresie NIR i MIR i wynika z re-emisji przez pył krzemianowy znajdujący sie w torusie energii pochodzącej z dysku akrecyjnego. Pył ten wytwarza kontinuum NIR-MIR o charakterystycznym dla AGN-ów wykładniczym kształcie $F_\nu \propto \nu^\alpha$ w przedziale 3–8 μm, z wartością parametru wykładniczego $\alpha < 0$ (Klaas i in., 2001; Alonso-Herrero i in., 2001; Alonso-Herrero i in., 2006a). W szczególności Alonso-Herrero i in. (2006a) stwierdzili, że galaktyki zdominowane przez AGN wykazują $-2,8 < alpha < -0,5$ w zakresie 3,6–8 μm. Ponadto, krzemianowy pył torusa wytwarza również dwie ważne cechy widmowe zlokalizowane przy 9,7 μm i 18 μm. Cechy te pochodzą odpowiednio od modów rozciągających i zginających SiO (Thompson i in., 2009). Innymi wyróżniającymi się cechami widmowymi AGN-ów obecnymi w średniej podczerwieni, które mogą mieć znaczenie dla metod selekcji, są silne linie [NeV] przy 14,3 μm i [OIV] przy 25,9 μm. Są one związane z jonizacją gazu przez fotony pochodzące z kontinuum i najprawdopodobniej powstają w obszarze NLR (Lutz i in., 2003). Opisane wyżej linie mogą mieć znaczący wkład do jasności obiektu w szerokich pasmach MIR.

Badania emisji AGN w podczerwieni pozwoliły ustalić podstawy zunifikowanego modelu (np. badania emisji w podczerwieni wraz z polaryzacją optyczną potwierdziły naturę źródeł ukrytego typu I jako AGN-u z torusem znajdującym się na linii widzenia obserwatora), a także umożliwiły badanie rozkładu gazu i pyłu w torusie (Thompson i in., 2009; Sirocky i in., 2008; Marin i in., 2018; Lopez-Rodriguez i in., 2018). Istnieją trzy główne grupy modeli rozkładu materii w torusie, które starają się

[2] W niektórych publikacjach można spotkać się z alternatywnym podziałem, w którym NIR obejmuje zakres 1–3 μm, a MIR odpowiednio 3–50 μm.

dopasować do obserwowanych spektralnych rozkładów energii AGN-ów: modele rozkładu ciągłego (ang. *continuous distribution models*, Pier i Krolik, 1992; Fritz, Franceschini i Hatziminaoglou, 2006), modele torusa skłębionego (ang. *clumpy torus models*, Nenkova, Ivezić i Elitzur, 2002; Nenkova i in., 2008a) oraz modele pośrednie łączące cechy obu poprzednich (Stalevski i in., 2012). Porównanie własności i przewidywań poszczególnych modeli zostało zaprezentowane w Feltre i in. (2012) i Lira i in. (2013) oraz Netzer (2015).

Szereg wyników obserwacyjnych zdecydowanie zaprzecza modelowi ciągłego rozkładu. Gdyby materia w torusie była rozłożona w sposób ciągły, powstałby silny gradient temperatury wzdłuż promienia torusa skutkujący obecnością wyraźnych cech absorpcyjnych w widmie. Wynikałoby to z tego, że strumień emitowany przez wewnętrzny region torusa byłby pochłaniany przez materiał pyłowy znajdujący się w większej odległości od centrum. Zatem analiza gorącego wewnętrznego obszaru torusa powinna wykazywać silną linię emisyjną krzemu przy 9,7 μm i mniejszą linię krzemu przy 18 μm. Z kolei analiza zewnętrznych regionów torusa powinna wykazać absorpcję w tych cechach. W związku z tym dla modelu torusa o ciągłym rozkładzie materii, obraz AGN-ów typu I w MIR, gdzie widoczny jest wewnętrzny region torusa, powinien pokazywać linie krzemu w emisji. Z drugiej strony, w AGN-ach typu II, gdzie obserwator widzi tylko zewnętrzny region, linie krzemu w MIR powinny być obecne w absorpcji. Obserwacje wykazują jednak znacznie słabszą absorpcję linii krzemu 9,7 μm niż przewidywana przez model rozkładu ciągłego, co wskazuje na zbitą strukturę torusa (Shi i in., 2006; Nenkova i in., 2008b; Martínez-Paredes i in., 2020). Ponadto porównanie intensywności linii 9,7 μm z intensywnością drugiej linii krzemu przy 18 μm pozwala jeszcze lepiej rozróżnić dwa modele rozkładu pyłu i określić skład pyłowy torusa (Hao i in., 2005; Thompson i in., 2009). Czytelnika możemy jednak również odesłać do dyskusji na temat wpływu różnego składu chemicznego modeli na przewidywania cech krzemianowych (Sirocky i in., 2008). Innym dowodem obserwacyjnym przeczącym modelom gładkich torusów jest silnie izotropowa emisja AGN-ów w MIR (Horst i in., 2006). Modele rozkładu ciągłego przewidują znacznie większą jasność AGN-ów typu I w MIR dla ustalonej jasności całkowitej (Thompson i in., 2009). Wreszcie obserwowany szeroki zakres temperatur pyłu obserwowany w określonej odległości od obszaru centralnego (Jaffe i in., 2004; Beckert i in., 2008) nie jest możliwy w przypadku gładkiego torusa, gdzie powinien występować silny gradient temperatury. Dlatego obserwowane zachowanie może występować jedynie w przypadku struktury zbitej, w której przerwy między chmurami pyłu pozwalają centralnemu obszarowi ogrzewać bardziej odległą materię. Zgodnie z sugestią zawartą w pracy Netzer (2015), prawdopodobnie bardziej realistycznie strukturę torusa opisuje dwufazowy model pośredni, w którym przestrzeń pomiędzy skupiskami pyłu wypełniona jest rozrzedzonym gazem pyłowym powodującym dodatkowe tłumienie padającego promieniowania. Jednym z powodów, dla których w torusie powinien być obecny rozrzedzony ośrodek, są nieuniknione zderzenia obłoków pyłowych. Jednak mimo obserwacyjnych wskazówek, problem rzeczywistej struktury torusa pozostaje otwarty (González-Martín i in., 2019a; González-Martín i in., 2019b; Victoria-Ceballos i in., 2022). Oprócz powyższej dyskusji na temat geometrii torusa, składnik pyłowy wydaje się być również obecny na obrzeżach NLR poza główną płaszczyzną. Ta struktura pyłowa może mieć duży udział w emisji kontinuum w zakresie 10–30 μm (Schweitzer i in., 2008).

Selekcja AGN-ów w podczerwieni skupia się na sygnaturach emisji pyłu w pasmach NIR i MIR. Emisja pyłu nadaje AGN-om charakterystyczne czerwone kolory

w zakresie NIR i krótszych falach MIR, odróżniając je od gwiazd i zwykłych galaktyk [3]. Emisja FIR jest rzadko wykorzystywana do selekcji AGN, ponieważ jest ona spowodowana głównie aktywnością gwiazdotwórczą galaktyki-gospodarza (Netzer i in., 2007; Hatziminaoglou i in., 2010; Mullaney i in., 2011). Przesłonięcie regionu centralnego nie ogranicza możliwości fotometrycznej selekcji w podczerwieni tak bardzo, jak ma to miejsce w przypadku selekcji optycznej. Co więcej, ma ona również przewagę nad selekcją rentgenowską w postaci znacznie większej szybkości obserwacji, co przekłada się na większe objętości katalogów (Padovani i in., 2017; Gorjian i in., 2008). Selekcja AGN-ów oparta na kolorach NIR-MIR została zastosowana we wszystkich głównych teleskopach podczerwonych, takich jak Spitzer (Werner i in., 2004; Stern i in., 2005), WISE (Wright i in., 2010; Stern i in., 2012; Assef i in., 2013) oraz AKARI (Murakami i in., 2007; Lee i in., 2007; Oyabu i in., 2011).

Istnieje jednak kilka źródeł zanieczyszczeń i statystycznych obciążeń selekcji obecnych w metodach opartych na danych podczerwonych. Podstawowe zanieczyszczenie na niższych wysokościach galaktycznych pochodzi od brązowych karłów (Stern i in., 2007) i młodych obiektów gwiazdowych (ang. *young stellar objects*, Koenig i in., 2012), które mogą naśladować podczerwone kolory AGN-ów w określonych zakresach przesunięciac ku czerwieni. Problem poprawnej identyfikacji tych obiektów staje się mniej istotny przy większej odległości kątowej od płaszczyzny Galaktyki, gdzie głównym źródłem zanieczyszczenia katalogu są galaktyki aktywne gwiazdotwórczo (ang. *star-forming galaxies*, SFG).

Aby lepiej zrozumieć naturę błędnej klasyfikacji SFG, musimy zrozumieć, w jaki sposób składnik gwiazdowy przyczynia się do widmowego rozkładu energii galaktyki w podczerwieni. Przy krótszych długościach fal NIR, kluczową rolę odgrywa cecha widmowa zlokalizowana w pobliżu 1,6 μm zwana z ang. *stellar bump* (Padovani i in., 2017; Stern i in., 2005; Alonso-Herrero i in., 2006b; Donley i in., 2012).

Cecha ta jest charakterystyczna dla atmosfer chłodnych gwiazd i stanowi ważny element widmowego rozkładu energii prawie wszystkich populacji gwiazdowych. Brakuje jej natomiast w przypadku bardzo młodych gorących gwiazd ($\sim$1 Myr). W ich przypadku widmowy rozkład energii przyjmuje wykładniczy kształt krzywej Rayleigha-Jeansa. Może ona być również mniej widoczna w gwiazdach o niskiej metaliczności (John, 1988; Sawicki, 2002). Innym zjawiskiem, które może wystąpić, jest przesunięcie obserwowanego położenia danej cechy widmowej w kierunku czerwieni, spowodowanego absorpcją części emisji występującej na krószych długościach fali przez pył ośrodka międzygwiazdowego (Sawicki, 2002). Obecność cechy widmowej 1,6 μm pochodzącej od emisji gwiazd w galaktykach, może silnie wpływać na selekcję AGN-ów. Szczególnie w przypadku SFG o wyższych przesunięciach ku czerwieni ($z \geq 1$), cecha 1,6 μm przesuwa się ku dłuższym falom w zakresie NIR. Widmo AGN w tym zakresie charakteryzuje się silnym spadkiem emisji dysku akrecyjnego, jak również niską emisją z torusa pyłowego. W rezultacie, cecha 1,6μm może naśladować w przestrzeni kolorów obecność wykładniczego kształtu widmowego rozkładu energii, typowego dla AGN-ów. Powoduje to, że SFG znajdujące się na dużych przesunięciach ku czerwieni mogą być łatwo mylone z AGN-ami w katalogach fotometrycznych (Stern i in., 2005; Donley i in., 2012; Assef i in., 2013).

Widmowy rozkład energii galaktyk aktywnych gwiazdotwórczo bez aktywnego jądra wykazuje charakterystyczne lokalne minimum między cechą 1,6 μm a emisją w FIR (i dalszym zakresem MIR), która pochodzi od pyłu podgrzanego przez składnik gwiazdotwórczy. W przypadku AGN ten zakres długości fal jest zajmowany przez

[3] Inna istotna cecha selekcji AGN-ów w NIR-MIR opiera się na możliwości znalezienia najbardziej odległych kwazarów o wysokim przesunięciu ku czerwieni, które są niewykrywalne w pasmach optycznych z powodu absorpcji Lyα (Bañados i in., 2016).

emisję kontinuum termicznego o wykładniczym kształcie, powstającą w torusie (Donley i in., 2012). Podczas gdy ogólna emisja z galaktyki-gospodarza staje się mniej znacząca w zakresie MIR i dłuższych falach NIR, znaczący wkład może pochodzić od kilku cech emisjnych cech widmowych pyłu związanych z obecnością ośrodka międzygwiazdowego. Najbardziej widoczne są cechy widmowe związane z emisją z dużych cząsteczek wielopierścieniowych węglowodorów aromatycznych (ang. *polycyclic-aromatic-hydrocarbon*, PAH), zawierających ∼50 atomów węgla (Leger i Puget, 1984; Allamandola, Tielens i Barker, 1985; Allamandola, Tielens i Barker, 1989). Emisja PAH pochodzi z re-emisji fotonów z zakresu ultrafioletu oraz do pewnego stopnia zakresu światła widzialnego (Uchida, Sellgren i Werner, 1998) i jako taka jest często używana jako wskaźnik obecności procesów gwiazdotwórczych. Przez długi czas pytanie o to, jak dokładnie cechy widmowe związane z emisją PAH śledzą procesy gwiazdotwórcze, pozostawało otwarte. Analiza danych spektroskopowych MIR z teleskopu Spitzera, jak również wcześniejsze badania teleskopu Infrared Space Observatory (ISO; Kessler i in., 1996) pokazały, że emisja PAH nie śledzi silnych procesów gwiazdotwórczych, których sygnaturą jest obecność gwiazd populacji O (Haas, Klaas i Bianchi, 2002). Zamiast tego jest najbardziej wrażliwa na obecność populacji B, która stanowi główny składnik gwiazdowy galaktyk (Peeters, Spoon i Tielens, 2004; Smith i in., 2007; Maiolino i in., 2007).

Główne cechy PAH znajdują się w zakresach NIR i MIR i są ulokowane na 3,3 μm, 6,2 μm, 7,7 μm, 8,6 μm i 11,3 μm. Emisja z samych PAH może stanowić do 20% całkowitej jasności galaktyki w podczerwieni (L_{IR}) w przypadku galaktyk, w których zachodzą silne procesy gwiazdotwórcze (Smith i in., 2007). Emisja PAH może więc znacząco wpływać na rozkład SFG w przestrzeni kolorów w podczerwieni. Może ona również ukrywać obecność AGN w widmach zdominowanych przez PAH w przypadku selekcji opartej na podczerwonych kolorach. Najbardziej widoczna cecha widmowa znajdująca się przy 7,7 μm odegrała ponadto ważną rolę w ustaleniu obserwacyjnych dowodów na istnienie zunifikowanego modelu AGN. W pracy Clavel i in. (2000), autorzy porównali widma MIR galaktyk Seyferta typu I i typu II. Analizując właściwości cech PAH w zakresie 7,7mum wykazali, że właściwości galaktyki-gospodarza nie są związane z typem AGN. Wykazali oni również podobieństwa we właściwościach podczerwonej części widmowego rozkładu energii pomiędzy galaktykami Seyferta II i galaktykami aktywnymi gwiazdotwórczo. Wyjątek tutaj stanowi zakres NIR-MIR zdominowany przez emisję torusa (Peeters, Spoon i Tielens, 2004).

Emisja PAH, wraz z własnościami AGN w podczerwieni, została również wykorzystana do badania natury galaktyk o bardzo dużej jasności w podczerwieni (ang. *ultra-luminous infrared galaxies*, ULIRGs): $L_{IR} > 10^{12} L_{\odot}$. Populacja ULIRG-ów stanowi ważną grupę źródeł przy analizie metod selekcji AGN-ów w zakresie podczerwonym. Badania morfologiczne tych obiektów wykazały, że większość z nich to układy galaktyk będących w stadium łączenia (Sanders i Mirabel, 1996). Na podstawie analizy emisyjnych cech widmowych PAH oraz absorpcyjnych cech krzemianowych obecnych w widmach obiektów ULIRG stwierdzono, że galaktyki z dominacją zarówno AGN, jak i składnika gwiazdotwórczego mogą występować jako obiekty typu ULIRG (Spoon i in., 2007). Nie znaleziono również konkluzywnych dowodów potwierdzających klasyczne przewidywania modeli łaczenia galaktyk, według których późniejsze etapy łączenia się galaktyk są zdominowane przez AGN (Rigopoulou i in., 1999).

Analizując statystyczne obciążenie selekcji AGN-ów w podczerwieni, należy rozważyć kilka ważnych mechanizmów. Jeden z nich wynika ze zmiany koloru AGN

w podczerwieni w przypadku silnej, szerokiej emisji H, która może wystąpić w pasmach NIR i zredukować czerwony kolor obiektu. Problem ten występuje w zakresie przesunięcia ku czerwieni $3.5 \leq z \leq 5$ i można go złagodzić stosując kombinację fotometrii podczerwonej i optycznej do selekcji AGN (Richards i in., 2009). Innym źródłem statystycznego obciążenia selekcji jest trudność identyfikacji AGN-ów o niskim stosunku emisji składnika pyłowego AGN do emisji pyłu galaktyki-gospodarza (Hao i in., 2010). Metody selekcji kolorów w podczerwieni mogą nie wychwycić tych obiektów ze względu na brak głównego składnika rozróżniającego AGN-y i normalne galaktyki - emisji pyłu AGN w MIR. Inne, bardziej subtelne zakłócenie selekcji pochodzi z wzajemnego oddziaływania galaktyki-gospodarza i aktywnego jądra galaktyki. W szczególności istotną rolę odgrywa korelacja pomiędzy jasnością składnika sferoidalnego galaktyki-gospodarza a jasnością AGN (Marconi i Hunt, 2003). Zależność ta przekłada się na charakterystyczną zależność stosunku jasności AGN do jasności galaktyki-gospodarza $L_{\nu,AGN}/L_{\nu,host}$ od współczynnika Eddingtona. Przekłada się to na obniżoną liczbę AGN-ów o niskim współczynniku Eddingtona w katalogach podczerwonych (Padovani i in., 2017). Stanowi to najpoważniejsze ograniczenie technik selekcji w podczerwieni. Selekcja w podczerwieni próbkuje jedynie kilka procent górnej granicy rozkładu L/L_{Edd} obecnego w katalogach rentgenowskich AGN (Hickox i in., 2009; Mendez i in., 2013). Efekt ten przekłada się również na systematyczny błąd w rozkładzie mas SMBH M_{BH}. W związku z tym w katalogach IR brakuje dużej liczby galaktyk spiralnych z wyraźnym składnikiem dyskowym i niższymi masami M_{BH} (Padovani i in., 2017; Magorrian i in., 1998).

Opisując własności katalogu i skuteczność selekcji, należy wziąć pod uwagę głębokość przeglądu. Chociaż selekcja AGN-ów w MIR jest uważana za bardzo efektywną, wykazuje ona różne własności dla płytkich i głębokich przeglądów. Jak pokazano w Assef i in. (2013), w przypadku płytkich przeglądów, selekcja w MIR może odzyskać katalog AGN-ów charakteryzujący się zarówno wysoką czystością, jak i kompletnością. Jednak w przypadku głębokich przeglądów pojawia się potrzeba silnego kompromisu między kompletnością a czystością.

2.3 Wpływ aktywnego jądra na fizykę galaktyk oraz jego związek z kosmologią obserwacyjną

Badania własności statystycznych katalogów AGN-ów dają nam wgląd w ewolucję galaktyk, wpływ aktywności AGN na procesy zachodzące w galaktyce-gospodarzu, a także związek między rozkładem przestrzennym AGN-ów a wielkoskalową strukturą Wszechświata. Zjawisko wpływu AGN na powstawanie gwiazd w galaktyce-gospodarzu znane jest jako sprzężenie zwrotne AGN (ang. *AGN feedback*). Proces ten zachodzi poprzez oddziaływanie pomiędzy promieniowaniem pochodzącym z akrecji materii na SMBH a składnikiem gazowym sferoidalnej środkowej części galaktyki-gospodarza. Ciśnienie promieniowania pochodzące z AGN wypycha gas ośrodka międzygwiazdowego poza sferoidalny centralny obszar galaktyki-gospodarza, kończąc w nim proces formowania się gwiazd. Równocześnie brak gazu w obszarze sferoidalnym ogranicza ilość materiału potrzebnego do procesu akrecji, kończąc aktywność silnika centralnego w AGN. W szczególności, zależność pomiędzy M_{BH} a dyspersją prędkości gwiazd w obszarach centralnych, w tym w zgrubieniu centralnym galaktyki-gospodarza (Gültekin i in., 2009) jest często traktowana jako ważny przypadek pośrednich obserwacji sprzężenia zwrotnego AGN spowodowanego przez AGN-y w trybie radiacyjnym. Kolejny dowód obserwacyjny na istnienie sprzężenia zwrotnego pochodzi z pomiarów mas najjaśniejszych galaktyk w gromadach

(ang. *brightest cluster galaxies*, BCG). Bez wkładu energii pochodzącego ze sprzężenia zwrotnego AGN w trybie kinetycznym, powinny one być jeszcze bardziej masywne, i widzielibyśmy je nadal jako olbrzymie galaktyki gwiazdotwórcze. Czytelnik może znaleźć obszerną dyskusję na temat obu powyższych faktów obserwacyjnych w Fabian (2012). Oba tryby AGN pełnią zatem odrębne fundamentalne role. AGN w trybie radiacyjnym ma tendencję do wypychania gazu ze struktury sferoidalnej galaktyki-gospodarza. Natomiast AGN pracujący w trybie kinetycznym utrzymuje gaz w wystarczająco wysokiej tempreraturze, by nie zasilał on procesów gwiazdotwórczych w pobliżu SMBH.

Wiele badań nad sprzężeniem zwrotnym AGN sugeruje, że aktywność AGN może odgrywać główną rolę w ukształtowaniu bimodalnej natury populacji galaktyk i wygaszaniu procesów powstawania gwiazd, w szczególności w galaktykach należących do populacji niebieskiej (Croton i in., 2006). Bimodalność populacji galaktyk definiuje dwie główne grupy galaktyk (Kauffmann i in., 2003a; Blanton i in., 2003). Jedna z nich, zwana niebieską populacją albo niebieską chmurą, ang. *blue cloud*, składa się z galaktyk późnego typu Hubble'a, charakteryzujących się silnymi procesami gwiazdotwórczymi, małymi masami gwiazdowymi i ścisłą zależnością między tempem tworzenia gwiazd a masą gwiazdową (Brinchmann i in., 2004; Noeske i in., 2007). Druga grupa, zwana czerwoną populacją lub czerwonym ciągiem, ang. *red sequence*, obejmuje galaktyki wczesnego typu Hubble'a ze znacznie słabszą składową gwizadotwórczą i zazwyczaj większą masą gwiazdową.

Uproszczony scenariusz ewolucji galaktyk (Lilly i in., 2013; Hickox i in., 2009) można opisać następująco. Najpierw galaktyka ewoluuje wzdłuż niebieskiego ciągu głównego formowania gwiazd, zwiększając masę poprzez akrecję materii z ośrodka międzygalaktycznego oraz, w mniejszym stopniu, poprzez łączenie się z innymi galaktykami. Po osiągnięciu masy krytycznej zarówno akrecja zimnego gazu, jak i procesy gwiazdotwórcze zostają wygaszone. Ten etap wyznacza moment wejścia galaktyki do populacji czerwonej. Zatem, zgodnie z obecnym paradygmatem, sprzężenie AGN odgrywa istotną rolę w tym procesie, przyczyniając się do powstrzymania galaktyk wchodzących do czerwonej populacji przed dalszym tworzeniem gwiazd. Jednakże badania nad sprzężeniem AGN są stosunkowo nową dziedziną i niezbędne są dalsze badania, aby lepiej zrozumieć ten proces. Kluczem do uzyskania pełnego obrazu interakcji pomiędzy aktywnością AGN a ewolucją galaktyk jest analiza dużych zbiorów danych astronomicznych. W tym celu łączy się badanie własności galaktyk-gospodarzy dla różnych typów AGN-ów z analizą własności samych AGN-ów i ich grupowania.

Wyniki badań przedstawione w Hickox i in. (2009) pozwalają wyróżnić kilka istotnych obserwacji. Analiza grupowania różnych typów AGN-ów i ich lokalnych środowisk pokazuje związek pomiędzy trybem AGN-u a masą i wiekiem galaktyki-gospodarza. Wiadomo, że galaktyki czerwone wykazują większe amplitudy grupowania i znajdują się w gęstszym otoczeniu niż obiekty z populacji niebieskiej (Zehavi i in., 2005; Meneux i in., 2006; Coil i in., 2008; Zehavi i in., 2011). AGN-y wyselekcjonowane przy pomocy metod rentgenowskich zajdują się przeważnie w galaktykach będących w okresie przejściowym między populacjami niebieską i czerwoną. AGN-y te charakteryzują się na ogół większą amplitudą grupowania i zamieszkują gęste środowiska (Gilli i in., 2005; Miyaji i in., 2007). Jeszcze silniejsze grupowanie wykazują AGN-y wyselekcjonowane w zakresie radiowym. Znajdują się one w galaktykach należących do czerwonej sekwencji, które występują w grupach i gromadach galaktyk oraz wykazują właściwości grupowania podobne do lokalnych galaktyk eliptycznych (Wake i in., 2008; Mandelbaum i in., 2009). AGN-y

wyselekcjonowane w podczerwieni rezydujące w bardziej niebieskich galaktykach-gospodarzach wykazują znacznie słabsze właściwości grupowania. Wynika z tego, że aktywność jasnych w podczerwieni AGN-ów może być wywoływana nie tylko przez właściwości galaktyki-gospodarza, ale także przez lokalne, najczęściej słabo zagęszczone środowisko (Hickox i in., 2009).

Tego typu badania otwierają możliwość wprowadzenia AGN-ów w kontekst kosmologii obserwacyjnej. Analiza grupowania w skali ∼Mpc pozwala oszacować masy halo ciemnej materii (Sheth i Tormen, 1999) i powiązać je z różnymi typami AGN-ów, podobnie jak to się robi w przypadku normalnych galaktyk (Faber i in., 2007). W szczególności, wyselekcjonowane radiowo AGN-y na stosunkowo małych przesunięciach ku czerwieni ($0.25 < z < 0.8$), które cechuje najsilniejsze grupowanie, zajmują najbardziej masywne halo ciemnej materii ($M_{\rm halo} \sim 3 \times 10^{13} h^{-1} M_{\odot}$), z rosnącym odchyleniem w mniejszych skalach ($< 0.5 h^{-1}$Mpc). Halo o takich masach charakteryzują duże grupy galaktyk lub małe gromady. Rosnąca skośność w mniejszych skalach sugeruje, że korelacja krzyżowa wybranych radiowo AGN-ów z normalnymi galaktykami w tych skalach będzie miała tendencję do wybierania par obiektów zajmujących to samo halo ciemnej materii (Zehavi i in., 2004; Coil i in., 2008; Brown i in., 2008). Szacunkowa masa halo ciemnej materii otrzymana z grupowania AGN-ów wyselekcjonowanych w świetle rentgenowskim daje mniejszą wartość $M_{\rm halo} \sim 10^{13} h^{-1} M_{\odot}$, typową dla mniejszych grup galaktyk. Co ciekawe, korelacja krzyżowa rentgenowskich AGN-ów z galaktykami normalnymi na mniejszych skalach ($1 h^{-1}$ Mpc) wykazuje mniejsze nachylenie w porównaniu z funkcją autokorelacji rentgenowskich AGN-ów. Podobny wynik uzyskano na optycznie wyselekcjonowanej próbce (Li i in., 2006), pokazując słabsze grupowanie par AGN-galaktyka normalna na małych skalach. Autorzy wyjaśnili to zachowanie jako tendencję AGN-ów do występowania w centralnych galaktykach. Jednakże, w przypadku rentgenowskich AGN-ów, mały rozmiar katalogu AGN-ów nie pozwolił na wyciągnięcie przekonujących wniosków. AGN-y wyselekcjonowane w podczerwieni zajmują jeszcze mniejsze halo ciemnej materii o masie $M_{\rm halo} \leq 10^{12} h^{-1} M_{\odot}$ masy. W tym przypadku autorzy postawili również hipotezę, że AGN-y selekcjonowane w zakresie podczerwieni są obecne w centralnych galaktykach małych halo.

Jak już wspomniano w rozdziale 2.2, metody selekcji AGN-ów w zakresie radiowym, rentgenowskim i podczerwonym próbkują różne zakresy współczynnika Eddingtona. W artykule Hickox i in. (2009) autorzy wykazali, że AGN-y selekcjonowane metodą rentgenowską próbkują prawie cały zakres współczynnika Eddingtona wynoszący $10^{-3} \leq L/L_{Edd} \leq 1$, podczas gdy AGN-y wyselekcjonowane radiowo reprezentują dolną granicę rozkładu ($L/L_{Edd} \leq 10^{-3}$), a AGN-y wyselekcjonowane w podczerwieni reprezentują górną granicę ($L/L_{Edd} \geq 10^{-2}$).

Powyższe obserwacje zostały zinterpretowane przez autorów omawianej pracy Hickox i in. (2009) jako uzupełnienie wcześniej przedstawionego modelu ewolucji galaktyk o wpływ pochodzący od AGN. W tym scenariuszu faza radiacyjna AGN i formowanie się centralnej sferoidalnej struktury galaktyki-gospodarza zachodzi, gdy halo ciemnej materii osiąga masę 10^{12} - $10^{13} M_{\odot}$. Po tym epizodzie procesy gwiazdotwórcze w galaktyce-gospodarzu zostają zahamowane, a tryb AGN zmienia się z radiacyjnego (reprezentowanego przez AGN-y jasne w zakresie optycznym i w podczerwieni) na kinetyczny (tj. źródła świecące radiowo). Podsumujmy pokrótce opisany w tym rozdziale obraz współewolucji AGN-ów i galaktyk z punktu widzenia niniejszej pracy. Różne metody selekcji AGN-ów pozwalają nam próbkować AGN-y w różnych stadiach trybów akrecji, próbkować różne galaktyki-gospodarze oraz różne środowiska i rozmiary halo ciemnej materii. Te własności sprawiają, że

statystyczna analiza katalogów AGN-ów jest bardzo potężnym narzędziem astrofizycznym i kosmologicznym. Dlatego zwiększenie objętości katalogów o wielu długościach fali oraz zachowanie wysokiej czystości i kompletności tych katalogów jest kluczowe dla współczesnych badań kosmologicznych.

3

Dane

3.1 Przeglądy nieba w polu północnego bieguna ekliptycznego

Region Północnego Bieguna Ekliptycznego (ang. North Ecliptic Pole, NEP; $\alpha(\mathrm{J2000}) = 18^h00^m00^s$, $\delta(\mathrm{J2000}) = 66°33'88''$, we współrzędnych galaktycznych $l = 96°.4$, $b = 29°.8$) jest jednym z kilku pól głębokiego nieba obserwowanych przez bezprecedensowo dużą liczbę instrumentów, co daje nam szeroki, wielozakresowy pogląd na naturę obiektów astronomicznych. Obserwacyjne misje kosmiczne często wybierają region NEP ze względu na wygodę obserwacji z orbity geocentrycznej i jego właściwości w odniesieniu do badań pozagalaktycznych: NEP jest regionem nieba na stosunkowo dużej wysokości galaktycznej, charakteryzującym się stosunkowo niską obecnością pyłu Galaktycznego. Taka kombinacja pozwala na przeprowadzenie głębokiego i zarazem stosunkowo szerokiego przeglądu i zmniejszenie wariancji kosmicznej (ang. cosmic variance).

Region NEP, obserwowany w różnych zakresach długości fal, daje nam szeroki, panchromatyczny obraz głębokiego nieba, od promieniowania rentgenowskiego do danych radiowych. Wysokoenergetyczne obserwacje obszaru NEP zostały po raz pierwszy wykonane przez rentgenowską misję kosmiczną ROSAT (Truemper, 1982). Przegląd ROSAT NEP (Henry i in., 2001; Voges i in., 2001; Henry i in., 2006) obserwował szeroki obszar 80,6 deg^2 skupiony wokół NEP (patrz także niedawne wydanie zaktualizowanego katalogu ROSAT NEP w Hasinger i in., 2021). Katalog ten, wraz z optycznymi obserwacjami uzupełniającymi (Gioia i in., 2003), został wykorzystany do wielu badań nad gromadami galaktyk (Gioia i in., 2001; Mullis i in., 2001; Gioia i in., 2004; Pratt i Bregman, 2020), badaniami aktywnych jąder galaktyk (Wolter i in., 2005)) i badań wzajemnego oddziaływania AGN-ów i gromad galaktyk (Branchesi i in., 2006; Cappelluti i in., 2007). Teleskopy rentgenowskie kolejnych generacji również prowadziły obserwacje regionu NEP jako uzupełnienie istniejących (lub przyszłych) przeglądów. Teleskop Chandra (Weisskopf i in., 2000) wykonał obserwacje uzupełniające w polu AKARI NEP-Deep (Krumpe i in., 2015). Teleskop NuSTAR (Harrison i in., 2013) wykonał obserwacje w planowanym obecnie polu James Webb Space Telescope North Ecliptic Pole Time-domain Field (Zhao i in., 2021), które jest małym polem o powierzchni 0.16 deg^2 w pobliżu centrum NEP. W najbliższej przyszłości region NEP będzie również obserwowany za pomocą teleskopu eROSITA (Predehl i in., 2021), jeśli eROSITA powróci do trybu operacyjnego.

Optyczne i ultrafioletowe obserewacje regionu NEP były wykonywane prawie wyłącznie jako uzupełnienie obserwacji poszczególnych pól przez inne instrumenty,

głównie instrumenty podczerwone. Pole NEP było obserwowane przez kilka dużych misji podczerwonych. Pierwsze obserwacje w głębokiej podczerwieni zostały wykonane przez satelitę IRAS (Neugebauer i in., 1984) w zakresie MIR-FIR (Hacking i Houck, 1987) z następującą po nich uzupełniającą obserwacją radiową za pomocą instrumentu VLA (Hacking i in., 1989). Katalog ten dostarczył obserwacji do pionierskich prac nad własnościami galaktyk w podczerwieni i modelami ewolucji galaktyk (Hacking, Condon i Houck, 1987; Hacking i Soifer, 1991; Ashby, Houck i Hacking, 1992). Następnie, obserwacje w zakresie NIR-MIR zostały wykonane przez sondę AKARI (Murakami i in., 2007; Matsuhara i in., 2006), która wykonała dwa najważniejsze i najgłębsze przeglądy w podczerwieni regionu NEP: AKARI NEP-Wide survey obejmujący region 5.4 $\deg^2$ (Lee i in., 2009) oraz AKARI NEP-Deep survey, który objął obszar $\sim$0.6 $\deg^2$ (Wada i in., 2008). Pola obu przeglądów były następnie obserwowane przez liczne teleskopy optyczne, gromadząc pokaźną ilość danych fotometrycznych i spektroskopowych (Jeon i in., 2010; Goto i in., 2017; Hwang i in., 2007; Huang i in., 2020; Oi i in., 2021). Uzupełniające obserwacje regionu AKARI NEP w dalekiej podczerwieni zostało wykonane przez satelitę Herschel (Pilbratt i in., 2010). Obserwował on dwa pola AKARI NEP za pomocą dwóch różnych instrumentów: pole AKARI NEP-Wide było obserwowane przez instrument SPIRE (Pearson i in., 2017) pokrywający zakres dłuższych fal (Griffin i in., 2010), natomiast pole AKARI NEP-Deep było obserwowane przez instrument PACS (Poglitsch i in., 2010; Pearson i in., 2019). Pole AKARI NEP było dodatkowo obserwowane przez satelitę GALEX (Martin i in., 2005), który dostarczył danych w bliskim i dalekim ultrafiolecie (Burgarella i in., 2019). Kolejne obserwacje uzupełniające pola AKARI NEP-Deep zostały wykonane przez JCMT/SCUBA2 (Holland i in., 1999) w zakresie submm (Geach i in., 2017; Shim i in., 2020) oraz przez The Westerbork Radio Synthesis Telescope w zakresie fal radiowych (White i in., 2010).

Po zakończonych obserwacjach AKARI, dane NIR-MIR regionu NEP zostały zebrane przez dwa inne duże kosmiczne obserwatoria podczerwieni. Pierwszym z nich był teleskop WISE (Wright i in., 2010). Dzięki kilkuset skanom obszaru NEP udało mu się stworzyć katalog głębokiego pola o powierzchni 1.5 $\deg^2$ (Jarrett i in., 2011). Drugim był teleskop Spitzera (Werner i in., 2004), który objął większy obszar o powierzchni 7.04 $\deg^2$ (Nayyeri i in., 2018). Pole NEP było również obserwowane przez specjalistyczny instrument CIBER (Bock i in., 2013) używany do badań kosmicznego tła w bliskiej podczerwieni (Zemcov i in., 2014).

Dane wielozakresowe (ang. *multiwavelenght data*) zebrane w regionie NEP zostały wykorzystane w wielu badaniach pozagalaktycznych. Były to liczne badania własności AGN (Wang i in., 2020; Yang i in., 2020; Chiang i in., 2019; Santos i in., 2021; Barrufet de Soto i in., 2017), ewolucji galaktyk (Kim i in., 2015; Oi i in., 2017; Kim i in., 2019; Goto i in., 2019; Barrufet i in., 2020; Kim i in., 2021a) czy badania grupowania gromad galaktyk (Solarz i in., 2015; Seo i in., 2019) i własności gromad galaktyk (Huang i in., 2021), by wymienić tylko kilka z nich. Oprócz badań stricte astrofizycznych, dane NEP były wykorzystywane do badań z zakresu zastosowania uczenia maszynowego. Były to m.in. metody automatycznej selekcji AGN w danych podczerwonych i wielozakresowych (Poliszczuk i in., 2019; Poliszczuk i in., 2021; Chen i in., 2021), separacji gwiazd i galaktyk (Solarz i in., 2012), czy zastosowanie głębokich sieci neuronowych do selekcji łączących się układów galaktyk w danych optycznych (Pearson i in., 2022).

3.2 Wielozakresowy katalog pola AKARI NEP-Wide

Jak wspomniano w poprzednich rozdziałach, zrozumienie własności w podczerwieni różnych populacji galaktyk jest kluczowe dla zrozumienia historii formowania się gwiazd, ewolucji galaktyk i procesów sprzężenia zwrotnego AGN. Doskonałym narzędziem do tych celów jest niedawno opublikowany wielozakresowy katalog pola AKARI NEP-Wide (Kim i in., 2021b). Katalog ten jest oparty na katalogu źródeł AKARI NEP-Wide (Lee i in., 2009; Kim i in., 2012), połączony z optycznymi obserwacjami fotometrycznymi wykonanymi przez instrument SUBARU/HSC (Miyazaki i in., 2012; Oi i in., 2021). Ostateczny, łączony katalog 91 861 źródeł AKARI został uzupełniony o istniejące pomiary z innych instrumentów w szerokim zakresie widma, tworząc bogaty obraz pola NEP.

Teleskop kosmiczny AKARI został uruchomiony w 2006 roku i wykonał trzy główne przeglądy: przegląd całego nieba w zakresie MIR-FIR (Ishihara i in., 2010; Doi i in., 2015), a także dwa głębokie przeglądy w polu NEP w zakresie NIR-MIR (Matsuhara i in., 2006). Większy obszar $\sim$5.4 deg^2 skupiony wokół środka NEP został zmierzony przez przegląd AKARI NEP-Wide (Lee i in., 2009; Kim i in., 2012). Z kolei mniejszy obszar $\sim$0.6 deg^2 został zmierzony w przeglądzie AKARI NEP-Deep (Wada i in., 2008). Pole AKARI NEP-Deep znajduje się w pewnej odległości od środka NEP, by możliwym było dopasowanie danych z istniejących wcześniej optycznych obserwacji uzupełniających (Takagi i in., 2012).

Na pokładzie satelity AKARI znajdowały się dwa oddzielne instrumenty obserwacyjne. Pierwszym z nich był Far-Infrared Surveyor (FIS; Kawada i in., 2007) stworzony do badań w zakresie FIR w czterech szerokich pasmach obejmujących zakres 50-180 μm. Drugim instrumentem był InfraRed Camera (IRC; Onaka i in., 2004), obejmujący zakres 2–26 μm za pomocą trzech oddzielnych kanałów: NIR (2–5 mum), MIR-S (5–12 mum) i MIR-L (12–26 μm).

Pole NEP było również obserwowane za pomocą teleskopu Spitzera. Jednak satelita AKARI ma kilka unikalnych własności, dzięki którym lepiej nadaje się do badań NEP w zakresie NIR-MIR. Po pierwsze, bardzo ważną właściwością jest unikalne ciągłe fotometryczne pokrycie zakresu 2-26 μm. Pozwala to wypełnić lukę 8–24 μm obecną w fotometrii Spitzera, co jest kluczowe dla badania własności SFG i AGN w zakresie MIR. Ponadto, pole widzenia AKARI/IRC jest prawie cztery razy większe niż pole widzenia instrumentu Spitzer/IRAC, co pozwoliło znacznie zwiększyć szybkość wykonywania przeglądu (około 100 razy szybciej niż IRAC) i osiągnąć podobną głębokość przeglądu, pomimo krótszego czasu funkcjonowania instrumentu AKARI w fazie kriogenicznej (Matsuhara i in., 2006).

Wstępna szerokopasmowa optyczna obserwacja pola AKARI NEP-Wide została przeprowadzona za pomocą dwóch teleskopów: Canada-France-Hawaii Telescope (CFHT; Hwang i in., 2007) and Maidanak (Jeon i in., 2010). Te optyczne katalogi miały dwa główne ograniczenia: po pierwsze, głębokość obu katalogów była niewystarczająca, aby dopasować optyczne odpowiedniki do podczerwonych źródeł w katalogu AKARI. Co więcej, te dwa przeglądy optyczne charakteryzowały się różnymi głębokościami i krzywymi transmisji pasmowej, co uniemożliwiło jednorodną analizę wielozakresowego katalogu AKARI NEP-Wide (Goto i in., 2017; Kim i in., 2021b). Aby rozwiązać ten problem, zaproponowano głęboki jednolity przegląd optyczny za pomocą instrumentu SUBARU/HSC (Goto i in., 2017). Obserwacje zostały wykonane w dwóch kampaniach. Pierwsze obserwacje zostały wykonane w paśmie r w sierpniu 2015 roku i ucierpiały z powodu złej widoczności, jak również z powodu błędów otwarcia kopuły, dając płytsze obserwacje w porównaniu z pozostałymi pasmami (g, i, z i Y), które zostały wykorzystane w drugiej kampanii obserwacyjnej w 2018

roku (Oi i in., 2021). Redukcja danych została przeprowadzona za pomocą oficjalnego oprogramowania teleskopu hscPipe 6.5.3 (Bosch i in., 2018). Uzyskane dane zostały przedstawione w fotometrycznym układzie magnitud Cmodel AB, który jest stosunkowo odporny na fluktuacje spowodowane warunkami widoczności (Huang i in., 2018). Porównanie ostatecznego katalogu SUBARU/HSC z wcześniejszymi danymi CFHT i Maidanak pokazuje, że nowe obserwacje HSC są o 1,7 - 2,5 mag głębsze w pasmach g, r, i i z. Nowy połączony katalog jest bogatszy o ∼20 000 źródeł AKARI dopasowanych do odpowiedników optycznych. Większość z tych obiektów to przypuszczalnie galaktyki o niskiej jasności (Kim i in., 2021b).

Łącznie 111 535 źródeł AKARI zostało dopasowanych z promieniem 3*sigma* (odpowiadającym 1".78) do optycznych odpowiedników SUBARU/HSC. Spośród tych obiektów wybrano katalogi, czystych dopasowań (91 861) i niepoprawnych dopasowań (19 674). Oprócz ograniczeń otrzymanych z jakości dopasowania, do pasm NIR AKARI zastosowano limity jasności obserwowanej, jasny koniec rozkładu jest bowiem zajęty w dużej mierze przez źródła fałszywe oraz źle zmierzone gwiazdy (Kim i in., 2012). Nowy, połączony katalog został również uzupełniony o poprawione fotometryczne estymacje przesunięcia ku czerwieni (Ho i in., 2021), spektroskopowe dane uzupełniające (Shim i in., 2013) oraz dane wielozakresowe od promieniowania rentgenowskiego do submilimetrowego, jak zostało to opisane w poprzednim rozdziale.

3.3 Próbki treningowe i generalizacyjne

W tym rozdziale opisano metody tworzenia próbki treningowej i wyznaczania limitu próbki generalizacyjnej. Obie te metody, zaproponowane, zaplanowane i zaimplementowane przez autora tak, żeby osiągnąć jak największą dokładność klasyfikacji, zostały opisane i opublikowane w Poliszczuk i in. (2021). Zaproponowany tu specyficzny sposób tworzenia próbki treningowej (patrz rozdział 3.3.1) pozwala na pośrednie wprowadzenie informacji MIR do struktury modelu klasyfikacyjnego w fazie uczenia. W pierwszej publikacji Poliszczuk i in. (2019) przeprowadzono badania nad ryzykiem ekstrapolacji podczas klasyfikacji podczerwonych danych fotometrycznych. Wykazała ona trudności w kontrolowaniu wydajności modelu w obszarach próbki generalizacyjnej, które nie były reprezentowane przez dane treningowe. Aby uniknąć ekstrapolacji i związanego z nią obniżenia niezawodności modelu, na próbkę generalizacyjną narzucono ograniczenia stworzone przez algorytm MCD (patrz rozdział 3.3.2). Metoda ta nigdy wcześniej nie była stosowana w astronomii. Pozwala ona na precyzyjne kontrolowanie wydajności modelu i uniknięcie ryzyka ekstrapolacji.

3.3.1 Próbka treningowa

W niniejszej pracy selekcja AGN-ów dokonywana jest na podstawie szerokopasmowej fotometrii w zakresach optycznym i NIR. W przypadku nadzorowanych algorytmów uczenia maszynowego (patrz rozdział 3), potrzebna jest również oznaczona próbka ze znanymi etykietami klas, zwana *próbką treningową*, aby wytrenować model klasyfikacyjny i później przewidzieć etykiety na próbce bez etykiet za pomocą procedury generalizacji. Ta ostatnia nieoznakowana próbka jest nazywana *próbką generalizacyjną*. Większość etykiet klas pochodzi z optycznych obserwacji spektroskopowych wykonanych w ramach przeglądu AKARI NEP-Wide (Shim i in., 2013) przez instrumenty MMT/HECTOSPEC (Fabricant i in., 2005) i WYIN/HYDRA (Barden i in., 1993), wraz z niewielką liczbą dodatkowych widm otrzymanych przy pomocy

spektrografów Keck/DEIMOS (Faber i in., 2003), GTC/OSIRIS (Cepa i in., 2000) i SUBARU/FMOS (Kimura i in., 2010), przez członków zespołu AKARI NEP.

W głównym przeglądzie spektroskopowym opisanym w Shim i in. (2013) obserwowano dwie kategorie obiektów. Pierwszą grupę stanowiły główne cele spektroskopowe, które zostały wybrane jako obiekty jasne w pasmach MIR AKARI ($S11 < 18.5$ mag i $L15 < 17.9$ mag). Na ten podzbiór nałożono dodatkowy limit optyczny w filtrze R teleskopu Maidanak ($16 < R < 21-22.5$ mag w zależności od spektrografu), aby wybrać obiekty wystarczająco jasne w świetle optycznym do obserwacji spektroskopowych. Drugą grupę stanowiły cele drugorzędne określonych klas. W szczególności kandydaci na AGN w Shim i in. (2013) zostali wybrani zgodnie z metodą ograniczeń kolorów NIR i MIR opracowaną dla danych AKARI NEP-Deep (Lee i in., 2007). Metoda ta definiuje przestrzeń kolorów zdominowaną przez AGN-y jako:

$$\begin{cases} S11 < 18.5\,\mathrm{mag}_{AB}, \\ N2 - N4 > 0, \\ S7 - S11 > 0. \end{cases} \tag{3.1}$$

Granice te pozwoliły na wyselekcjonowanie obiektów, które z dużym prawdopodobieństwem charakteryzują się silną emisją pyłową z torusa oraz wykładniczym kształtem widma w zakresie NIR-MIR. W dalszej części pracy obiekty te, z potwierdzoną klasą spektroskopową, jak również z co najmniej jedną linią emisyjną o szerokości połówkowej (ang. *full width at half maximum*, FWHM) większej niż 1000 km/s (Shim i in., 2013) będą określane jako *AGN1*. Należy zwrócić uwagę na dwa ważne aspekty konstrukcji próbki treningowej. Po pierwsze, gwiazdy obecne w polu NEP (głównie w jasnej części rozkładu NIR) zostały usunięte z połączonego katalogu AKARI/HSC za pomocą procedur opisanych w Kim i in. (2021b), dlatego w tej pracy wykonujemy dwuklasową separację galaktyk i AGN-ów. Po drugie, optycznie potwierdzona próbka treningowa AGN-ów została wybrana przy użyciu metod opartych na kolorach MIR. Tak więc własności MIR tych AGN-ów pośrednio wpływają na własności i rozkład próbki w pasmach optycznych i NIR.

Dodatkowy zestaw etykiet klasowych został zaczerpnięty z obserwacji rentgenowskich wykonanych przez teleskop Chandra w polu AKARI NEP-Deep (Krumpe i in., 2015). W tym przypadku obiekty o dużej jasności w zakresie rentgenowskim, zdefiniowanej jako $\log L_X > 41.5$ erg/s w zakresie 0.5-7 keV, zostały określone jako rentgenowskie AGN-y. Taka granica jasności pozwala na zaliczenie do tej grupy zarówno galaktyk Seyferta, jak i kwazarów. Klasa ta będzie określana jako XAGN.

Ostateczne dane treningowe (i generalizacyjne) były wybrane z dodaktowym warunkiem detekcji we wszystkich filtrach SUBARU/HSC (g, r, i, z, Y), oraz wszystkich filtrach NIR AKARI/IRC ($N2$, $N3$, $N4$). W rezultacie otrzymano próbkę treningową składającą się z 1547 obiektów: 1348 galaktyk i 199 AGN-ów (163 AGN1 i 36 XAGN). Tabela 3.1 pokazuje, jak zmienia się liczba obiektów wykrytych w poszczególnych pasmach przy przechodzeniu z zakresu optycznego do MIR. O ile zmiana ta nie wydaje się aż tak istotna w przypadku próbki oznaczonej (wstępna selekcja obiektów do obserwacji spektroskopowych była ograniczona do jasnych źródeł), o tyle cały katalog doznaje poważnego ograniczenia przy przechodzeniu do pasm MIR. Liczba detekcji w pasmie $L24$ stanowi zaledwie $\sim$3% katalogu źródeł zaobserwowanym w paśmie optycznym g. Inną widoczną tendencją jest znaczny spadek liczby obiektów przy przejściu z zakresu NIR do MIR. Tabela ta doskonale ilustruje konieczność znalezienia bardziej efektywnego mechanizmu selekcji AGN-ów niż metody oparte na danych MIR.

Filtr	Cały katalog	Próbka oznaczona
g	**89 835**	**1 870**
r	**89 431** (88 642)	**1 869** (1 867)
i	**87 385** (86 186)	**1 864** (1 860)
z	**89 028** (86 023)	**1 871** (1 859)
Y	**86 587** (84 874)	**1 862** (1 856)
N2	**61 679** (59 845)	**1 650** (1 637)
N3	**74 475** (54 152)	**1 743** (1 598)
N4	**66 134** (45 841)	**1 722** (1 547)
S7	**5 041** (4 168)	**998** (918)
S9	**9 316** (3 536)	**1 360** (882)
S11	**9 147** (3 167)	**1 320** (843)
L15	**8 688** (2 404)	**1 070** (729)
L18	**10 258** (2 294)	**1 131** (704)
L24	**2 450** (1 208)	**520** (437)

TABLICA 3.1: Liczba obiektów wykrytych w poszczególnych pasmach SUBARU/HSC i AKARI/IRC. Liczby w nawiasach oznaczają źródła z pomiarami istniejącymi we wszystkich poprzednich pasmach odpowiadających krótszym długościom fali. Cały katalog odnosi się do wszystkich obiektów obecnych w katalogu czystych źródeł (Kim i in., 2021b). Próbka oznaczona odnosi się do obiektów z istniejącą klasą spektroskopową (Shim i in., 2013) lub z silną emisją promieniowania rentgenowskiego wykrytą przez teleskop Chandra (Krumpe i in., 2015).

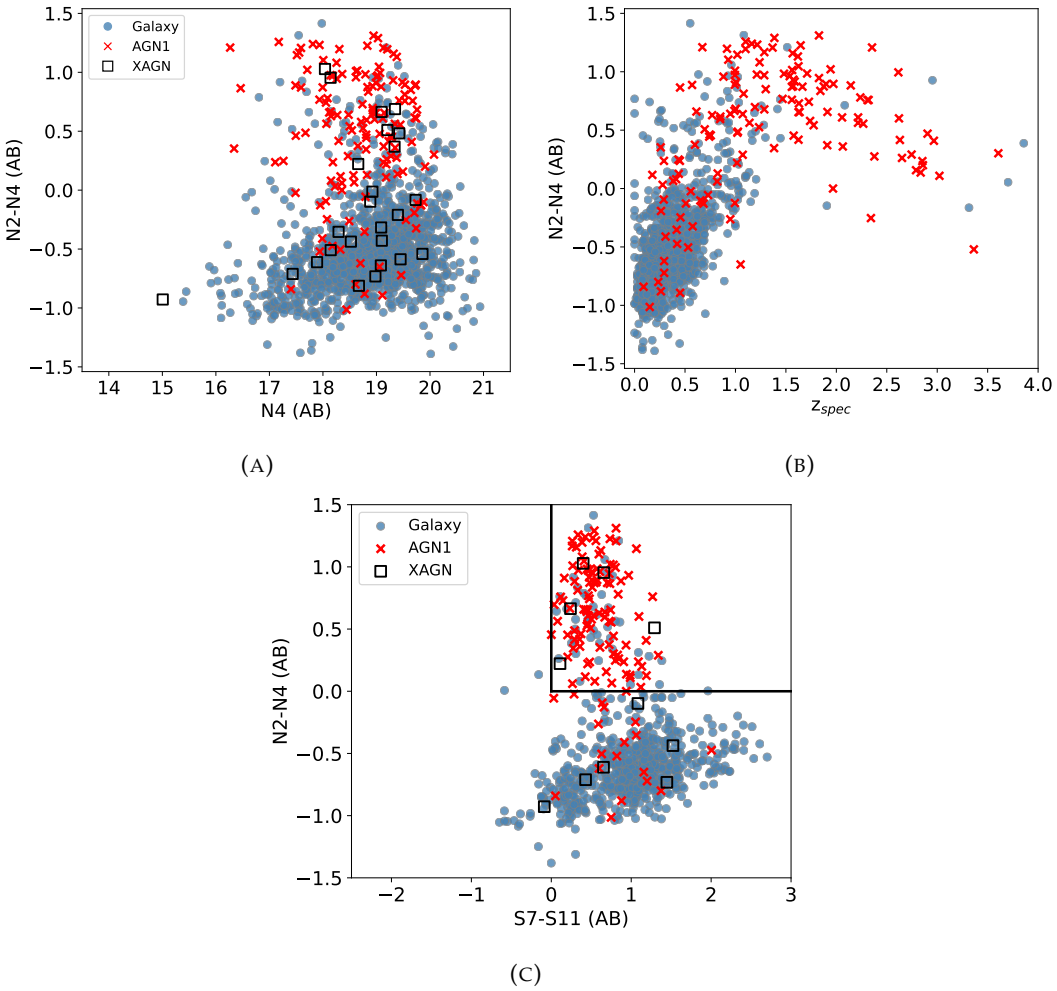
Aby przeanalizować główne własności próbki treningowej, spójrzmy na Rysunek 5.3, gdzie rozkłady klas są analizowane pod względem wartości koloru $N2$–$N4$. Wybór tego konkretnego koloru nie jest przypadkowy. Różnica pomiędzy jasnością obiektu w pasmach $N2$ i $N4$ może pokazać nam dwie ważne własności: stromość wykładniczego kształtu widma AGN w zakresie 3–8 μm, jak również obecność cechy widmowej 1,6 μm w przypadku SFG o wysokim przesunięciu ku czerwieni. Ze względu na wrażliwość na powyższe cechy obiektów jak i dobrą separację między klasami AGN-ów i galaktyk, wykres koloru $N2$–$N4$ połączony z wielkością gwiazdową w pasmie $N4$ jest często używany do wizualizacji danych AKARI (Lee i in., 2007; Lee i in., 2009). Jak widzimy na Rysunku 3.1a, $N2$–$N4$ daje dobre rozróżnienie pomiędzy głównym położeniem klasy AGN1 a centrum rozkładu galaktyk. Ze względu na brak gwiazd w katalogu, obszar zajmowany przez gwiazdy jest pusty na tym wykresie. Klasa gwiazdowa powinna znajdować się w lewym dolnym rogu, czyli w miejscu dla obiektów charakteryzujących się dużym strumieniem w paśmie NIR i niebieskim kolorem $N2$–$N4$. Porównując Rysunek 3.1a z Rysunkiem 3.1b widzimy, że główne zanieczyszczenie obszaru zajmowanego przez AGN-y w rozkładzie $N2$–$N4$ pochodzi od SFG o wysokim przesunięciu ku czerwieni, jak to zostało omówione w poprzednim rozdziale. Emisja termiczna pyłu z torusa jest podstawową cechą wykorzystywaną w tradycyjnej technice selekcji MIR AGN zastosowanej dla danych z AKARI (Eq. 3.1). Wyniki tej klasyfikacji są przedstawione na Rysunku 3.1c. W tym przypadku warunki selekcji AGN-ów zostały nieco złagodzone, tzn. usunięto wymóg $S11 < 18,5$, dzięki czemu dodatkowa niewielka próbka potwierdzonych spektroskopowo AGN-ów i XAGN-ów została włączona do schematu klasyfikacji. Brak tego ograniczenia nie wpływa silnie na efektywność klasyfikacji i nadal można zauważyć wyraźny podział na klasy galaktyk i AGN-ów.

Osobny problem stanowi rozkład próbki XAGN. Jak widać, nie jest ona zlokalizowana w dobrze określonym miejscu. Źródła XAGN są rozrzucone po całej przestrzeni przypisanej zarówno do klas AGN-ów, jak i galaktyk. Główną przyczyną takiego zachowania jest różnica między selekcją AGN w zakresie promieniowania rentgenowskiego i podczerwieni. Jak zostało to omówione w poprzednim rozdziale, selekcja AGN-ów w zakresie NIR-MIR może zidentyfikować tylko AGN-y o wysokim współczynniku Eddingtona (tzn. $L/L_{Edd} \geq 0.01$). Tak więc AGN-y, które są zasilane przez nieefektywne radiacyjnie przepływy akrecyjne (często nie wykazują one również obecności szerokich linii emisyjnych Trump i in., 2009), mogą wymykać się selekcji NIR-MIR, jak ma to również miejsce w przypadku klasyfikacji optycznej (Donley i in., 2012).

Ostatnim zagadnieniem, które zostanie poruszone w tym rozdziale, jest dokładność estymacji fotometrycznych przesunięć ku czerwieni. Fotometryczne przesunięcia ku czerwieni użyte w tej pracy zostały zamieszczone w połączonym katalogu SUBARU/HSC - AKARI/IRC i zostały opisane w Ho i in. (2021). Zostały one uzyskane za pomocą metody dopasowania χ^2 szablonów widmowych rozkładów energii kodu Le Phare (Arnouts i in., 1999; Ilbert i in., 2006) i bazowały na wielu pasmach fotometrycznych z zakresu UV-IR. Porównanie spektroskopowego przesunięcia ku czerwieni i jego fotometrycznego oszacowania zostało przedstawione na rysunku 3.2. Jest ono oparte na metodach opisanych w Ilbert i in. (2006). Widzimy dobrą zgodność między spektroskopowym i fotometrycznym przesunięciem ku czerwieni przy niższych wartościach ($z < 1,5$) zarówno dla galaktyk, jak i AGN-ów. Dalszy zakres przesunięć ku czerwieni, zajmowany głównie przez klasę AGN, wykazuje znaczne odchylenie w kierunku niższych przesunięć ku czerwieni. Rozbieżność ta jest spowodowana brakiem odpowiednio dobrych szablonów widmowego rozkładu energii AGN-ów używanych do fotometrycznej estymacji przesunięć ku czerwieni, co sprawia, że analiza własności przesunięć ku czerwieni w katalogu kandydatów na AGN-y jest zagadnieniem bardzo trudnym.

3.3.2 Próbka generalizacyjna i ograniczenie MCD

Z samej natury uczenia nadzorowanego wynika, że model klasyfikacyjny może rozpoznawać tylko te obserwacje, które są reprezentowane w próbce treningowej. Ta cecha prowadzi do dwóch ważnych konsekwencji. Po pierwsze, możemy manipulować własnościami próbki treningowej, aby uwypuklić pewne specyficzne zachowanie klasyfikatora. Do tej klasy manipulacji należą zastosowania wag do obiektów w próbce treningowej oraz zastosowanie technik selekcji AGN-ów w oparciu o dane MIR do stworzenia próbki treningowej. Drugą konsekwencją jest niezdolność klasyfikatora do właściwej ekstrapolacji przewidywań poza obszar przestrzeni cech zajmowany przez próbkę treningową. Ta druga konsekwencja nakłada ważne ograniczenie na próbkę generalizacyjną. Jej właściwości w przestrzeni cech nie powinny znacząco różnić się od właściwości próbki treningowej. Sytuacja, w której próbki są znacząco różne jest problematyczna, ponieważ klasyfikator nie posiada informacji o regionach poza próbką treningową i nie można kontrolować jego zachowania w tych regionach. Bez obserwacji danych treningowych nie można obliczyć żadnej metryki oceny klasyfikacji. Tak więc zarówno model, jak i naukowiec, który go trenuje, pozostają ślepi na te zewnętrzne regiony próbki generalizacyjnej. Ograniczenia nakładane na dane astronomiczne mają zwykle postać cięć dokonywanych w przestrzeni kolorów i wielkości gwiazdowych. Jednak choć takie podejście pozwala zachować prostą postać funkcji selekcji, nie spełnia ono ogólnych wymagań uczenia nadzorowanego. Głównym powodem jest to, że takie proste cięcia nie mogą poprawnie określić granic

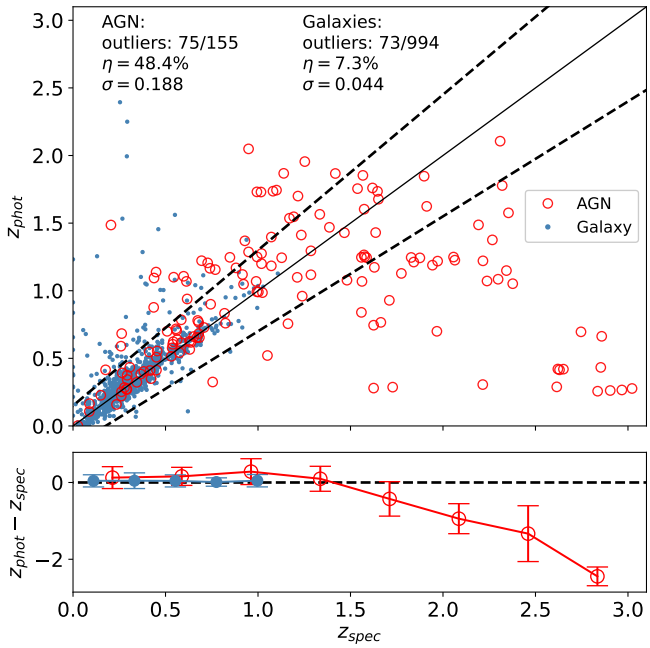


RYSUNEK 3.1: Właściwości próbki treningowej. *Panel A:* Wykres kolor-wielkość gwiazdowa danych treningowych $N2$–$N4$ vs $N4$. *Panel B:* Wykres koloru $N2$–$N4$ względem spektroskopowego przesunięcia ku czerwieni. *Panel C:* Wykres kolorów $N2$–$N4$ vs $S7$–$S11$ użyty w Lee i in. (2007) do selekcji kandydatów na AGN-y w danych AKARI MIR.

zbioru w wielowymiaroweej przestrzeni cech wyznaczonych przez próbkę treningową. Z tego powodu w próbce generalizacyjnej mogą pozostawiać obszary, w których klasyfikator musiałby dokonywać ekstrapolacji, aby przypisać obiekty do poszczególnych klas.

Niniejsza praca jest próbą spełnienia obu tych wymagań: skonstruowania efektywnego ograniczenia dla próbki generalizacyjnej w wielowymiarowej przestrzeni cech oraz zachowania względnie prostej formy tego ograniczenia, która umożliwiłaby rekonstrukcję funkcji selekcji. Aby osiągnąć oba te cele, wykorzystano algorytm estymacji minimalnego wyznacznika kowariancji (ang. *minimum covariance determinant estimator algorithm*, MCD, Rousseeuw i Driessen, 1999). Metoda MCD pozwala dopasować do zbioru danych treningowych wielowymiarową elipsoidę i ograniczyć kształt próbki generalizacyjnej do zakresu tej elipsoidy. Metoda MCD ma jeden wolny parametr α zwany *współczynnikiem zanieczyszczenia*. Parametr ten kontroluje ilość danych (w naszym przypadku obserwacji z próbki treningowej) dopuszczalnych poza elipsoidą podczas procesu jej dopasowania.

Aby uzyskać kształt odzwierciedlający rozkład próbki treningowej, dopasowano



RYSUNEK 3.2: Porównanie między spektroskopowym przesunięciem ku czerwieni a jego fotometrycznym oszacowaniem na podstawie danych z Ho i in. (2021) dla oznaczonych galaktyk (niebieskie kropki) i AGN-ów (czerwone kółka). Stożek utworzony przez linie przerywane odnosi się do $z_{phot} = z_{spec} \pm 0.15 \times (1 + z_{spec})$. Wartość η opisuje frakcję wartości odstających (lub błędów katastrofalnych) zdefiniowanych jako obiekty poza stożkiem. Sigma jest znormalizowaną medianą odchylenia bezwzględnego zdefiniowaną jako $\sigma = 1.48 \times \text{median}(|\Delta z|/(1+z))$. Dolny wykres przedstawia średnie wartości różnic między wartością rzeczywistą a oszacowaniem wraz z odchyleniami standardowymi. Pokazane są tylko obiekty $z < 3$. Wykres pochodzi z pracy Poliszczuk i in. (2021).

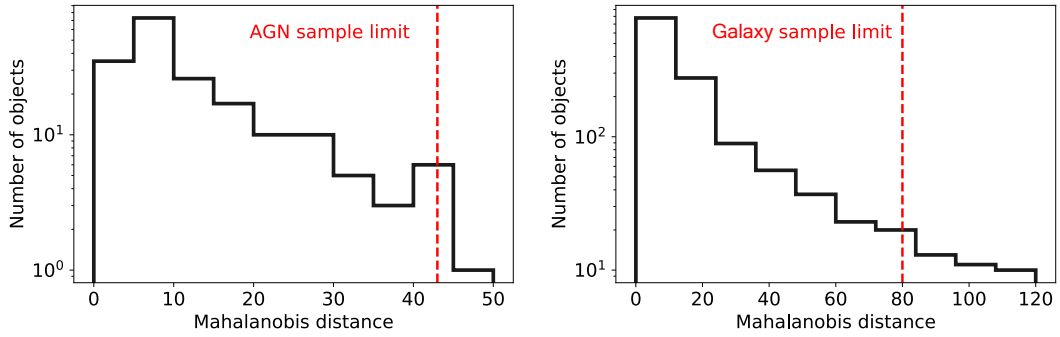
dwie oddzielne elipsoidy osobno dla klasy galaktyk i klasy AGN-ów. Zrobiono to w celu uniknięcia dwóch problemów. Po pierwsze, klasy galaktyk i AGN mogą mieć bardzo różne rozkłady w przestrzeni wielowymiarowej i efektywne dopasowanie jednej elipsoidy może być niemożliwe. Po drugie, duża różnica w liczebności obu klas może prowadzić do dalszego zmniejszania się liczby obserwacji AGN-ów podczas procesu dostrajania parametrów . Poszukiwanie odpowiedniej wartości parametru przeprowadzono poprzez analizę rozkładu odległości Mahalanobisa. Odległość Mahalanobisa d_M jest zdefiniowana jako:

$$d_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}, \tag{3.2}$$

gdzie $\vec{x}$ wskazuje ku miejscu obserwacji, $\vec{mu}$ jest wektorem średnich wartości cech, a Σ jest macierzą kowariancji. Na rysunku 3.3 przedstawiono histogramy odległości Mahalanobisa dla klas AGN i galaktyk. Wybór wartości α jest subiektywny i zależy od konkretnego zastosowania algorytmu MCD. Nie było potrzeby stosowania dużych wartości , co odpowiada bardzo konserwatywnym limitom. Głównym celem limitu MCD było uniknięcie ekstrapolacji w fazie generalizacji. Dlatego wartość została dobrana tak, aby odpowiadała zakresowi d_M, w którym zaczynają występować nieciągłości histogramu lub duży spadek liczby obiektów. Wybrano wartość $\alpha = 0.065$ ($d_M \simeq 43$) oraz $\alpha = 0.05$ ($d_M \simeq 80$) odpowiednio dla klas AGN-ów i galaktyk. Dwie

różne elipsoidy z odpowiadającymi im wartościami zostały dopasowane do klas AGN-ów i galaktyk, a kształt próbki generalizacyjnej został ograniczony do tych elipsoid. W ten sposób ostateczna próbka generalizacyjna składa się z obiektów, które należą do co najmniej jednej z elipsoid. Granice te zostały utworzone w przestrzeni wielkości gwiazdowych w pasmach optycznych i NIR, a nie w przestrzeni cech końcowych, w której dokonano przewidywań, aby zachować prosty kształt narzuconych ograniczeń.

Zastosowanie limitu MCD wraz z wymogiem detekcji w optycznych pasmach HSC i pasmach NIR IRC (który ogranicza rozmiar katalogu do 45 841 źródeł) daje nam ostateczną próbę generalizacyjną składającą się z 33 119 obiektów. Porównanie statystycznych właściwości próbek treningowej i generalizacyjnej jest przedstawione w tabeli 3.2. Widzimy tu, że wartości mediany w próbie generalizacyjnej są przesunięte w kierunku ciemnego końca rozkładu. Jest to spowodowane właściwościami próbki treningowej, w której nałożono dodatkowe warunki jasności w pasmach optycznych w celu wykonania pomiarów spektroskopowych. Jednocześnie zakresy rozkładów w poszczególnych pasmach próbki generalizacyjnej pozostają w granicach próbki treningowej (z wyjątkiem kilku przypadków, takich jak ciemny koniec pasma g). Inna analiza własności próbki generalizacyjnej przedstawiona jest na rysunkach 3.4, 3.5. Na rysunku 3.4 przedstawiono rozkład wielkości gwiazdowych w pasmach optycznych i NIR-MIR. Widzimy, że granica MCD odcina słaby koniec rozkładu, szczególnie w optycznej części widma elektromagnetycznego. Na Rysunku 3.5 pokazano rozkład koloru $N2$–$N4$. Widzimy tu bardzo silną redukcję obiektów charakteryzujących się wartością $N2$–$N4 = 0$, czyli część rozkładu koloru zajmowaną głównie przez galaktyki nieaktywne o niskim przesunięciu ku czerwieni.

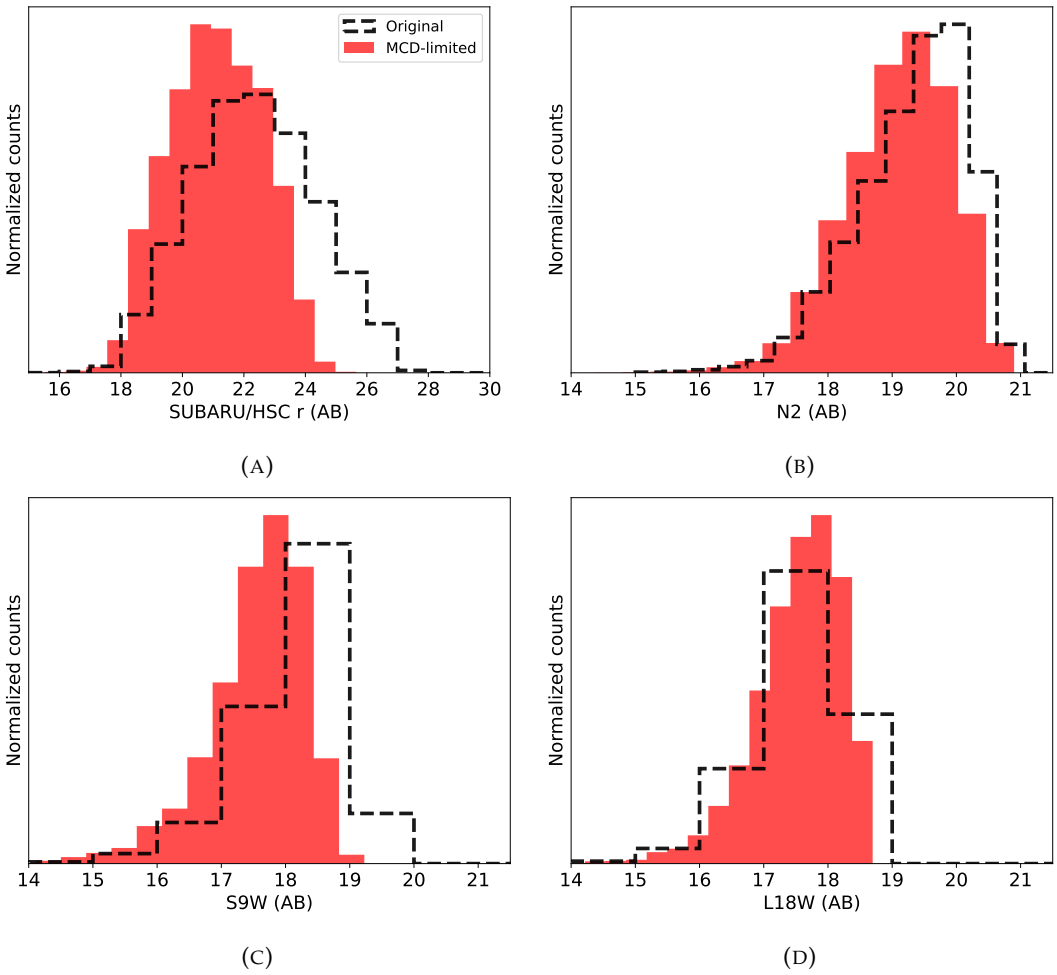


(A) (B)

RYSUNEK 3.3: Histogramy odległości Mahalanobisa dla próbek treningowych AGN-ów i galaktyk. *Panel a:* Próbka AGN-ów. *Panel b:* Próbka galaktyk. Przerywana czerwona linia odpowiada wartościom parametru α użytym do stworzenia ograniczających elipsoid. Wykresy pochodzą z pracy Poliszczuk i in. (2021).

	mediana	**MAD**	**min.**	**max.**
		próbka treningowa		
z_{spec}	0.339	0.308	0.001	4.320
z_{phot}	0.387	0.289	0.002	2.394
g	21.075	1.313	16.224	27.109
r	20.126	1.188	15.594	26.264
i	19.610	1.101	15.254	25.214
z	19.296	1.066	15.056	24.781
Y	19.119	1.059	14.850	24.278
N2	18.543	0.859	14.079	20.814
N3	18.692	0.732	14.528	20.638
N4	18.951	0.711	15.007	20.935
		próbka generalizacyjna		
z_{phot}	0.484	0.241	0.005	2.841
g	22.353	1.380	16.425	28.073
r	21.096	1.252	15.529	25.654
i	20.327	1.123	15.065	24.467
z	19.945	1.050	14.779	23.708
Y	19.752	1.015	14.586	23.430
N2	19.158	0.662	14.366	20.899
N3	19.325	0.554	14.829	20.979
N4	19.723	0.539	15.289	20.999

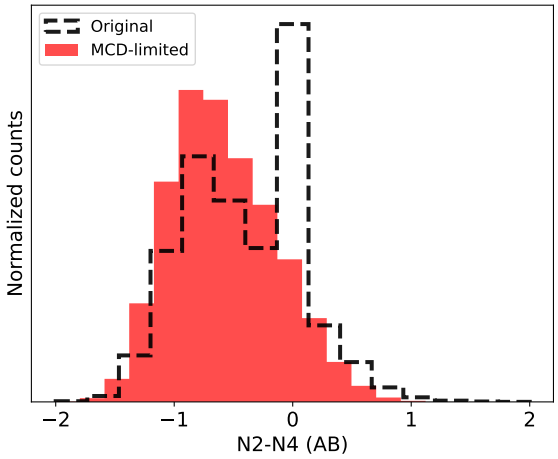
TABLICA 3.2: Własności statystyczne próbki treningowej i próbki generalizacyjnej. Przedstawione są wartości mediany, mediany odchylenia bezwzględnego (ang. *median absolute deviation*, MAD), minimalne i maksymalne wartości przesunięcia ku czerwieni i wielkości gwiazdowych w pasmach optycznych i NIR użytych podczas treningu i generalizacji.



(A) (B)

(C) (D)

RYSUNEK 3.4: Znormalizowane histogramy rozkładu wielkości gwiazdowych w pasmach optycznych, NIR i MIR w próbce generalizacyjnej z ograniczeniem MCD (czerwony, wypełniony histogram) i w oryginalnym wielozakresowym katalogu (Kim i in., 2021b) SUBARU/HSC-AKARI/IRC (czarna, przerywana linia). *Panel A:* SUBARU/HSC pasmo *r*. *Panel B:* AKARI/IRC pasmo *N*2. *Panel C:* AKARI/IRC pasmo *S*9*W*. *Panel D:* AKARI/IRC pasmo *L*18*W*.



RYSUNEK 3.5: Znormalizowany histogram rozkładu koloru $N2 - N4$ w próbce generalizacyjnej z ograniczeniem MCD (czerwony, wypełniony histogram) i w oryginalnym wielozakresowym katalogu (Kim i in., 2021b) SUBARU/HSC-AKARI/IRC (czarna, przerywana linia).

4

Techniki uczenia maszynowego

4.1 Klasyfikacja nadzorowana

W tym rozdziale omówimy koncepcję uczenia nadzorowanego i przedstawimy kilka algorytmów uczenia maszynowego stosowanych w niniejszej pracy. Sposób przedstawienia materiału został zaczerpnięty z Hastie, Tibshirani i Friedman (2001) oraz Bishop (2006). Uczenie nadzorowane (ang. supervised learning) jest dziedziną uczenia maszynowego, w której model jest trenowany na oznaczonych danych ze znanymi etykietami. *Etykiety* te (ang. *label* albo *target*), mogą być dyskretnymi wartościami w przypadku klasyfikacji lub ciągłymi w przypadku regresji. W uczeniu nadzorowanym dane oznaczone są reprezentowane przez dwa zestawy parametrów. Jeden zestaw odnosi się do etykiet danych. Może to być etykieta jednowymiarowa lub wektor etykiet w przypadku bardziej złożonych zadań uczenia. Etykieta będzie oznaczana przez y. Inny zestaw parametrów nazywany jest *zestawem cech*, jest on wspólny dla danych oznaczonych i nieoznaczonych i będzie oznaczany przez x. Tak więc każdy obiekt w danych oznaczonych, który będzie również nazywany *obserwacją*, jest określony przez parę (x_i, y_i).

Model nadzorowany próbuje nauczyć się zależności między cechami a etykietami na danych oznaczonych. Ten proces nazywamy *treningiem*. Model nauczony tego odwzorowania, może przewidywać etykiety dla danych nieoznaczonych, wykorzystując reprezentację tych danych w przestrzeni cech. Ten proces przewidywania etykiet na danych nieoznaczonych nazywamy *generalizacją*. Dane opatrzone etykietami nazywamy *próbką treningową*, zaś dane nieoznaczone używane podczas generalizacji - *próbką generalizacyjną*. W problemie klasyfikacji model nadzorowany próbuje nauczyć się, jak przypisywać dyskretne etykiety klasowe danym na podstawie ich właściwości w przestrzeni cech. Podstawową formą klasyfikacji, która jest wykorzystywana w tej pracy, jest rozróżnianie obiektów dwóch klas. Ten typ klasyfikacji jest znany jako klasyfikacja binarna, a klasy są często określane jako "pozytywna"(ang. *positive*) i ńegatywna"(ang. *negative*) z przypisanymi etykietami odpowiednio $y = 1$ i $y = -1$.

W celu znalezienia najbardziej odpowiedniego modelu dla konkretnego zadania uczenia się, należy przeprowadzić proces *wyboru modelu*. Po pierwsze, algorytm uczenia maszynowego powinien być testowany z różnymi kombinacjami hiperparametrów modelu. Hiperparametry określają specyficzne, przestrajalne właściwości algorytmu. Dobrą praktyką jest również testowanie różnych typów algorytmów uczenia maszynowego w celu znalezienia najlepszego z nich. Potrzebę dostrajania hiperparametrów i testowania różnych typów algorytmów uczenia można wyjaśnić

używając zagadnienia kompromisu między obciążeniem a wariancją (ang. *bias-variance tradeoff*). *Obciążenie modelu* opisuje siłę założeń, jakie model przyjmuje w odniesieniu do wykonywanych przez niego przewidywań. Na przykład modele liniowe mają tendencję do większego obciążenia niż bardziej elastyczne modele nieliniowe. Zbyt silne obciążenie prowadzi do sztywności modelu i niemożności dostosowania się do danych treningowych. Zjawisko to jest znane jako niewystarczające dopasowanie (ang. underfitting). Jednak zmniejszenie obciążenia modelu nie może być traktowane jako uniwersalne remedium na słabe wyniki klasyfikatora. Model, który będzie zbyt elastyczny, może zbyt dokładnie nauczyć się rozkładu danych treningowych. W tym przypadku model uczy się nie tylko rzeczywistych właściwości populacji klas, ale również szumu statystycznego obecnego w próbce treningowej. Zjawisko to znane jest jako nadmierne dopasowanie modelu (ang. overfitting)). Nadmiernie dopasowany model charakteryzuje się niskim obciążeniem i wysoką wariancją. Model o wysokiej wariancji jest bardzo wrażliwy na każdą zmianę w danych treningowych, co czyni go podatnym na uczenie się składnika szumu w rozkładzie danych. Taki model, mimo że ma dobre wyniki na danych treningowych, nie będzie dobrze generalizował. Trening modelu i późniejszy wybór modelu są ważnymi krokami pozwalającymi znaleźć kompromis pomiędzy obciążeniem i wariancją ostatecznego klasyfikatora.

4.1.1 Modele liniowe i regresja logistyczna

Jeżeli granice decyzyjne tworzone przez model między klasami są liniowe, to model taki nazywamy liniowym. W najprostszym podejściu liniowa granica decyzyjna może być utworzona poprzez dopasowanie modelu liniowego dla zmiennej docelowej:

$$f_i(x) = \theta_{i0} + \theta_{i1}^T x, \tag{4.1}$$

gdzie i to etykieta klasy, x to obserwacja, a θ_{i0}, θ_{i1} to parametry modelu. Hiperpłaszczyzna decyzyjna w przestrzeni cech między klasami 1 i 2 jest tworzona przez punkty, dla których $f_1(x) = f_2(x)$, czyli:

$$\{x : (\theta_{10} - \theta_{20}) + (\theta_{11} - \theta_{21})^T x = 0\}. \tag{4.2}$$

W tej pracy wykorzystano algorytm regresji logistycznej jako reprezentację rodziny liniowych modeli klasyfikacyjnych. Algorytm regresji logistycznej modeluje prawdopodobieństwo a posteriori jako liniową funkcję obserwacji. W przypadku klasyfikacji binarnej prawdopodobieństwo a posteriori jest modelowane jako:

$$\begin{aligned} P(k = 1|X = x) &= \frac{\exp(\theta_0 + \theta_1^T x)}{1 + \exp(\theta_0 + \theta_1^T x)} = p(x; \theta) \\ P(k = 2|X = x) &= \frac{1}{1 + \exp(\theta_0 + \theta_1^T x)} = 1 - p(x; \theta), \end{aligned} \tag{4.3}$$

gdzie $\theta = (\theta_0, \theta_1)$. W celu zachowania liniowych granic decyzyjnych można zastosować monotoniczną transformację logit:

$$\log(\frac{p}{1-p}) = \theta_0 + \theta_1^T x. \tag{4.4}$$

Tak więc hiperpłaszczyznę rozdzielającą dwie klasy (ang. *separating hyperplane*) tworzy zbiór punktów, dla których transformacja logitowa jest równa zeru. Model regresji

logistycznej jest dopasowywany metodą największej wiarygodności, wykorzystując warunkową funkcję wiarygodności w formie logarytmicznej (ang. *log-likelihood*) dla klasy k, względem obserwacji x. Z punktu widzenia uczenia maszynowego, zadaniem jest minimalizacja ujemnej funkcji logarytmu wiarygodności, co pozwala otrzymać funkcję błędu zwaną entropią krzyżową *cross-entropy*. Pracując na przykładzie klasyfikacji binarnej, gdzie $k_i = 1$ daje wartość docelową odpowiedzi $y_i = 1$, a $k_i = 2$ daje $y_i = 0$, logarytm funkcji wiarygodności dla N obserwacji definiuje się jako:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \log p_{k_i}(x_i, \theta) \\ &= \sum_{i=1}^{N} \left[y_i \log p(x_i, \theta) + (1 - y_i) \log(1 - p(x_i, \theta)) \right] \\ &= \sum_{i=1}^{N} \left[y_i \theta^T x_i - \log(1 + \exp(\theta^T x_i)) \right]. \end{aligned} \tag{4.5}$$

Tutaj zakładamy, że n-wymiarowy wektor wejściowy x_i staje się (n+1)-wymiarowy, gdzie w dodatkowym wymiarze zawarta jest wartość 1. Taka modyfikacja jest wprowadzana po to, aby łatwiej można było uwzględnić parametr $theta_0$. W ten sposób funkcja błędu może być zdefiniowana jako

$$E(\theta) = -l(\theta). \tag{4.6}$$

Aby zminimalizować funkcję błędu (i zmaksymalizować funkcję wiarygodności w formie logarytmicznej) przyjmujemy wartość pochodnej funkcji błędu równą zeru, otrzymując n+1 równań nieliniowych:

$$\frac{\partial E(\theta)}{\partial \theta} = \sum_{i=1}^{N} x_i \left[\log p(x_i, \theta) - y_i \right] = 0, \tag{4.7}$$

Funkcja błędu jest minimalizowana za pomocą techniki optymalizacji iteracyjnej Netwona-Raphsona (Fletcher, 1987).

$$\theta^{\text{new}} := \theta^{\text{old}} - \mathbf{H}^{-1} \partial_\theta E(\theta), \tag{4.8}$$

gdzie $\mathbf{H}$ jest macierzą Hessego utworzoną z drugich pochodnych funkcji błędu:

$$\mathbf{H} := \frac{\partial^2 E(\theta)}{\partial \theta \partial \theta^T} = \sum_{i=1}^{N} x_i x_i^T p(x_i, \theta)(1 - p(x_i, \theta)). \tag{4.9}$$

Po optymalizacji parametrów modelu, może być on użyty do ostatecznych przewidywań na próbce. Model można dodatkowo zmodyfikować, dodając człon regularyzacyjny, który często opiera się na normach L_1 lub L_2. Dodanie normalizacji nakłada ograniczenia na model w fazie uczenia i poprawia jego wydajność. W tym przypadku funkcja błędu zyskuje dodatkowy termin regularyzacyjny $R(\theta)$, którego postać zależy od strategii regularyzacji:

$$E(\theta) = \sum_{i=1}^{N} \left[y_i \theta^T x_i - \log(1 + \exp(\theta^T x_i)) \right] - C\, R(\theta). \tag{4.10}$$

Parametr C kontroluje siłę regularyzacji i podlega dalszemu dostrojeniu. Ponadto model regresji logistycznej może być dodatkowo optymalizowany za pomocą wag

opartych na stosunkach rozmiarów klas lub na specyficznych właściwościach poszczególnych obiektów, wprowadzając stałe wagi pod sumą występującą w funkcji wiarygodności.

4.1.2 Maszyna wektorów nośnych

Inna grupa metod klasyfikacji opiera się na poszukiwaniu hiperpłaszczyzny, która maksymalizuje margines między klasami w danych treningowych. Popularnym algorytmem, który wykorzystuje to podejście jest klasyfikator wektorów nośnych (ang. *support vector classifier*, SVC). Dla przypadku, w którym pełne rozdzielenie klas nie jest możliwe, problem optymalizacji SVC można zdefiniować jako

$$\begin{aligned} \max_{\theta_1,\theta_0,\|\theta_1\|=1} \quad & M \\ \text{subject to} \quad & y_i(\theta_1^T x_i + \theta_0) \geq M(1-\xi_i), \quad i = \overline{1,N} \\ & \forall i : \xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i < \text{const}, \end{aligned} \tag{4.11}$$

gdzie M oznacza margines między klasą a hiperpłaszczyzną rozdzielającą (margines między dwiema klasami jest równy $2M$). Parametr $\xi = \{\xi_i\}_{i=\overline{1,N}}$ kontroluje współczynnik błędnych klasyfikacji modelu. Dany punkt x_i jest błędnie zaklasyfikowany, gdy $xi_i > 1$. Wartość xi_i jest również proporcjonalna do odległości przesunięcia predykcji po niewłaściwej stronie marginesu w stronę błędnej klasy. Wprowadzając warunek $\sum_{i=1}^{N} \xi_i < \text{const}$, możemy kontrolować całkowitą ilość błędnych klasyfikacji na danych treningowych.

Równania 4.11 opisują standardowe sformułowanie klasyfikatora wektorów nośnych i tworzą wypukły problem optymalizacyjny (ang. *convex optimization problem*). Można go rozwiązać za pomocą metody mnożników Lagrange'a. Usuwamy wymóg unormowania parametru θ_1, definiując

$$M = \frac{1}{\|\theta_1\|}, \tag{4.12}$$

co przekształca problem 4.11 w minimalizację $\|theta_1\|$. Poprzez dodatkowe zastąpienie stałej, która kontroluje liczbę błędnych klasyfikacji przez przestrajalny parametr C, otrzymujemy równoważną formę problemu optymalizacyjnego 4.13, która jest bardziej użyteczna z obliczeniowego punktu widzenia:

$$\begin{aligned} \min_{\theta_1,\theta_0} \quad & \frac{1}{2}\|\theta_1\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & \forall i : \xi_i \geq 0, \quad y_i(\theta_1^T x_i + \theta_0) \geq 1 - \xi_i. \end{aligned} \tag{4.13}$$

Teraz możemy skonstruować pierwotną formę Lagrangianu

$$L_p = \frac{1}{2}\|\theta_1\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i[y_i(\theta_1^T x_i + \theta_0) - (1-\xi_i)] - \sum_{i=1}^{N} \mu_i \xi_i. \tag{4.14}$$

Po zminimalizowaniu L_p względem θ_1, θ i ξ_i oraz ustawieniu odpowiednich pochodnych na zero, otrzymujemy trzy warunki:

$$\begin{aligned} \theta_1 &= \sum_{i=1}^{N} \lambda_i y_i x_i, \\ 0 &= \sum_{i=1}^{N} \lambda_i y_i, \\ \lambda_i &= C - \mu_i. \end{aligned} \tag{4.15}$$

Stosując te warunki wraz z dodatkowym warunkiem dodatniej wartości mnożników Lagrange'a i xi_i, otrzymujemy dualną formę Lagrangianu:

$$L_d = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j, \tag{4.16}$$

co daje ograniczenie na problem minimalizacji przedstawiony w równaniu 4.13. Po zmaksymalizowaniu L_d z zastrzeżeniem $0 \leq \lambda_i \leq C$ i drugiego warunku z 4.15 otrzymujemy trzy dodatkowe warunki:

$$\begin{aligned} \lambda_i[y_i(\theta_1^T x_i + \theta_0) - (1 - \xi_i)] &= 0, \\ y_i(\theta_1^T x_i + \theta_0) - (1 - \xi_i) &\geq 0, \\ \mu_i \xi_i &= 0. \end{aligned} \tag{4.17}$$

Oba zestawy ograniczeń, 4.15 i 4.17 tworzą tak zwane warunki Karusha-Kuhna-Tuckera (KKT). Warunki te, wraz z 4.18 jednoznacznie charakteryzują rozwiązanie problemu optymalizacji. Ponadto warunki KKT definiują ważny podzbiór obserwacji charakteryzujących się niezerową wartością parametru λ_i, zwanych wektorami nośnymi (ang. *support vectors*). Niektóre z tych punktów z $\xi_i = 0$ znajdują się na marginesie i są wykorzystywane do stworzenia hiperpłaszczyzny separacji, podczas gdy wektory nośne z $\xi_i > 0$ będą znajdować sie poza marginesem.

Maszyna wektorów nośnych (ang. *support vector machine*, SVM Cortes i Vapnik, 1995) jest rozszerzeniem klasyfikatora wektorów nośnych (SVC). Klasyfikator wektorów nośnych jest metodą liniową, która działa w niezmodyfikowanej przestrzeni cech. Metoda ta może stać się bardziej elastyczna poprzez stworzenie sztucznych wymiarów składających się z kombinacji pierwotnych cech. W ten sposób model zacznie działać w sposób nieliniowy. W SVM te sztuczne cechy są wprowadzane za pomocą funkcji jądrowej (ang. *kernel function*) $k(x, x')$ zawierającej pierwotne cechy tylko w postaci iloczynów. Zastosowanie funkcji jądrowych do rozszerzenia przestrzeni cech jest znane w literaturze jako *kernel trick* (ang.) i nie wymaga wiedzy o dokładnym odwzorowaniu między początkową przestrzenią cech a nową, wielowymiarową przestrzenią cech (patrz Cortes i Vapnik (1995) w celu dogłębnego omówienia właściwości zastosowania funkcji jądrowych w algorytmie SVM). Wprowadzenie funkcji jądrowej przekształca dualną postać Lagrangianu w

$$L_d = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j k(x_i, x_j). \tag{4.18}$$

W danej pracy użyto gaussowskiej funkcji jądrowej (ang. *radial basis kernel function*, RBF), zdefiniowanej jako

$$k(x, x') = exp(-\gamma \|x - x'\|^2), \tag{4.19}$$

gdzie γ jest hiperparametrem modelu. Model SVM został wykorzystany w tej pracy w wersji tradycyjnej opisanej powyżej, jak również w wersji ważonej zwanej maszyną

wektorów nośnych opartą na logice rozmytej (ang. *fuzzy SVM*, Lin i Wang, 2002). W tym przypadku do problemu optymalizacji wprowadza się dodatkowe wagi s_i (ang. *fuzzy membership*):

$$\begin{aligned} \min_{\theta_1,\theta_0} \quad & \frac{1}{2}\|\theta_1\|^2 + C\sum_{i=1}^{N} s_i\xi_i \\ \text{s.t.} \quad & \forall i : \xi_i \geq 0, \quad y_i(\theta_1^T x_i + \theta_0) \geq 1 - \xi_i. \end{aligned} \tag{4.20}$$

W ten sposób różne obserwacje obecne w danych treningowych mogą mieć różny wpływ na proces uczenia. Takie wagi mogą być stosowane w postaci osobnych wag (ang. *instance weights*), gdzie każdy obiekt ma inną wagę. Mogą być również stosowane w formie wag klasowych (ang. *class weights*), gdzie obiekty mniejszej klasy mają większe wagi, aby zwiększyć ich znaczenie i przesunąć granicę decyzyjną modelu dalej od mniejszej klasy.

Oszacowanie prawdopodobieństwa a posteriori dla SVM uzyskuje się w podobny sposób, jak w przypadku metody regresji logistycznej. Klasyfikator SVM podaje odległość obiektu od hiperpłaszczyzny separującej. Aby skalibrować te odległości i przekształcić je w oszacowanie prawdopodobieństwa, do danych odległości dopasowywana jest funkcja sigmoidalna, podobnie jak w przypadku równania 4.3. Szczegółowy opis tej procedury można znaleźć w Platt (1999).

4.1.3 Metody zespołowe i drzewa decyzyjne

Inna popularna rodzina metod stosowanych do problemów klasyfikacji opiera się na konstrukcji drzew decyzyjnych. Większość algorytmów drzew decyzyjnych (w tym te, które są wykorzystywane w danej pracy) dzieli przestrzeń cech na prostokątne pola z przypisanymi etykietami poprzez zastosowanie rekurencyjnego podziału binarnego. W ten sposób każdy podział uzyskany przez węzeł zewnętrzny m (czyli węzeł bez istniejących węzłów potomnych) w drzewie może być utożsamiany z uzyskanym regionem R_m. Ten typ drzewa decyzyjnego jest określany jako *CART* (drzewo klasyfikacji i regresji, ang. *classification and regression tree*) i został opisany w Breiman i in. (1984).

Predykcja w danym regionie R_m jest modelowana jako etykieta klasy większościowej występującej w tym regionie w zbiorze treningowym. Zdefiniujmy N_m jako liczbę obserwacji w R_m i p_m jako prawdopodobieństwo zaobserwowania przedstawiciela pozytywnej klasy w R_m, które można oszacować poprzez odsetek obserwacji klasy pozytywnej w węźle m:

$$\hat{p}_m = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = 1). \tag{4.21}$$

Aby zdefiniować strukturę drzewa, musimy uruchomić dwa mechanizmy: jednym jest poszukiwanie optymalnego podziału binarnego dokonywanego w węźle, drugim zaś sposób na ustalenie optymalnej głębokości drzewa. Drzewo decyzyjne buduje się poprzez utworzenie węzła podziału, który wykorzystuje wybraną cechę z określoną wartością progową. Wartość progowa zastosowana do cechy dzieli region przestrzeni cech na dwa oddzielne podregiony, powyżej i poniżej wartości progowej. W najbardziej podstawowej formie, cecha i próg rozdzielający są wybierane w sposób obliczeniowo zachłanny (ang. *greedy method*) poprzez minimalizację miary nieczystości w węźle. Zazwyczaj wykorzystuje się do tego celu wskaźnik Giniego. Dla problemu dwuklasowego można go uprościć do postaci:

$$Q_m^{GiniIndex}(T) = \sum_{k \neq k'}^{K} \hat{p}_{mk}\hat{p}_{mk'} = 2\hat{p}_m(1 - \hat{p}_m), \tag{4.22}$$

gdzie $\hat{p}_{mk}$ to udział klasy k w węźle m.

Tworzenie w pełni rozwiniętego drzewa nie zawsze jest optymalną strategią. Wielkość drzewa jest traktowana jako hiperparametr modelu, który kontroluje złożoność (a tym samym wariancję) modelu i może być dostrajany. Prostym podejściem może być zakończenie wzrostu drzewa, gdy dokładność przewidywań nie poprawia się przy dodatkowych podziałach. Taka metoda może nie dawać optymalnych wyników, ponieważ czasami znaczna poprawa wydajności może nastąpić dopiero po kilku podziałach. Aby uniknąć tego problemu i lepiej kontrolować parametr wielkości drzewa, stosuje się metodę zwaną ograniczeniem w oparciu o koszt i złożoność modelu (ang. *cost-complexity pruning*). Oznaczmy w pełni rozwinięte drzewo jako T_0, a T jako dowolną przyciętą wersję drzewa T_0. Niech $|T|$ oznacza liczbę węzłów końcowych w T, gdzie *węzeł końcowy* (lub zewnętrzny) odnosi się do węzła bez węzłów potomnych. Zdefiniujmy kryterium kosztu złożoności (ang. *cost-complexity criterion*) jako

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|, \tag{4.23}$$

gdzie m odnosi się do regionu R_m, N_m jest liczbą obserwacji w tym regionie, Q_m jest miarą zanieczyszczenia węzła, a C jest parametrem podlegającym dostrojeniu. Metoda ograniczenia w oparciu o koszt i złożoność modelu poszukuje takiego poddrzewa, które może najbardziej efektywnie zminimalizować C. Parametr dostrajania pozwala na kontrolę kompromisu między obciążeniem modelu a jego wariancją. Kontrola nad wielkością drzewa i jego ogólnymi właściwościami może być dalej rozwijana poprzez zastosowanie wag opartych na klasach i pojedynczych obiektach. Można to zrobić w trakcie procedury wzrostu drzewa. Prawdopodobieństwo $\hat{p}_m$ uzyskuje się przez obliczenie obiektów z określonej klasy w regionie m. Można więc nadać wagi tym obiektom, zanim zostaną one użyte do oszacowania $\hat{p}_m$. Taka modyfikacja wpłynie na miarę zanieczyszczenia podczas wzrostu drzewa, a w konsekwencji zmodyfikuje właściwości klasyfikatora. Takie samo podejście do ważenia stosuje się w przypadku zespołów drzew, o których mowa w kolejnych akapitach.

Chociaż drzewa decyzyjne osiągają bardzo dobre wyniki na danych treningowych, to w ogólności charakteryzują się wysoką wariancją i tendencją do nadmiernego dopasowania (Breiman i in., 1984). W celu przezwyciężenia tego problemu często stosuje się metodę uśredniania wyników z zespołu drzew decyzyjnych, aby zmniejszyć wariancję ostatecznego modelu. Metoda ta funkcjonująca w literaturze anglojęzycznej pod nazwą *bagging* polega na trenowaniu zespołu modeli na podzbiory próbek treningowych losowanych z zastąpieniem (ang. *bootstrap samples*) i uśrednianiu przewidywań modeli w celu uzyskania ostatecznej predykcji. Jeżeli więc zdefiniujemy zbiór B próbek losowanych z zastąpieniem i z każdego drzewa otrzymamy predykcję dla danej obserwacji, to ostateczne przewidywanie klasy otrzymamy poprzez głosowanie większością głosów.

Zadanie redukcji wariancji modelu opartego na drzewach decyzyjnych można rozwiąć jeszcze bardziej, tak jak to zrobiono w algorytmie lasu losowego (ang. *random forest*, Breiman, 2001). Wariancję średniego wyniku uzyskanego z B identycznie rozłożonych (ang. *identically distributed*) drzew definiuje się jako

$$\sigma_B^2 = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2, \tag{4.24}$$

gdzie σ^2 to wariancja pojedynczego drzewa, a ρ to korelacja między parami obiektów. Pierwszy człon nie może być zredukowany przez zwiększenie liczebności drzew w zespole B. W związku z tym korelacja między drzewami w zespole ogranicza ich zdolność do redukcji wariancji. Główną ideą lasu losowego jest połączenie metody *bagging-u* z dekorelacją drzew składowych w celu dalszej redukcji wariancji. Taka dekorelacja odbywa się poprzez losowy wybór podzbioru cech m przed każdym podziałem w drzewie (tzn. ten podzbiór cech jest używany do uzyskania najlepszej cechy i punktu podziału w każdym węźle). Dla zadań klasyfikacji domyślna wielkość podzbioru cech wynosi $\lfloor\sqrt{p}\rfloor$ gdzie p jest całkowitą liczbą cech wejściowych. Losowość w konstrukcji drzewa może być posunięta jeszcze dalej, tak jak to jest zrobione w algorytmie wyjątkowo losowych drzew (ang. *extremely randomized trees*, Geurts, Ernst i L., 2006), gdzie ostateczny próg wartości cechy używany w podziale w węźle uzyskuje się poprzez ustawienie losowych progów dla cech w podzbiorze m i przyjęcie najlepszego progu jako ostatecznego. Warto tutaj zauważyć, że zarówno struktura zespołów drzew decyzyjnych zarówno w algorytmie lasu losowego jak i w algorytmie wyjątkowo losowych drzew pozwala na uzyskanie prostego sposobu klasyfikacji probabilistycznej. W większości implementacji, szacowanie prawdopodobieństwa odbywa się poprzez głosowanie większością głosów w zespole drzew.

W tradycyjnym sformułowaniu metody *bagging*, wartość oczekiwana średniej z każdej z próbek uzyskanych metodą *bootstrap* jest taka sama, ponieważ składowe drzewa decyzyjne mogą być traktowane jako zmienne losowe o identycznym rozkładzie. Dlatego też obciążenie modelu w przypadku zespołu jest takie same jak w przypadku poszczególnych drzew składowych i nie może być obniżone. Jest to również prawdziwe dla lasów losowych i wyjątkowo losowych drzew. Poprawa predykcji następuje w tych modelach tylko poprzez redukcję wariancji.

Aby zmniejszyć obciążenie statystyczne modelu, należałoby zbudować zespół z drzew, które nie są identycznie rozłożone. Takie podejście jest stosowane w metodzie *boosting*. Algorytmy tego typu wykorzystują *słabe klasyfikatory* - tj. modele, które są tylko nieznacznie lepsze niż losowe przypisywanie etykiet klas do obserwacji. Takie słabe klasyfikatory są następnie ustawiane w sekwencję M, w której kolejne modele starają się skorygować przewidywania swoich poprzedników. Ostateczną predykcję modelu można wyrazić jako ważoną kombinację modeli składowych:

$$f(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m f_m(x, \beta_m).\right) \tag{4.25}$$

Dla przejrzystości tego równania zmieniliśmy etykiety klas z $\{0,1\}$ na $\{\pm 1\}$. Waga α_m opisuje udział modelu f_m w ostatecznej predykcji.

Podczas każdego kroku obserwacje w próbce treningowej są ponownie ważone, co zmusza kolejny klasyfikator f_m do skupienia się na problematycznych przypadkach. Tutaj $beta_m$ jest zbiorem parametrów opisujących model f_m. W przypadku drzewa decyzyjnego byłyby to cechy używane w węzłach oraz ich wartości używane do popdziału. Wersja modelu opartego na metodzie *boosting*, która wykorzystuje drzewa decyzyjne jako modele bazowe, jest określana literaturze anglojęzycznej jako *boosted trees*. Tradycyjna procedura tworzenia zespołów słabych klasyfikatorów z iteracyjnym ważeniem próbki treningowej, jaka została opisana powyżej jest procedurą bardzo kosztowną obliczeniowo. Dlatego w praktyce używa się innych metod opartych na optymalizacji numerycznej, takich jak *gradient boosting* (Hastie, Tibshirani i Friedman,

2001). W danej pracy wykorzystano bardzo popularną i efektywną implementację tego algorytmu o nazwie XGBoost (Chen i Guestrin, 2016).

Ponadto, niniejsza praca zawiera dwie dodatkowe metody wykorzystujące zespoły silnych klasyfikatorów, tj. w pełni rozwiniętych i dostrojonych modeli. Takie modele wejściowe mogą być klasyfikatorami dowolnego typu, które wykazały najlepszą wydajność. Te dodatkowe metody będą nazywane systemami głosującymi (ang. *voting schemes*). Pierwszy typ systemu głosującego będzie określany jako system twardego głosowania (ang. *hard voter*). Przyjmuje on predykcję większości modeli wejściowych jako ostateczne wynik klasyfikacji. Drugi typ systemu głosującego będzie określany jako klasyfikator stosowy (ang. *stacked classifier*). Jest to prosty model regresji logistycznej, który wykorzystuje oszacowania prawdopodobieństwa z modeli wejściowych jako zestaw cech i dokonuje predykcji na podstawie ich rozkładów. Podstawową ideą tych metod jest dalsza redukcja wariancji modelu końcowego i stabilizacja przewidywań w przypadku małej próbki danych treningowych.

4.2 Ocena wydajności

Skuteczność klasyfikacji dokonanej przez model uczenia maszynowego na nieoznaczonych danych, przeznaczonych do generalizacji, jest zwykle oceniana za pomocą metryki oceny (ang. *evaluation metric*) obliczonej na danych oznaczonych. Metryki obliczone na oznaczonych danych, które posłużyły do wytrenowania modelu, mogą wykazywać zbyt wysokie wartości z powodu możliwości wystąpienia nadmiernego dopasowania. Dlatego, aby uzyskać wiarygodne wyniki metryk, powinny one być obliczane na danych z etykietami, które są usuwane z procesu treningu modelu. Taka procedura jest określana jako podział na próbki treningową i testową (ang. *train-test split*). W takim podziale, klasyfikator jest trenowany na jednym podzbiorze danych z etykietami, zwanym zbiorem treningowym, i oceniany na innym, zwanym zbiorem testowym. W celu dalszego zwiększenia dokładności oceny i zmniejszenia wpływu losowego podziału danych z etykietami, procedura ta jest powtarzana kilka razy. Ta wieloetapowa ocena znana jest jako n-krotna walidacja krzyżowa (ang. *n-fold cross-validation*). W tym przypadku próbka oznaczona jest dzielona na n podzbiorów. Model trenowany jest na $(n-1)$ podzbiorach i sprawdzany na ostatnim z nich. Proces ten jest powtarzany n razy za każdym razem sprawdzając poprawność działania modelu na innym podzbiorze. Średnia wartość wszystkich wyników walidacji jest brana jako ostateczne oszacowanie wyniku metryki. Walidacja krzyżowa jest szczególnie ważna w przypadku małej ilości danych treningowych, gdzie ryzyko nadmiernego dopasowania jest większe. Dokładność oszacowania wyniku metryki na małych zbiorach danych można dodatkowo poprawić przeprowadzając walidację krzyżową wiele razy, tasując dane przed każdą iteracją podziału na n podzbiorów.

Innym problemem poruszonym w tej pracy jest uczenie się na danych o niezrównoważonych rozmiarach klas. Różna wielkość klas w próbce treningowej wpływa na dwa aspekty procesu uczenia modelu. Po pierwsze, ma to duży wpływ na sposób, w jaki model rozdziela klasy. Zazwyczaj, jeżeli jedna klasa jest znacznie mniejsza od drugiej, model ma tendencję do przesunięcia płaszczyzny decyzyjnej w kierunku mniejszej klasy. Może to przynieść korzyści ze względu na mniejsze zanieczyszczenie mniejszej klasy kosztem zmniejszenia jej kompletności. Po drugie, wpływa to na wartości metryk, przez co często zawyżają one efektywność działania modelu, nawet gdy wydajność modelu w małej klasie jest bardzo słaba. Dlatego w przypadku niezrównoważonych danych należy wybrać takie metryki oceny, które nie będą podatne na różnice w rozmiarach klas (dogłębną dyskusję na ten temat można znaleźć np. w:

Fernández i in., 2018). W tej pracy głównym celem było uzyskanie wiarygodnego katalogu kandydatów na AGN-y. Jak zostało to omówione w Rozdziale 3, próbka treningowa AGN-ów jest znacznie mniejsza niż próbka galaktyk. Dlatego położono nacisk na właściwą ocenę wydajności klasyfikatora na mniejszej klasie AGN-ów.

Od tej pory klasa AGN-ów będzie traktowana jako *pozytywna*, zaś klasa galaktyk jako *negatywna*. Z punktu widzenia astrofizyki, najbardziej powszechnymi miarami opisującymi właściwości katalogu są jego czystość i kompletność. Czystość klasy pozytywnej (AGN) jest znana w uczeniu maszynowym jako *precision* i jest definiowana jako

$$\text{Precision} = \frac{T_p}{T_p + F_p}. \tag{4.26}$$

Pokazuje ona, jaką część wszystkich obiektów zaklasyfikowanych jako pozytywne, tzn. prawdziwe pozytywne obiekty (ang. *true positives*) T_p i obiekty niepoprawnie zaklasyfikowane jako pozytywne (ang. *false positives*) F_p, stanowią prawdziwe obserwacje z klasy pozytywnej (T_p). Kompletność katalogu AGN-ów jest określana w literaturze jako *recall* i definiuje się ją jako

$$\text{Recall} = \frac{T_p}{T_p + F_n}. \tag{4.27}$$

Kompletność klasy pozytywnej (albo recall) definiuje stosunek prawidłowo sklasyfikowanych obserwacji z klasy pozytywnej (T_p) do całej próbki klasy pozytywnej, prawdziwie pozytywnych i fałszywie negatywnych. Aby zoptymalizować zarówno czystość, jak i kompletność pozytywnej klasy podczas treningu modelu oraz mieć skuteczny sposób na porównanie wydajności różnych modeli, można użyć bardziej złożonych metryk, które są zbudowane na podstawie czystości i kompletności. Aby móc efektywnie kontrolować wydajność modelu w klasie pozytywnej, popularnym i dobrze sprawdzonym wyborem jest metryka F1, która jest średnią harmoniczną czystości (precision) i kompletności (recall) pozytywnej klasy:

$$\text{F1} = 2 \times \frac{\text{Precison} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.28}$$

Inną metryką ściśle związaną z czystością i kompletnością pozytywnej klasy jest pole powierzchni pod krzywą stworzoną z tych dwóch metryk (ang. *precision-recall area under curve*, PR AUC). Krzywą taka jest tworzona poprzez obliczenie wyników czystości i kompletności dla różnych probablistycznych progów decyzyjnych. Tutaj pojawia się ważne ograniczenie tej metryki. Aby zdefiniować różne progi decyzyjne oparte na estymacji prawdopodobieństwa przynależności obserwacji do danej klasy, trzeba być w stanie uzyskać oszacowanie prawdopodobieństwa z modelu. Dlatego ta metryka może być stosowana tylko do algorytmów, które pozwalają na estymację prawdopodobieństwa. Obszar pod krzywą odnosi się do ogólnej skuteczności modelu, nie skupiając się na żadnym konkretnym progu decyzyjnym. Im większy obszar pod krzywą, tym lepszy jest model. Ostatnia metryka wykorzystana w tej pracy znana jest jako zrównoważona dokładność (ang. *balanced accuracy*, bACC) i jest mniej związana zarówno z czystością, jak i kompletnością pozytywnej klasy. Została ona użyta w celu sprawdzenia ogólnej wydajności modelu. Zrównoważoną dokładność definiuje się jako

$$\text{bACC} = 0.5 \times (\text{Precison} + \text{TNR}). \tag{4.29}$$

Skrót TNR w równaniu odnosi się do *true negative rate* i jest odpowiednikiem *precision* dla klasy negatywnej:

$$\text{TNR} = \frac{T_n}{T_n + F_n}. \tag{4.30}$$

4.3 Wykrywanie obserwacji odstających i wizualizacja wielowymiarowa

4.3.1 Wykrywanie obserwacji odstających za pomocą algorytmu lasu izolującego

Oprócz klasyfikacji nadzorowanej, niniejsza praca zajmuje się również problemem nienadzorowanego wykrywania obserwacji odstających. Nieco upraszczając, obserwację odstającą (ang. *outlier* lub *novelty*) można zdefiniować jako obserwację, która nie pasuje do ogólnych trendów w rozkładzie danych. Poszukiwanie tego typu obiektów może stanowić wyzwanie, szczególnie w przypadku wielowymiarowej przestrzeni cech lub dużego rozmiaru próbki. Tym problemem zajmuje się gałąź technik uczenia maszynowego stworzonych do wykrywania obiektów odstających.

Większość algorytmów wykrywania obserwacji odstających opiera się na tej samej ogólnej zasadzie. Najpierw model jest dopasowywany do danych, na których podstawie tworzy on profil typowych obserwacji. Następnie model próbuje znaleźć obserwacje, które nie pasują do tego profilu. W tej pracy wykorzystano zasadniczo inne podejście zastosowane w algorytmie lasu izolującego (ang. *Isolation Forest*, Liu, Ting i Zhou, 2008). Zamiast budować typowy profil obserwacji, dany algorytm bezpośrednio wykrywa obserwacje odstające. Ta własność sprawia, że algorytm Isolation Forest jest bardzo szybkim i skalowalnym narzędziem, często stosowanym w nowoczesnych metodach przetwarzania potokowego danych opartych na metodach uczenia maszynowego.

Podstawowa właściwość algorytmu Isolation Forest opiera się na prostym założeniu, że obserwacje odstające są, ogólnie rzecz biorąc, łatwiejsze do oddzielenia od reszty danych w porównaniu z obiektami typowymi dla danej próbki. Ta właściwość jest wykorzystywana w następujący sposób. Tworzy się las w pełni rozwiniętych binarnych drzew decyzyjnych. Podczas wzrostu drzewa tworzony jest podział węzłów poprzez losowy wybór cech podziału i odpowiadającej im wartości progowej. Aby uzyskać skuteczne wykrywanie obserwacji odstających, "stopień odstawania" (ang. *degree of anomaly*) od reszty próbki jest związany z głębokością drzewa i względną pozycją obserwacji w danej strukturze klasyfikacyjnej. W szczególności, długość ścieżki $h(x)$ jest zdefiniowana jako liczba podziałów, które obserwacja przechodzi przed osiągnięciem zewnętrznego węzła drzewa. Obserwacje są sortowane według odpowiadających im długości ścieżek, a obiekty o najkrótszych ścieżkach mają większe prawdopodobieństwo bycia obserwacją odstającą. Ostateczna miara "stopnia odstawania" uzyskana na podstawie $h(x)$ została opisana w oryginalnej pracy Liu, Ting i Zhou (2008) w następujący sposób. Biorąc pod uwagę zbiór N obserwacji, miara "stopnia odstawania" dla obserwacji x jest zdefiniowana jako

$$s(x, n) = 2^{-\overline{h}(x)/c(n)}, \tag{4.31}$$

gdzie $\overline{h}(x)$ jest średnią długością ścieżki otrzymaną ze zbioru drzew modelu, a $c(n)$ jest średnią długością ścieżki nieudanych poszukiwań zdefiniowanych w Liu, Ting i Zhou (2008). Taki wynik miary "stopnia odstawania" jest monotoniczny względem

długości ścieżki. Te dwie miary mogą zajmować zakresy wartości odpowiednio $0 < s(x, N) \leq 1$ oraz $0 < h(x) < N - 1$. W tych warunkach autorzy definiują reguły wykrywania w następujący sposób. Jeżeli obserwacja x ma wartość $s(x, N) \simeq 1$, to x jest identyfikowana jako obserwacja odstająca. Obserwacje z $s(x, N) \leq 0.5$ są identyfikowane jako obiekty typowe. Wreszcie, jeżeli cała próbka charakteryzuje się $s \simeq 0,5$, to w danych nie ma wyraźnych obserwacji odstających.

4.3.2 Wizualizacja wielowymiarowa za pomocą algorytmu tSNE

Mechanizm wykrywania obserwacji odstających za pomocą algorytmu Isolation Forest, pomimo swojej wysokiej skuteczności, daje nam ograniczone informacje o kontekście, w którym występuje dana obserwacja odstająca. Nie wiemy, jak ten obiekt ma się w stosunku do innych obiektów znajdujących się być może w różnych próbkach i jakie cechy fizyczne zadecydowały o zaklasyfikowaniu go jako obiektu odstającego. W rozprawie zastosowano algorytm Isolation Forest w celu identyfikacji różnych źródeł zanieczyszczenia katalogu AGN-ów, i w niektórych przypadkach kontekst występowania obserwacji jest kluczowy. Z tego powodu dołączono dodatkowy krok w postaci wysokowymiarowej wizualizacji za pomocą algorytmu tSNE (ang. *t-distributed stochastic neighbor embedding*, Hinton i Roweis, 2002; Maaten i Hinton, 2008).

Algorytm tSNE jest nieliniową techniką redukcji wymiarowości, która jest zwykle stosowana do wizualizacji danych wielowymiarowych. Algorytm tSNE składa się z dwóch głównych kroków. Po pierwsze, wspólny rozkład prawdopodobieństwa jest konstruowany z podobieństw pomiędzy obiektami w wielowymiarowej przestrzeni cech. Podobieństwo obiektów w podstawowej wersji algorytmu tSNE jest oparte na odległości euklidesowej pomiędzy obiektami w przestrzeni cech. Następnie algorytm tSNE próbuje nauczyć się niskowymiarowej reprezentacji, które może zachować podobieństwa obecne w reprezentacji wysokowymiarowej. Algorytm tSNE minimalizuje dywergencję Kullbacka-Leiblera (ang. *Kullback-Leibler divergence*) metodą gradientu prostego (ang. *gradient descet*) pomiędzy wysokowymiarową reprezentacją danych w oryginalnej przestrzeni cech a niskowymiarową reprezentacją utworzonym przez tSNE. Taka minimalna wartość odpowiada pozycji obiektów w niskowymiarowej przestrzeni.

5

Budowa modelu klasyfikującego i wynikowe katalogi aktywnych jąder galaktyk

5.1 Budowa modelu klasyfikującego opartego na technikach uczenia maszynowego

W tym rozdziale opisano ogólną budowę układu potokowego przetwarzania danych opartego na technikach uczenia maszynowego, który został zastosowany w rozprawie. Schemat sekwencji głównych kroków jest przedstawiony na rysunku 5.1 i przebiega w sposób opisany poniżej.

W pierwszej kolejności przeprowadzono procedury przygotowania danych. Należała do nich selekcja cech używanych przez modele uczenia maszynowego. Cechy zostały wybrane za pomocą porównania wartości statystyki Kołomogorowa-Smirnova dla różnych cech. Ta metoda jest opisana w rozdziale 5.2. Drugą częścią przygotowania danych było ograniczenie próbki generalizacjnej za pomocą algorytmu MCD. Ta metoda została opisana w rozdziale 3.3.2. W ten sposób trening modelu został przeprowadzony na oznaczonych danych reprezentowanych przez wybrany zestaw cech. Generalizacja przeprowadzona na nieoznakowanych danych z ograniczeniem MCD została wykonana przy użyciu tego samego zestawu cech.

Podczas procedury treningowej wytrenowano i zbadano szereg rodzajów nadzorowanych algorytmów uczenia maszynowego opisanych w rozdziale 4. Były to regresja logistyczna, maszyna wektorów nośnych, las losowy (*random forest*), algorytm wyjątkowo losowych drzew (*extremely randomized trees*), XGBoost oraz stworzone na ich podstawie dwa schematy głosujące. Każdy z podstawowych modeli był testowany w dwóch głównych wersjach: w wersji niezmodyfikowanej oraz w wersji z zastosowaniem wag klasowych. *Ważenie klasowe* (lub zrównoważenie klasowe) polega na przypisaniu wszystkim obiektom klasy mniejszej wagi będącej stosunkiem rozmiaru klasy większej do klasy mniejszej. Dzięki temu wszystkie obiekty mniejszej klasy są traktowane jako bardziej istotne, co pozwala zredukować efekt nierównych rozmiarów klas. Obie te wersje modeli (zrównoważone i niezrównoważone klasowo) zostały dodatkowo przetestowane z zastosowaniem różnych strategii logiki rozmytej (tj. osobnego ważenia poszczególnych obserwacji w próbce treningowej). Podstawy zastosowania logiki rozmytej zostały opisane w rozdziale 5.3.

Podczas treningu, dla większości modeli zastosowano optymalizację hiperparametrów. Modele regresji logistycznej, SVM i XGBoost były optymalizowane za

pomocą losowego przeszukiwania siatki parametrów (ang. *randomized grid search*). W takim przeszukiwaniu zostało stworzonych 1000 różnych losowo wybranych kombinacji wartości hiperparametrów. Ten rodzaj przeszukiwania siatki hiperparametrów pozwala na znalezienie najbardziej odpowiedniego wariantu modelu do konkretnego zadania klasyfikacyjnego. Metody oparte na zespołach drzew decyzyjnych, tj. algorytmy Random Forest i Extremely Randomized Trees nie były optymalizowane poprzez przeszukiwanie siatki hiperparametrów. Pozostawiono je w domyślnych wersjach obecnych w bibliotece Scikit-learn (Pedregosa i in., 2011) ze względu na bardzo małą wrażliwość tych metod na dostrajanie hiperparametrów (Probst, Boulesteix i Bischl, 2019). Optymalna kombinacja hiperparametrów była poszukiwana poprzez maksymalizację metryki F1. Za każdym razem, gdy wybierano konkretną kombinację hiperparametrów, wartość metryki F1 była szacowana za pomocą 100-krotnego losowania 5-krotnych walidacji krzyżowych. Tak dokładna estymacja wartości metryk została użyta aby zminimalizować ryzyko nadmiernego dopasowania na małym zbiorze danych treningowych.

Tę samą technikę podziału zastosowano na ostatecznej, najlepszej kombinacji hiperparametrów, aby oszacować wartości pozostałych metryk oraz towarzyszące im niepewności. Wartości lub zakresy hiperparametrów, które zostały użyte do stworzenia siatki hiperparametrów, są przedstawione w tabeli 5.1. Opis większości tych parametrów można znaleźć w rozdziale 4.1. Wyjątkiem są parametry algorytmu XGBoost, z których część jest związana z konkretną optymalizacją numeryczną. Pełny opis hiperparametrów XGBoost można znaleźć w dokumentacji danej implementacji algorytmu [1]. W przypadku wszystkich algorytmów opartych na drzewach decyzyjnych, do budowy zespołu użyto zestawu 500 drzew. W tym przypadku, jeżeli liczba drzew okazałaby się zbyt duża, nie spowodowałoby to nadmiernego dopasowania modelu. Zamiast tego zbyteczne dodatkowe drzewa przestałyby poprawiać wydajność modelu. Kody napisane w języku Python 3 użyte do trenowania modeli są dostępne na portalu GitHub [2].

Następnie wytrenowany model z najlepszą kombinacją hiperparametrów był użyty do predykcji na zbiorach danych oznaczonych i próbce generalizacyjnej. Ze względu na niewielką ilość danych treningowych, nie utworzono oddzielnego zbioru testowego, który zostałby wykorzystany tylko do predykcji modelu końcowego. Zamiast tego przeprowadzono przybliżoną wersję testowania przewidywań modelu. W tym celu model z ustalonymi wybranymi kombinacjami hiperparametrów był poddawany dodatkowej 5-krotnej walidacji krzyżowej. Łączne przewidywania z 5-krotnej walidacji zostały wykorzystane do oszacowania przewidywań modelu na oznaczonych danych. Należy pamiętać, analizując wizualizację przewidywań modelu na danych oznaczonych, że tego typu procedura nieco zaniża rzeczywistą wydajność klasyfikatora. Po zakończeniu etapu predykcji na danych oznaczonych, model z ustalonymi wartościami hiperparametrów był ponownie trenowany na całości danych treningowych a następnie był użyty do klasyfikacji danych w próbce generalizacyjnej w celu utworzenia katalogu kandydatów na AGN-y. Oprócz tych metod, z najlepszych modeli utworzono dwa klasyfikatory głosujące (patrz rozdział 5.4). System twardego głosowania (ang. *hard voter*), który wykorzystał głosowanie większościowe zestawu modeli, nie musiał być trenowany. Klasyfikator stosowy (ang. *stacked classifier*), który wykorzystywał estymacje prawdopodobieństwa najlepszych modeli jako zestaw cech, był trenowany w taki sam sposób, jak inne modele.

[1] https://xgboost.readthedocs.io/en/stable/parameter.html
[2] https://github.com/ArtemPoliszczuk/NEPWide_AGN

hiperparametr	zakres
regresja logistyczna	
C	loguniform(1, 10^{-3})
Penalty	[L_1, L_2]
SVM	
C	loguniform(1, 10^{-3})
γ	loguniform(1, 10^{-3})
XGBoost	
learning rate	[0.01, 0.02, 0.03, 0.05, 0.08, 0.1]
γ	[0.5, 1, 1.5, 2, 5]
Min child weight	[1, 5, 10]
Training subsample ratio	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Max tree depth	[2, 3, 4, 5, 6, 10]
L_2 regularization weight	[1, 2, 4]
L_1 regularization weight	[0, 1, 2]

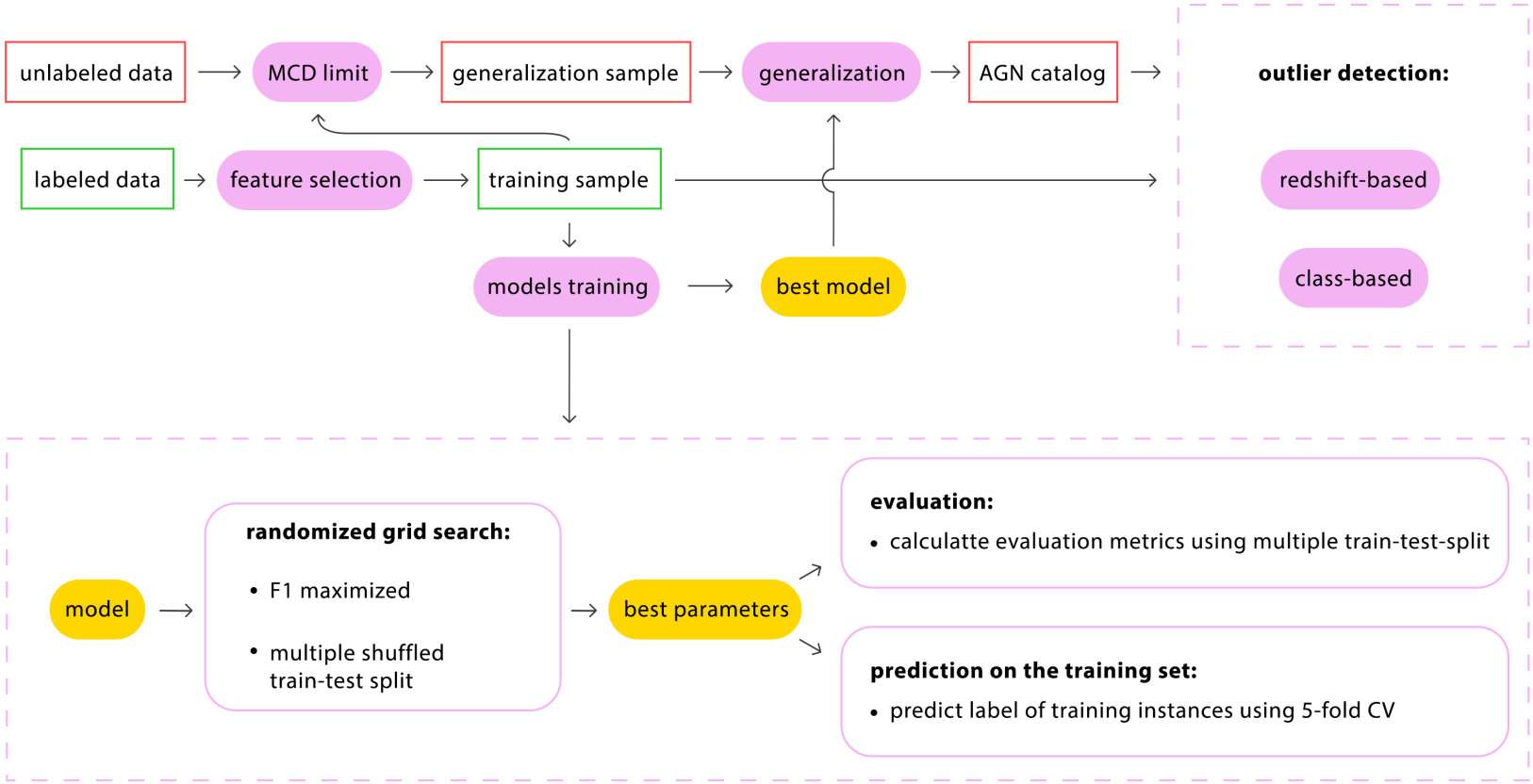
TABLICA 5.1: Siatka wartości hiperparametrów wykorzystywanych do optymalizacji modelu podczas treningu. Niektóre parametry regresji logistycznej i SVM były próbkowane z rozkładu logarytmicznego-jednostajnego (ang. *log-uniform*) o ustalonym zakresie.

Po wybraniu najlepszego modelu, otrzymany katalog AGN-ów został przeanalizowany za pomocą metod wykrywania obserwacji odstających opisanych w rozdziale 5.5. Jedna metoda, wykorzystująca algorytm Isolation Forest, została zastosowana do selekcji nieprawidłowo oszacowanych fotometrycznych przesunięć ku czerwieni. Druga metoda, składająca się z połączenia algorytmu Isolation Forest z wielowymiarową wizualizacją za pomocą algorytmu tSNE, została wykorzystana do znalezienia zanieczyszczeń w katalogu kandydatów na AGN-y.

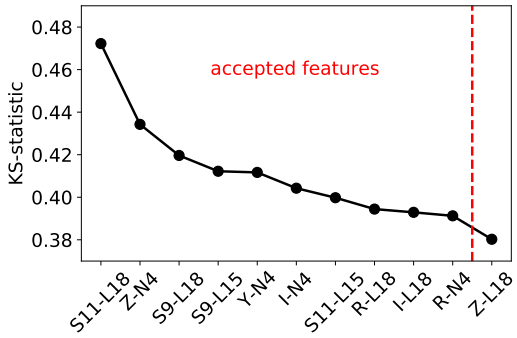
5.2 Wybór cech

Oryginalna reprezentacja danych często nie jest optymalna dla zastosowań uczenia maszynowego. Istnieją dwa główne powody. Po pierwsze, bardzo duża liczba cech może skutkować dużą rozproszonością danych występujących w przestrzeni wielowymiarowej - ta cecha znana jest jako tzw. przekleństwo wielowymiarowości, ang. *curse of dimensionality* (Bishop, 2006). Pojawienie się dużych odległości między obserwacjami w przestrzeni parametrów często skutkuje zmniejszoną skutecznością modelu. Po drugie, niereprezentatywny zestaw cech może wprowadzić znaczny szum statystyczny do danych, co prowadzi do nadmiernego dopasowania (*overfitting*). Aby przezwyciężyć ten problem, stosuje się różne metody inżynierii cech (ang. *feature engineering*). Częste podejście do tego problemu składa się z dwóch kroków. Najpierw odbywa się *tworzenie cech*, gdzie sztuczne cechy są tworzone na podstawie kombinacji cech oryginalnych, np. stosunku cech, ich mnożenia lub różnicy. Następnie odbywa się *wybór cech*, gdzie wybierany jest najlepszy podzbiór wszystkich dostępnych cech.

Metody selekcji cech można podzielić na dwa główne podejścia. Pierwszym z nich jest ogólne, nieuwzględniające specyfiki problemu podejście, koncentrujące się na redukcji wymiarowości przestrzeni cech. Przykładem tego podejścia jest wiele



RYSUNEK 5.1: Schemat układu potokowego przetwarzania danych opartego na metodach uczenia maszynowego, opisany w rozprawie. Górna część schematu pokazuje ogólny zarys procedury, dolna część schematu, pokazana w fioletowym prostokącie, odnosi się bezpośrednio do procesu trenowania modeli. Wykres jest zmodyfikowaną wersją wykresu przedstawionego w pracy Poliszczuk i in. (2021).



RYSUNEK 5.2: Wyniki selekcji cech w głównym procesie klasyfikacji metodą statystyki Kołomogorowa-Smirnowa. Pokazany jest tylko podzbiór cech z najwyższym wynikiem statystyki KS. Wykres pochodzi z pracy Poliszczuk i in. (2021).

popularnych metod, takich jak PCA, które mają tendencję do redukcji szumu w reprezentacji danych bez odniesienia do konkretnego problemu predykcji. Drugie podejście zastosowane w tej pracy ma na celu znalezienie cech najbardziej odpowiednich dla konkretnego problemu predykcji. W celu znalezienia zestawu cech, który pozwoli na skuteczną selekcję AGN-ów w danych, zastosowano statystykę Kołomogorowa-Smirnowa (KS). Statystyka KS, w tym przypadku, jest definiowana jako największa odległość pomiędzy empiryczną dystrybuantą rozkładów (ang.*empirical cumulative distribution function*) próbek treningowych AGN-ów i galaktyk. Statystykę KS można traktować jako miarę różnicy między tymi dwoma rozkładami. Tym samym większa wartość statystyki KS obliczona dla danej cechy odpowiada lepszej przydatności tej cechy do rozdzielenia AGN-ów i galaktyk na danym zbiorze obserwacji. Statystyka KS została obliczona dla pasm optycznych SUBARU/HSC (g, r, i, z, Y), pasm NIR AKARI/IRC ($N2$, $N3$, $N4$) oraz wszystkich możliwych kolorów stworzonych z tych ośmiu pasm. Rysunek 5.2 pokazuje wartości statystyki KS dla podzbioru najlepszych cech. W celu uniknięcia wzrostu wymiaru przestrzeni cech, liczba cech końcowych została ograniczona do liczby początkowych pasm optycznych i NIR. W rezultacie otrzymano zestaw cech składający się z ośmiu kolorów, wykorzystujący wszystkie dostępne informacje o pasmach. Ponadto można tutaj również zaobserwować szczególnie duży wkład informacyjny kolorów NIR do selekcji AGN-ów.

Statystyka KS ma kilka zalet i wad jako metoda selekcji cech. Z jednej strony, jej prostota jest jej główną zaletą. Nie wymaga dużej ilości danych i nie jest podatna na nadmierne dopasowanie modelu, w przeciwieństwie do bardziej złożonych, zależnych od modelu metod (tzw. *wrapper methods*, Jović, Brkić i Bogunović, 2015). Ponadto ta metoda jest w stanie uchwycić różnice zarówno w położeniu, jak i kształcie dwóch rozkładów. Z drugiej strony, takie podejście do selekcji cech koncentruje się na pojedynczych cechach i nie jest wrażliwe na wzajemne powiązania między nimi.

5.3 Logika rozmyta w uczeniu nadzorowanym

W tej pracy zastosowano dwa rodzaje logiki rozmytej (lub osobnego ważenia obserwacji w próbce treningowej): oparte na odległości od środka klasy i oparte na błędzie pomiarowym. W obu przypadkach waga obiektu s_i dla i-tej obserwacji została znormalizowana do przedziału [0, 1], gdzie większa wartość odpowiada większej istotności obserwacji. Wagi s_i były liczone na podstawie równania

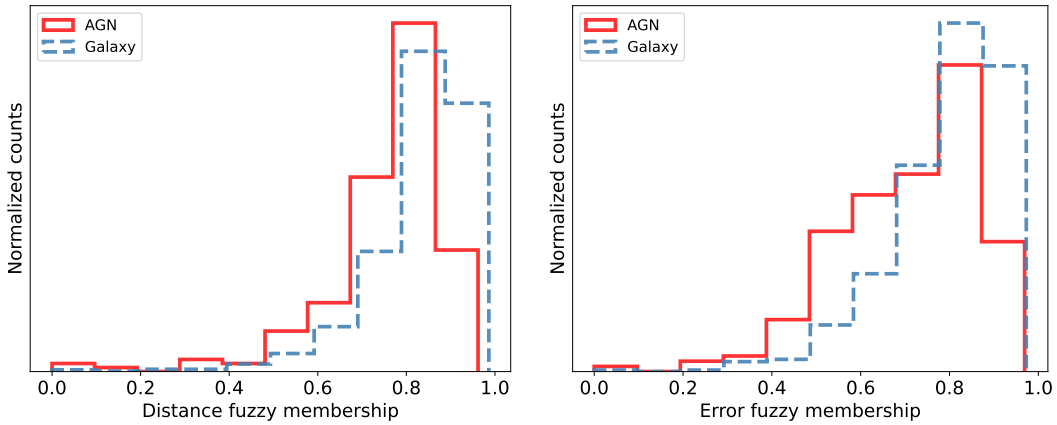
$$s_i = 1 - \frac{u_i}{u_{\max} + \delta}, \tag{5.1}$$

gdzie u_i jest wielkością charakterystyczną dla określonego typu wagi obiektu obliczoną dla i-tej obserwacji w zbiorze treningowym, $u_{\max}$ jest maksymalną wartością u_i w próbce, zaś δ jest małą wartością używaną w celu uniknięcia dzielenia przez zero. Ponieważ na ważność poszczególnych obserwacji w próbce treningowej wpływa tylko względna różnica między wagami, obecność parametru δ nie ma żadnego wpływu na proces treningowy i służy jedynie zapewnieniu bezpieczeństwa numerycznego. W naszym wypadku $\delta = 10^{-4}$.

Idea zastosowania logiki rozmytej opartej na odległości od środka klasy w przestrzeni cech, jak również definicja s_i przedstawiona w równaniu 5.1 została opisana w Lin i Wang (2002), gdzie była stosowana w algorytmie SVM. Podobne ważenie obserwacji może być stosowane również w innych rodzajach algorytmów uczenia (patrz rozdział 4.1). W przypadku wag opartych na odległości, u_i jest odległością euklidesową od środka klasy, zdefiniowanego w przestrzeni cech. Celem zastosowania takiej wagi jest zminimalizowanie wpływu obserwacji odstających na proces trenowania modelu. Aby prawidłowo skonstruować takie wagi, zostały one obliczone oddzielnie dla klas AGN-ów i galaktyk. Takie podejście jest podyktowane różnym rozkładem klas w przestrzeni cech oraz nierównym rozmiarem klas. Różnica w rozmiarze próbek powodowałaby systematyczne niedoszacowanie znaczenia mniejszej klasy, gdyby wagi były liczone na całym zbiorze treningowym. Histogramy rozkładu wag na podstawie odległości dla obu klas są pokazane na rysunku 5.3a. Tutaj widzimy, że klasę AGN-ów charakteryzuje duża liczba obserwacji odstających. Dlatego też zastosowanie logiki rozmytej opartej na odległości powinno zapewnić bardziej konserwatywną klasyfikację, charakteryzującą się wyższą czystością i niższą kompletnością katalogu wynikowego AGN-ów.

Metoda ważenia opartego na błędach pomiarowych opiera się na tych samych zasadach, ale koncentruje się na wpływie niepewności pomiaru na klasyfikację. To podejście zostało po raz pierwszy omówione we wstępnej pracy Poliszczuk i in. (2019) i dalej rozwinięte wraz z wagami opartymi na odległości w Poliszczuk i in. (2021). W tym podejściu u_i jest zdefiniowane jako suma wartości bezwzględnych niepewności pomiarowych obliczonych dla pasm optycznych i NIR. W tym przypadku dopuszczono specyficzne statystyczne obciążenie. Pomiary w pasmach optycznych charakteryzują się mniejszą niepewnością pomiaru ze względu na wyższą dokładność instrumentu jak i specyfikę oprogramowania do przetwarzania danych SUBARU/HSC, które ma tendencję do niedoszacowania błędu pomiarowego. Z tego powodu dokładność pomiaru w pasmach AKARI ma dominujący wkład do ostatecznej wartości s_i. Ze względów praktycznych nie zastosowano przeskalowania i ujednolicenia błędów pomiarowych. Pasma NIR mają główny wkład informacyjny do selekcji AGN i dlatego powinny być traktowane jako ważniejsze. Histogramy rozkładu wag logiki rozmytej opartej na błędach pomiarowych dla obu klas są pokazane na rysunku 5.3a. Ponownie widzimy tutaj większy wpływ wag na klasę AGN-ów.

Dalsze porównanie dwóch strategii ważenia jest przedstawione na rysunku 5.4. W szczególności na rysunku 5.4a widać silny wpływ wag opartych na odległości od środka klasy. Można tu zaobserwować spadek znaczenia SFG o dużych przesunięciach ku czerwieni, które charakteryzują się czerwonym kolorem $N2$–$N4$, jak również mniejszy wpływ XAGN i innych AGN, charakteryzujących się niebieskimi kolorami. Z drugiej strony, wagi oparte na błędach pomiarowych wykazują słabą, rozproszoną tendencję do zmniejszania znaczenia czerwonych obiektów $N2$–$N4$.



(A) (B)

RYSUNEK 5.3: Znormalizowane histogramy różnych typów ważenia opartego na logice rozmytej. *Panel a:* Wagi oparte na odległości od środka klasy. *Panel b:* Wagi oparte na niepewności pomiarowej. Wykresy pochodzą z pracy Poliszczuk i in. (2021).

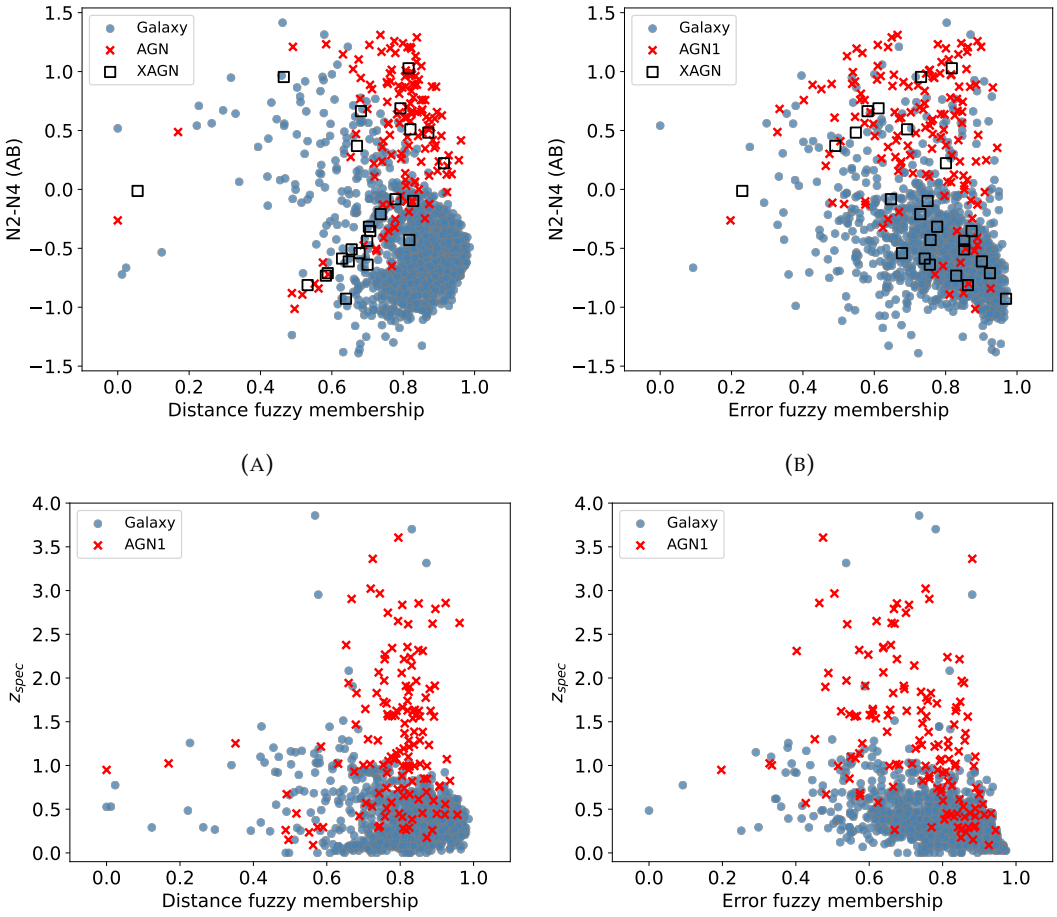
Zależność między wagami opartymi na odległości a rozkładem przesunięcia ku czerwieni w próbce treningowej jest pokazana na rysunku 5.4c. W tym przypadku nie można zaobserwować żadnej istotnej zależności. Analogiczna zależność dla wag logiki rozmytej opartych na błędzie pomiarowym jest przedstawiona na rysunku 5.4d. Tutaj widać słabą korelację, która jest spowodowana tym, że bardziej odległe obiekty na ogół cechuje mniejsza wielkość gwiazdowa.

5.4 Ocena jakości klasyfikacji poszczególnych modeli

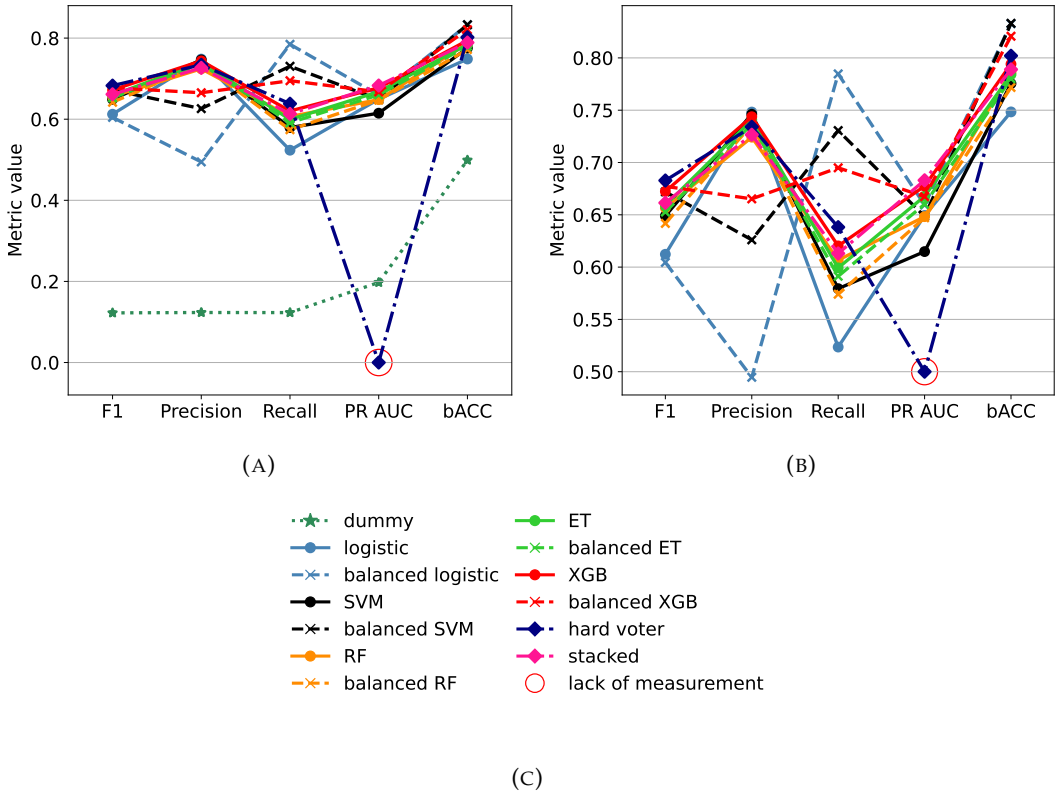
5.4.1 Wpływ zastosowania wag klasowych na jakość predykcji

Wynik treningu różnych modeli uczenia nadzorowanego opisanych w rozdziale 4.1 jest oceniany poprzez porównanie kilku metryk: metryki F1, czystości katalogu AGN-ów (*precision*), kompletności katalogu AGN-ów (*recall*), obszaru pod krzywą precision-recall (PR AUC) i zrównoważonej dokładności (bACC). Wszystkie te metryki zostały szczegółowo opisane w rozdziale 4.2.

Ogólna analiza jakości modelu jest przedstawiona na rysunku 5.5. Tutaj widzimy wartości metryk dla różnych modeli niezbalansowanych i zbalansowanych klasowo. Ponieważ zastosowanie ważenia klasowego jest uważane za bardziej fundamentalną zmianę w modelu niż bardziej subtelne ważenie poszczególnych obserwacji (logika rozmyta), na tym pierwszym etapie analizy jakości modeli, modele ważone metodami logiki rozmytej nie zostały użyte. Na tym etapie analizy można wyróżnić trzy główne cele: znalezienie podzbioru najlepszych modeli do stworzenia schematów głosujących, przeanalizowanie wpływu ważenia klasowego na jakość predykcji modelu i znalezienie, jeśli to możliwe, najlepszego modelu. Na rysunku 5.5a przedstawiono zestaw wszystkich klasyfikatorów wraz z tzw. klasyfikatorem naiwnym (ang. *dummy classifier*). Klasyfikator naiwny przypisuje etykiety klas w sposób losowy, gdzie wielkości przewidywanych próbek klas odpowiadają stosunkom klas w danych treningowych. Wydajność tego podstawowego modelu określa dolną granicę jakości predykcji w naszej analizie, tzn. jeżeli model był w stanie nauczyć się zadania klasyfikacji w procesie szkolenia, to powinien wykazywać wyższe wartości metryk



RYSUNEK 5.4: Wpływ logiki rozmytej na właściwości próbki treningowej. *Panel A:* Zależność między wagami opartymi na odległości od środka klasy a rozkładem koloru $N2–N4$. *Panel B:* Zależność między wagami opartymi na niepewności pomiarowej a rozkładem koloru $N2–N4$. *Panel C:* Zależność między wagami opartymi na odległości od środka klasy a rozkładem przesunięcia ku czerwieni. *Panel D:* Zależność między wagami opartymi na niepewności pomiarowej a rozkładem przesunięcia ku czerwieni.



(A) (B)

(C)

RYSUNEK 5.5: Ocena jakości predykcji różnych modeli klasyfikacyjnych. Przedstawione są tylko modele bez zastosowania logiki rozmytej oraz schematy głosujące. *Panel A:* Metryki oceny dla różnych modeli w porównaniu z klasyfikatorem naiwnym (*dummy*). *Panel B:* Metryki oceny dla różnych modeli. Klasyfikator naiwny nie jest uwzględniony. *Panel C:* Legenda. Pokazane metryki zostały opisane w rozdziale 4.2. Wykres pochodzi z pracy Poliszczuk i in. (2021).

niż klasyfikator naiwny. Na podstawie porównania z klasyfikatorem naiwnym, rysunek 5.5a pokazuje, że wszystkie przedstawione modele mogą skutecznie nauczyć się zadania klasyfikacji. Brak wyniku PR AUC dla systemu twardego głosowania (*hard voter*) wynika z faktu, że konstrukcja krzywej precision-recall wymaga użycia prawdopodobieństw klasyfikacji, które nie występują w przypadku głosowania większościowego tworzonego przez system twardego głosowania.

Rysunek 5.5b pokazuje te same wyniki w powiększeniu bez klasyfikatora naiwnego. Tutaj widzimy kilka interesujących tendencji. Po pierwsze, istnieje podział na dwa rodzaje wpływu, jaki ważenie klasowe może mieć na wydajność klasyfikatora. Ten podział, widoczny jest przede wszystkim w przypadku czystości (precision) i kompletności (recall) katalogu AGN-ów precyzji do przywołania. Jedna klasa zachowania reprezentowana przez regresję logistyczną (*logistic* na rysunku), SVM i XGBoost (*XGB* na rysunku). Zatem metody liniowe, nieliniowe oraz oparte na metodzie boosting wykazują znaczny wzrost kompletności przy jednoczesnym spadku czystości pozytywnej klasy, gdy zastosowane jest ważenie klasowe. Ten wpływ wag stosowanych w celu zniwelowania wpływu różnic wielkości klasy można interpretować intuicyjnie. W przypadku braku takiego ważenia, separacja pomiędzy klasami jest przesunięta w kierunku mniejszej klasy (tj. klasy AGN-ów lub klasy pozytywnej). W ten sposób obiektom znajdującym się w bardzo bliskim sąsiedztwie próbki treningowej mniejszej klasy przypisywana jest etykieta mniejszej klasy. Skutkuje to

predykcją, która przypisuje etykietę klasy pozytywnej tylko dla bardzo typowych obserwacji, które leżą blisko centrum klasy, dając katalog o wysokiej czystości i niskiej kompletności. Zastosowanie ważenia klasowego przesuwa granice klasyfikacji w przestrzeni cech z dala od regionu mniejszej klasy. Powoduje to wzrost kompletności oraz większe zanieczyszczenie katalogu klasy pozytywnej (tj. niższą wartość metryki *precision*). W przypadku pozostałych modeli, tj. algorytmu wyjątkowo losowych drzew (oznaczonego jako *ET* na rysunkach) i algorytmu lasu losowego (oznaczonego jako *RF* na rysunkach), nie widać znaczącego wpływu wagi klasy na predykcje modelu. Wynik ten nie może być wyjaśniony właściwościami modeli zespołu drzew i może wynikać ze specyfiki danego problemu klasyfikacyjnego, bowiem zgodnie z dotychczasowymi badaniami, ważenie klasowe powinno mieć istotny wpływ na predykcje modeli opartych na zespołach drzew decyzyjnych (patrz np. Chen, 2004).

Z podzbioru najlepszych modeli zostały utworzone oba klasyfikatory głosujące, tzn. system twardego głosowania (oznaczony na rysunkach jako *hard voter*) oraz klasyfikator stosowy (oznaczony na rysunkach jako *stacked*). Grupy modeli bazowych dla tego podzbioru zostały wybrane poprzez analizę modeli niezrównoważonych i zrównoważonych klasowo bez zastosowania logiki rozmytej. Innymi słowy, jeżeli pewien model z zastosowaniem wag klasowych został wybrany jako model bazowy klasyfikatora głosującego, to jako modele bazowe były brane wszystkie trzy wersje tego modelu oparte na logice rozmytej: bez logiki rozmytej, z ważeniem opartym na odległości od środka klasy i z ważeniem opartym na niepewnościach pomiarowych. Zrobiono tak, ponieważ wagi klasowe są uważane za bardziej podstawową metodę ważenia, podczas gdy wagi oparte na logice rozmytej są bardziej subtelne i odgrywają rolę pomniejszego dostrojenia. Wybór modeli dla klasyfikatorów głosujących został dokonany na podstawie porównania wartości metrycznych przedstawionych na rysunku 5.5. Aby odrzucić niektóre z klasyfikatorów, usunięto modele o znacznie niższych wartościach którejkolwiek z metryk. W ten sposób usunięto regresję logistyczną bez ważenia klasowego ze względu na bardzo niską czystość pozytywnej klasy (precision), klasowo zrównoważoną regresję logistyczną ze względu na bardzo niską wartość metryki recall oraz SVM bez ważenia klasowego ze względu na niską wartość metryki PR AUC. Klasowo zrównoważone modele SVM i XGBoost pozostawiono w ostatecznym zestawie pomimo nieco niższych wartości metryki precision, aby uzyskać różnorodność w działaniu poszczególnych modeli. System twardego głosowania i klasyfikator stosowy wykazują podobne tendencje w wartościach metryk do modeli bez wag klasowych. To odejście od zachowania obecnego w modelach z zastosowaniem wag klasowych wynika z faktu, że w niektórych przypadkach zastosowanie logiki rozmytej zbliżało modele z zastosowaniem wag klasowych do modeli bez tych wag (zjawisko to zostanie omówione w dalszej części tego rozdziału).

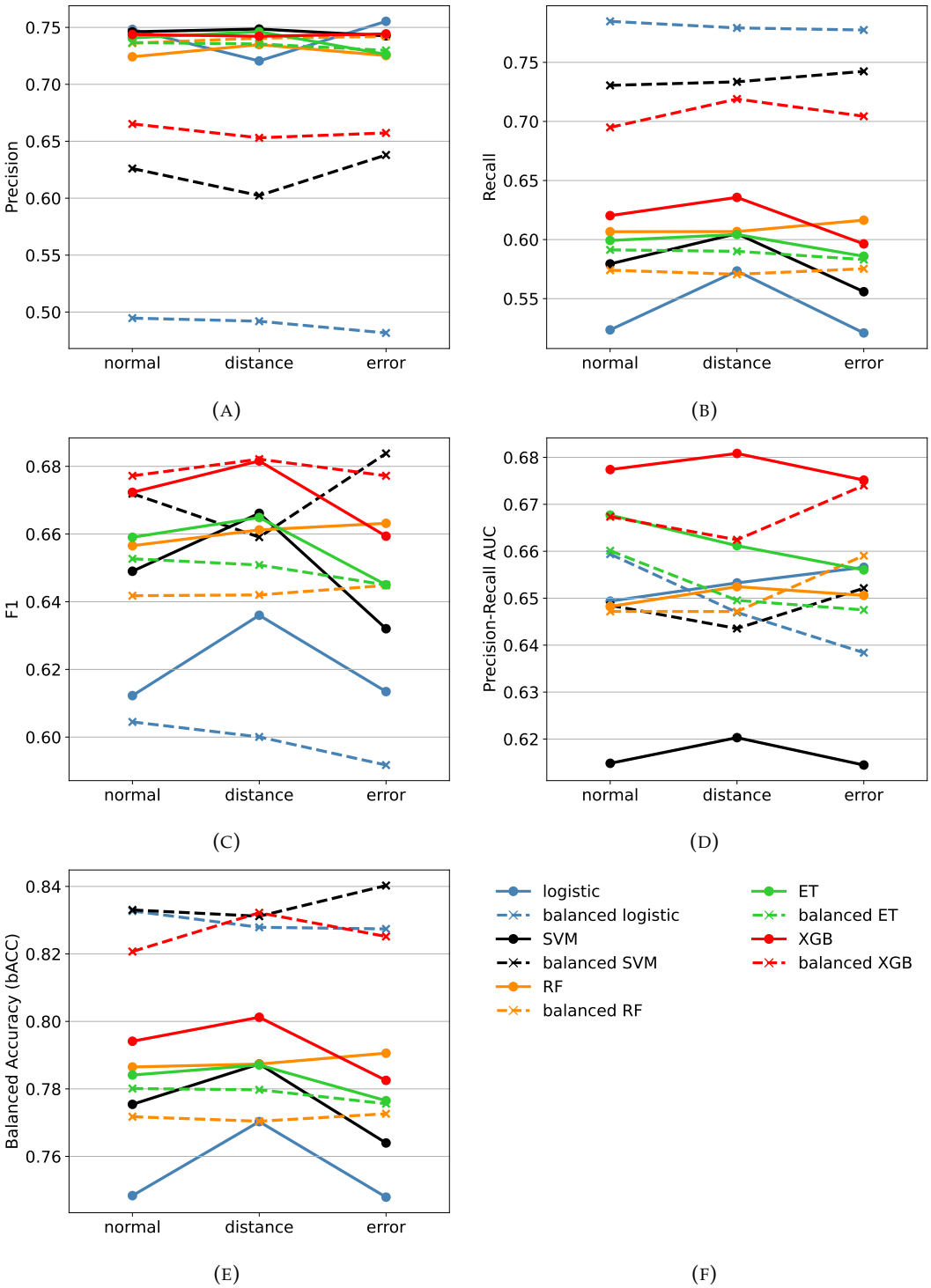
Jako najlepszy model końcowy wybrano system twardego głosowania. Ta decyzja ma kilka powodów. Na tym etapie porównanie wydajności modelu musi wykraczać poza analizę wyniku F1, który był główną metryką podczas procesu treningowego. Niewielka ilość danych szkoleniowych utrudnia precyzyjne oszacowanie wartości metryki, a wnioski na podstawie uzyskanych wyników należy wyciągać z ostrożnością. Ta niepewność, jak również znane zanieczyszczenia pochodzące z SFG o dużym przesunięciu ku czerwieni, doprowadziły do preferencji modeli o większej czystości i mniejszej kompletności katalogu AGN-ów podczas analizy jakości predykcji. Najlepsze wyniki wykazały dwa modele: algorytm XGBoost bez wag klasowych i z zastosowaniem rozmytej logiki opartej na odległości od środka klasy oraz system twardego głosowania. Różnice między wartościami większości metryk mieściły się w granicach niepewności (dokładne wartości metryk i odpowiadające im niepewności dla wszystkich modeli można znaleźć w Dodatku B, w Tabelach B.1 oraz B.2). Z tych

dwóch modeli wybrano system twardego głosowania, ze względu na ograniczoną możliwość kontroli wydajności klasyfikatora spowodowaną małą ilością danych treningowych. System twardego głosowania, ze względu na swój charakter zespołowy, pozwala nam jeszcze bardziej zmniejszyć wariancję modelu i uniknąć ryzyka nadmiernego dopasowania modelu. Jednak prostota i skuteczność tego modelu zostały uzyskane kosztem braku możliwości estymacji prawdopodobieństwa klasyfikacji. W rezultacie ostateczny model charakteryzował się wartościami 0,73 metryki precision (czystość katalogu AGN-ów) i 0,64 metryki recall (kompletność katalogu AGN-ów). Generalizacja przeprowadzona na nieoznakowanych danych dała nam katalog 465 kandydatów na AGN-y, co stanowi 1,4% całej próbki generalizacyjnej.

5.4.2 Wpływ logiki rozmytej na predykcje modeli klasyfikacyjnych

Zanim przejdziemy do analizy właściwości ostatecznego klasyfikatora i uzyskanego katalogu AGN-ów, zbadajmy wpływ różnych strategii logiki rozmytej na działanie modeli. Rysunek 5.6 przedstawia wartości metryk dla różnych strategii logiki rozmytej. Rysunki 5.6a i 5.6b pokazują wartości metryk precision i recall. Ponieważ obie te metryki są podstawą dla bardziej złożonych metryk F1, PR AUC i częściowo bACC, warto od nich zacząć analizę. Zastosowanie logiki rozmytej nie ma dużego wpływu na wartość metryki precision - podobne wartości są uzuskiwane w przypadku wszystkich wersji ważenia opartego na logice rozmytej. Dwa wyjątki stanowią regresja logistyczna i SVM zrównoważony klasowo, które wykazują znaczne różnice między różnymi typami wag i przeciwny wpływ zastosowania logiki rozmytej dla modeli z i bez zastosowania ważenia klasowego. W znacznie mniejszym stopniu ta tendencja jest widoczna również w przypadku modeli XGBoost. W ten sposób można zaobserwować podział modeli na dwie różne kategorie zachowań, które pokrywają się z ogólnymi różnicami między modelami widocznymi na rysunku 5.5, które zostały omówione już wcześniej. Ten podział jest obecny nie tylko w przypadku ważenia klasowego, ale jest również zachowany przy zastosowaniu logiki rozmytej. Porównanie wartości metryki recall pokazanej na rysunku 5.6b daje odmienny obraz. Widać tutaj, że logika rozmyta wpływa na algorytmy niezrównoważone klasowo, pozostawiając modele z zastosowaniem klasowych wag stosunkowo nienaruszone. Co więcej, zastosowanie logiki rozmytej opartej na odległości od środka klasy daje lepsze wyniki niż zastosowanie logiki rozmytej opartej na niepewnościach pomiarowych lub brak zastosowania logiki rozmytej. Tendencje widoczne w wartościach metryk precision i recall przekładają się na zachowanie obserwowane w wynikach F1 (rysunek 5.6c) i bACC (rysunek 5.6e). Wynik PR AUC, przedstawiony na rysunku 5.6d, często wykazuje odmienne tendencje, co sugeruje, że możliwym jest uzyskanie modelu o zupełnie innych własnościach, gdyby podczas procedury optymalizacji hiperparametrów wykorzystywano by wyniki PR AUC zamiast F1.

Ogólną tendencją, którą można zaobserwować podczas analizy wartości metryk, jest poprawa jakości predykcji modelu w przypadku logiki rozmytej opartej na odległości od środka klasy oraz stosunkowo subtelna różnica pomiędzy modelami z ważeniem opartym na niepewnościach pomiarowych a modelami bez zastosowania logiki rozmytej. Widoczny jest znaczący wzrost wartości metryki recall dla modeli bez równoważenia klasowego w których zastosowano ważenie oparte na odległości od środka klasy. Jednocześnie nie da się zaobserwować znaczącej zmiany w wartości metryki precision w danych modelach. Podobnego zachowania nie widać w modelach zrównoważonych klasowo. To zjawisko ma intuicyjne wyjaśnienie. Logika rozmyta oparta na odległości zmniejsza znaczenie obiektów odstających w procesie treningu, przy jednoczesnym wzroście znaczenia obiektów typowych dla

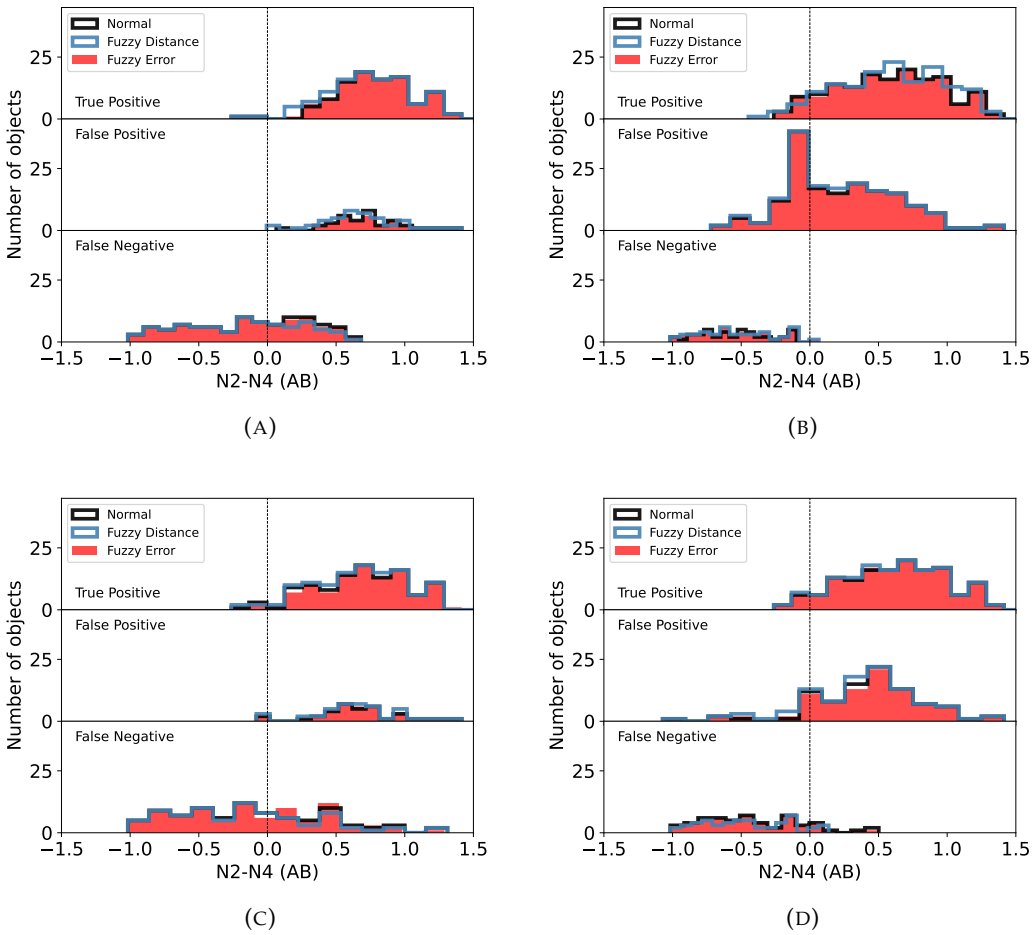


RYSUNEK 5.6: Wizualizacja wartości metryk dla różnych strategii ważenia opartych na logice rozmytej. *Normal* odpowiada modelom bez logiki rozmytej, *distance* odpowiada modelom z zastosowaniem logiki rozmytej opartej na odległości od środka klasy, *error* odpowiada modelom z zastosowaniem logiki rozmytej opartej na niepewnościach pomiarowych. *Panel A:* Precision (czystkość katalogu klasy pozytywnej). *Panel B:* Recall (kompletność katalogu klasy pozytywnej). *Panel C:* Metryka F1. *Panel D:* PR AUC. *Panel E:* Zrównoważona dokładność (bACC). Wykres na panelu A pochodzi z pracy Poliszczuk i in. (2021).

danej klasy. W ten sposób znaczenie treningowych AGN-ów leżących w obszarze zajmowanym przez galaktyki staje się niższe. Ta zmiana nie wpływa znacząco na obniżenie kompletności katalogu, ponieważ te regiony przestrzeni cech zajmują w przeważającej mierze galaktyki. Jednocześnie wzrasta czystość katalogu, ponieważ prawdopodobieństwo wyboru kandydata na AGN daleko od centrum klasy AGN-ów staje się niższe. Wzrost znaczenia typowych obiektów oddala granice separacji klasy od obszaru zajmowanego przez klasę AGN-ów. W ten sposób kompletność katalogu AGN-ów wzrośnie w regionie zajmowanym głównie przez AGN-y w zbiorze treningowym. Dodatkowo można zaobserwować wzrost zanieczyszczenia katalogu AGN-ów w regionie zajętym głównie przez AGN-y, ze względu na spadek znaczenia SFG o dużym przesunięciu, spowodowany ich niskimi wagami opartymi na odległości. Podsumowując, widzimy wyraźną tendencję wzrostu kompletności katalogu AGN (recall) i przeciwstawne mechanizmy, które mogą zwiększać lub zmniejszać czystość katalogu AGN (precision). W rezultacie mamy oparty na odległości model logiki rozmytej z wyższą wartością metryki recall i podobną wartością metryki precision w porównaniu z początkowym modelem bez zastosowania logiki rozmytej. Widzimy, że logika rozmyta ma największy wpływ na modele ważenia klasowego, gdzie granica decyzyjna modelu jest przesunięta w kierunku mniejszej klasy. Tutaj waga oparta na odległości od środka klasy może uzyskać swój skumulowany wpływ na klasyfikację. Jednak w przypadku modeli zrównoważonych klasowo, granica decyzjna jest już odsunięta od mniejszej klasy przez wagi klasowe. Główny wpływ, jaki ważenie oparte na odległości od środka klasy może mieć na modele zrównoważone klasowo, polega na zmniejszeniu znaczenia obserwacji odstających leżących daleko od środka klasy.

Logika rozmyta oparta na niepewnościach pomiarowych nie ma znaczącego wpływu na jakość klasyfikacji zarówno w modelach zrównoważonych klasowo, jak i niezrównoważonych klasowo. Pomimo fizycznej motywacji leżącej u podstaw logiki rozmytej opartej na błędach pomiarowych, może ona spowodować tylko niewielką zmianę we właściwościach predykcji klasyfikatora. Zjawisko to jest najprawdopodobniej spowodowane niską korelacją między konkretnym problemem klasyfikacyjnym a niepewnością pomiaru. Obserwacje, które powinny być traktowane jako najważniejsze ze względu na ich pozycję w przestrzeni cech, mogą nie charakteryzować się najwyższą dokładnością pomiaru. Dlatego ten rodzaj logiki rozmytej może powodować dwa sprzeczne zachowania. Z jednej strony, obiekty, które zostały źle zmierzone i w konsekwencji zostały przesunięte do innej części przestrzeni cech, nie stanowiłyby większego problemu, ponieważ wagi oparte na błędach zmniejszają ich wpływ na klasyfikację. Z drugiej strony, źle zmierzone obiekty, które są kluczowe dla klasyfikacji i znajdują się w odpowiedniej objętości przestrzeni cech, nie byłyby w stanie dostarczyć modelowi wystarczających informacji.

Teraz przeanalizujmy, jak logika rozmyta wpływa na rozkład predykcji kolorów $N2$–$N4$ przedstawiony na rysunkach 5.7 i 5.8. We wszystkich przedstawionych przypadkach można zauważyć hierarchiczny charakter wag klasowych i opartych na logice rozmytej. Wagi klasowe są bardziej fundamentalne i mają większy wpływ na predykcję, podczas gdy wagi oparte na logice rozmytej należy traktować jako bardziej subtelną technikę dostrajania. W przypadku regresji logistycznej, SVM i XGBoost zastosowanie wag klasowych ma kilka wspólnych implikacji, z których wszystkie są związane z przesunięciem granicy decyzjnej modelu w kierunku niebieskich wartości koloru $N2$–$N4$. Po pierwsze, widzimy, że wagi klasowe powodują przesunięcie obiektów *True Positive* (prawidłowo wybranych AGN-ów) w kierunku niebieskich wartości koloru $N2$–$N4$. Jak już wcześniej wykazano, niebieski zakres kolorów $N2$–$N4$ jest zajęty głównie przez galaktyki. Dlatego zwiększenie wagi obserwacji



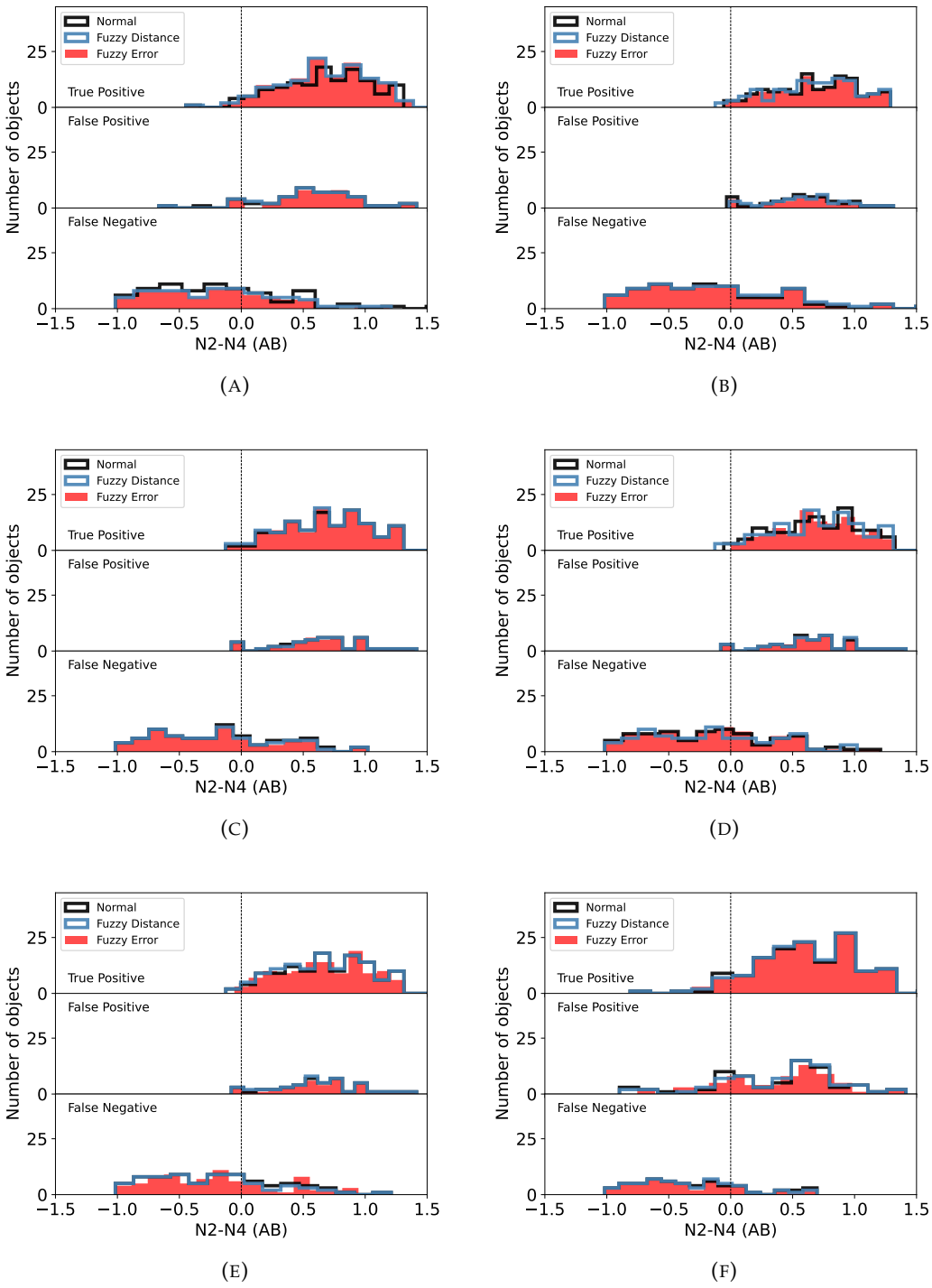
RYSUNEK 5.7: Rozkład koloru $N2$–$N4$ przedstawia wpływ różnych wag opartych na logice rozmytej na klasyfikację danych oznaczonych. Klasa pozytywna odnosi się do klasy AGN-ów, a klasa negatywna do klasy galaktyk. Obiekty *True Positive* to prawidłowo sklasyfikowane AGN-y. Obiekty *False Positive* to błędnie sklasyfikowane galaktyki, czyli zanieczyszczenie katalogu AGN. Obiekty *False Negative* to AGN-y błędnie sklasyfikowane jako galaktyki. *Normal* odpowiada modelom bez logiki rozmytej, *distance* odpowiada modelom z zastosowaniem logiki rozmytej opartej na odległości od środka klasy, *error* odpowiada modelom z zastosowaniem logiki rozmytej opartej na niepewnościach pomiarowych. *Panel A:* Regresja logistyczna bez wag klasowych. *Panel B:* Regresja logistyczna z wagami klasowych. *Panel C:* SVM bez wag klasowych. *Panel D:* SVM z wagami klasowymi.

mniejszej klasy przesuwa separację między klasami na zewnątrz od mniejszej klasy (AGN). Ponadto wzrost znaczenia AGN-ów powoduje, że obszar klasy AGN-ów (tzn. czerwony zakres kolorów $N2$–$N4$) jest zajmowany prawie wyłącznie przez AGN-y, przy jednoczesnym zmniejszeniu wpływu SFG o wysokim przesunięciu ku czerwieni na klasyfikację. W związku z tym w modelach zrównoważonych klasowo można zaobserwować znaczny wzrost zanieczyszczenia katalogu AGN-ów w czerwonym zakresie kolorów $N2$–$N4$, wraz ze spadkiem liczby obiektów *False Negative* (tzn. AGN-ów zaklasyfikowanych jako galaktyki). Jedynymi modelami, w których nie można zaobserwować znaczącego wpływu wag klasowych, są las losowy (RF) i algorytm wyjątkowo losowych drzew (ET), o czym była mowa wcześniej. Analiza wpływu

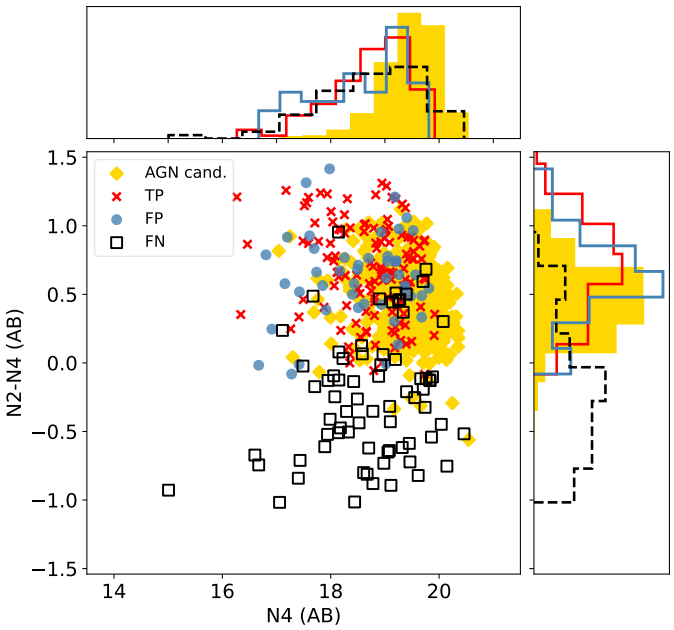
logiki rozmytej nie wykazuje żadnych znaczących zmian w rozkładzie kolorów $N2$–$N4$. Jednakże, wyciągając wnioski o niewielkim wpływie wag rozmytych, należy pamiętać, że kolor $N2$–$N4$, pomimo swojej użyteczności, nie jest cechą w pełni reprezentatywną, a bardziej znaczące zmiany mogą wystąpić w wielowymiarowej przestrzeni cech.

5.4.3 Końcowy model i katalog kandydatów na AGN-y

Skupmy się teraz na ocenie własności końcowego modelu, tj. systemu twardego głosowania. Znany już Czytelnikowi wykres $N4$ vs $N2$–$N4$ dla predykcji na oznaczonych danych jak i próbce generalizacyjnej, jest pokazany na rysunku 5.9. Można tu zaobserwować tendencje obecne w opisanych wcześniej modelach składowych. Jedną z takich wyraźnych tendencji jest to, że AGN-y są w przeważającej mierze wybierane w czerwonej części koloru $N2$–$N4$. Dwa czynniki powodują to zjawisko. Po pierwsze użyliśmy wyższego wyniku metryki precision jako ważniejszej właściwości klasyfikatora przy porównywaniu modeli poprzez stosunek metryk precision i recall. Dlatego ostateczny model ma mniejszą tendencję do wybierania kandydatów na AGN-y w regionach przestrzeni cech zdominowanej przez galaktyki (np. niebieska część rozkładu kolorów $N2$–$N4$). Obserwacje AGN-ów z próbki treningowej znajdujące się w niebieskiej części $N2$–$N4$ to głównie AGN-y wyselekcjonowane na podstawie promieniowania rentgenowskiego. Jak omówiono w poprzednich częściach tej pracy, AGN-y wyselekcjonowane w zakresie rentgenowskim są często trudne do odzyskania przy zastosowaniu selekcji w podczerwieni lub w pasmie optycznym. Główna przyczyna tej trudności wynika z faktu, że selekcja AGN-ów w podczerwieni i optyce próbkuje tylko koniec rozkładu stosunku L/L_{Edd}, podczas gdy selekcja rentgenowska obejmuje większość zakresu rozkładu. Główne zanieczyszczenie katalogu pochodzi z SFG znajdujących się na stosunkowo dużych przesunięciach ku czerwieni (patrz rysunek 3.1b) charakteryzujących się przeważnie czerwonym kolorem $N2$–$N4$. Bliższe spojrzenie na rozkład przesunięć ku czerwieni względem predykcji modelu przedstawionych na rysunku 5.10 pokazuje nam trzy interesujące cechy modelu klasyfikacyjnego. Po pierwsze, widzimy znaczącą różnicę między rozkładami przesunięć ku czerwieni True Positive (prawidłowo wyselekcjonowane AGN-y) i False Negative (AGN-y błędnie zaklasyfikowane jako galaktyki). To pokazuje nam, że główny składnik rozkładu AGN-ów, który wymyka się naszej selekcji, pochodzi z próbki o niskich przesunięciach ku czerwieni. Kolejna ważna obserwacja dotyczy właściwości przesunięć ku czerwieni obserwacji False Positive (tzn. galaktyk zaklasyfikowanych jako AGN-y). Widzimy, że zanieczyszczenie pochodzi nie tylko od SFG o wyższych przesunięciach ku czerwieni, ale również od galaktyk o niskich przesunięciach ku czerwieni, charakteryzujących się najprawdopodobniej znacznym składnikiem pyłowym. Wreszcie, porównanie spektroskopowego przesunięcia ku czerwieni obiektów True Positive i fotometrycznego przesunięcia ku czerwieni kandydatów na AGN-y pokazuje nam podobne właściwości tych dwóch próbek. Niewielkie przesunięcie rozkładu przesunięcia ku czerwieni kandydatów na AGN-y w kierunku mniejszych wartości jest najprawdopodobniej spowodowane systematycznym niedoszacowaniem fotometrycznych przesunięć ku czerwieni AGN-ów powyżej $z \geq 1.5$. Tak więc rzeczywisty rozkład przesunięć ku czerwieni w próbce kandydatów na AGN-y obejmuje prawdopodobnie znacznie szerszy zakres. Inną oznaką obecności AGN-ów o wysokim przesunięciu ku czerwieni w katalogu kandydatów są widoczne przy porównaniu rysunków 5.9 i 3.1b. Tutaj można zaobserwować spadek wartości koloru $N2$–$N4$ przy najwyższych przesunięciach ku czerwieni. Poza tym AGN-y o dużym przesunięciu ku czerwieni wydają się być



RYSUNEK 5.8: Rozkład koloru $N2–N4$ przedstawia wpływ różnych wag opartych na logice rozmytej (wagi instancji) na klasyfikację danych oznaczonych. Klasa pozytywna odnosi się do klasy AGN-ów, a klasa negatywna do klasy galaktyk. Obiekty *True Positive* to prawidłowo sklasyfikowane AGN-y. Obiekty *False Positive* to błędnie sklasyfikowane galaktyki, czyli zanieczyszczenie katalogu AGN. Obiekty *False Negative* to AGN-y błędnie sklasyfikowane jako galaktyki. *Normal* odpowiada modelom bez logiki rozmytej, *distance* odpowiada modelom z zastosowaniem logiki rozmytej opartej na odległości od środka klasy, *error* odpowiada modelom z zastosowaniem logiki rozmytej opartej na niepewnościach pomiarowych. *Panel A:* Random Forest bez wag klasowych. *Panel B:* Random Forest z wagami klasowymi. *Panel C:* Extremely Randomized Trees bez wag klasowych. *Panel D:* Extremely Randomized Trees z wagami klasowymi. *Panel E:* XGBoost bez wag klasowych. *Panel F:* XGBoost z wagami klasowymi.

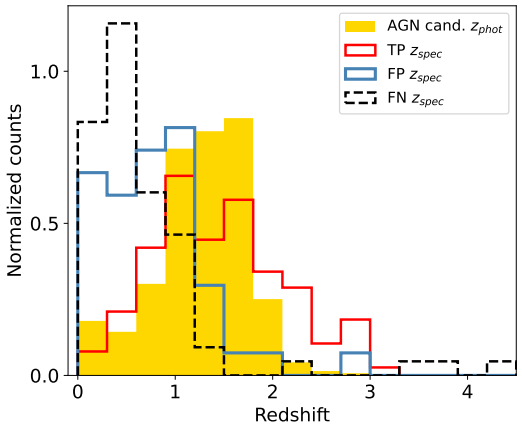


RYSUNEK 5.9: Wykres kolor-magnitudo $N2$–$N4$ vs. $N4$, wraz z histogramami gęstości odpowiadających im kolorów i wielkości gwiazdowych. Wykres przedstawia predykcje końcowego modelu na próbce oznaczonej i generalizacyjnej. True Positive (TP, czerwone krzyżyki) odnosi się do prawidłowo sklasyfikowanych AGN-ów w zbiorze danych z etykietami. False Positive (FP, niebieskie kropki) odnosi się do galaktyk błędnie sklasyfikowanych jako AGN-y. False Negative (FN, czarne kwadraty) odnosi się do AGN-ów nieprawidłowo zaklasyfikowanych jako galaktyki. Kandydaci na AGN-y, oznaczeni żółtymi rombami, odnoszą się do obserwacji z próbki generalizacyjnej, zaklasyfikowanych jako AGN-y. Kolory na znormalizowanych histogramach odpowiadają kolorom na wykresie kolor-magnitudo. Wykres pochodzi z pracy Poliszczuk i in. (2021).

ogólnie ciemniejsze. Łącząc te dwa fakty, zobaczymy podzbiór obiektów charakteryzujących się wysokimi wartościami $N4$ i $N2$–$N4 \in (0, 0.5)$. Dane treningowe słabo reprezentują ten region zajmowany przez kandydatów na AGN-y. Ich brak w próbce treningowej wynika głównie z warunków, jakie musiał spełniać obiekt treningowy, aby mógł być obserwowany przez spektrograf. Obiekty charakteryzujące się takimi właściwościami są potencjalnymi kandydatami na AGN-y o dużym przesunięciu ku czerwieni. Statystyczne właściwości ostatecznego katalogu kandydatów na AGN są przedstawione w tabeli 5.2.

5.4.4 Eksperyment ekstrapolacyjny

Jedną z wyróżniających się cech klasyfikacji dokonanej przez system twardego głosowania była trudność w selekcji AGN-ów w regionie przestrzeni cech charakteryzującym się niebieskim kolorem $N2$–$N4$. Natura tych obiektów i trudności w ich selekcji przy użyciu pasm optycznych, NIR i MIR zostały już omówione. W celu sprawdzenia, czy nowoczesne techniki ML, które operują w złożonych, wielowymiarowych



RYSUNEK 5.10: Znormalizowany histogram rozkładu przesunięcia ku czerwieni w odniesieniu do wyników predykcji końcowego modelu na zbiorze oznaczonym i próbce generalizacyjnej. True Postive (TP, kolor czerwony) odnosi się do prawidłowo sklasyfikowanych AGN-ów w zbiorze danych z etykietami. False Positive (FP, kolor niebieski) odnosi się do galaktyk błędnie zaklasyfikowanych jako AGN. False Negative (FN, kolor czarny) odnosi się do AGN-ów błędnie zaklasyfikowanych jako galaktyki. Kandydaci na AGN, oznaczeni żółtym kolorem, odnoszą się do obserwacji z próbki generalizacyjnej, zaklasyfikowanych jako AGN-y.

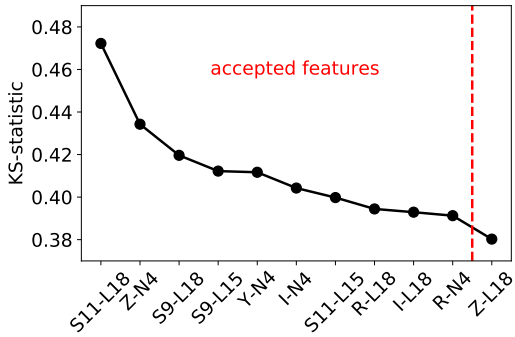
przestrzeniach, mogą przezwyciężyć ten problem, przeprowadzono dodatkowy eksperyment. Eksperyment ten będziemy nazywać *eksperymentem ekstrapolacyjnym*, a poprzednią klasyfikację będziemy nazywać *klasyfikacją główną*. W celu zwiększenia informacji dostępnej dla modelu, dodano pomiary z pasm MIR: *S*7, *S*9*W*, *S*11, *L*15 i *L*18*W* (nie użyto *L*24 ze względu na bardzo małą liczbę detekcji w tym paśmie).

Próbka treningowa została zmodyfikowana w celu stworzenia modelu skoncentrowanego na problematycznych przypadkach z głównej klasyfikacji. Obserwacje zaklasyfikowane przez system twardego głosowania podczas głównej klasyfikacji jako obiekty False Negative (tzn. błędnie zaklasyfikowane AGN-y) zostały użyte jako nowa, ograniczona próbka treningowa AGN-ów. Próba galaktyk pozostała niezmieniona. Ta modyfikacja wraz z ograniczeniami narzuconymi przez wymóg detekcji MIR dała ostateczną próbkę treningową składającą się z 705 galaktyk i 39 AGN-ów. Próbka treningowa galaktyk nie została zmodyfikowana na podpróbkę obserwacji False Positive, aby nie zmniejszać jeszcze bardziej wielkości próbki treningowej. Nowa próbka generalizacyjna została utworzona z próbki generalizacyjnej klasyfikacji głównej, poprzez dodanie prostego wymogu detekcji MIR. Nie zastosowano nowego limitu MCD na nieoznakowanych danych, aby móc porównać wyniki uogólnienia z głównej klasyfikacji i eksperymentu ekstrapolacji. W ten sposób uzyskano próbkę generalizacyjną składającą się z 2207 obiektów.

Wybór cech był oparty na tej samej metodzie, co poprzednio. Do wyboru najlepszych cech dla tego konkretnego zadania wykorzystano statystykę KS. Wartości statystyki KS są pokazane na rysunku 5.11. Podjęto subiektywną decyzję, aby wybrać dziesięć cech o najwyższych wartościach KS jako ostateczny zestaw cech. Ponieważ niektóre z pasm występowały tylko w cechach o niskiej wartości statystyki KS, pominięto wcześniej stosowany wymóg stosowania wszystkich pasm, w ostatecznym zestawie cech. Warto zauważyć, że wiele z wybranych cech w eksperymencie ekstrapolacyjnym składa się z pasm MIR.

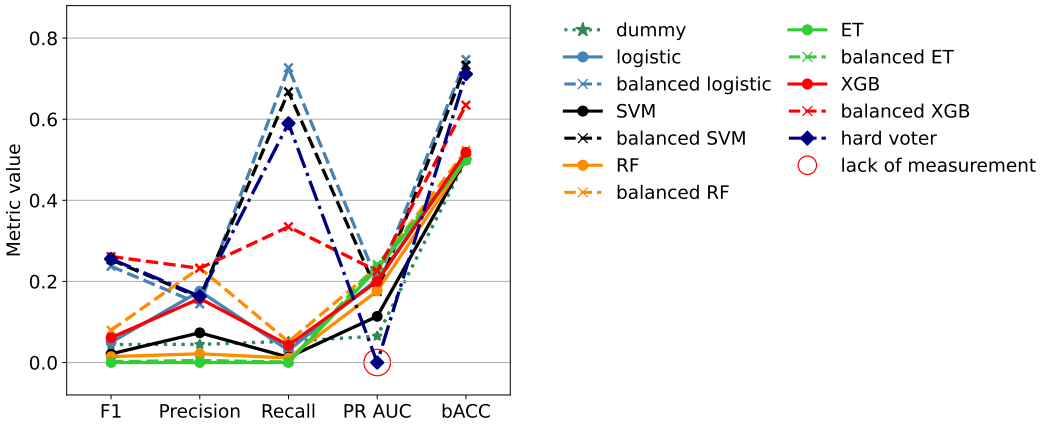
filtr	mediana	MAD	min.	max.
z_{phot}	1.264	0.376	0.005	2.841
g	22.613	1.139	18.768	27.586
r	22.061	0.963	18.764	25.544
i	21.569	0.815	18.532	24.347
z	21.292	0.725	18.165	23.708
Y	21.137	0.674	18.283	23.430
N2	19.974	0.415	17.341	20.846
N3	19.687	0.416	17.289	20.648
N4	19.515	0.405	17.041	20.546

TABLICA 5.2: Własności statystyczne katalogu kandydatów na AGN-y. Przedstawione są wartości mediany, mediany odchylenia bezwzględnego (ang. *median absolute deviation*, MAD), minimalne i maksymalne wartości przesunięcia ku czerwieni i wielkości gwiazdowych w pasmach optycznych i NIR.



RYSUNEK 5.11: Wyniki selekcji cech metodą statystyki Kołomogorowa-Smirnowa zastosowanej w eksperymencie ekstrapolacji. Pokazany jest tylko podzbiór cech z najwyższym wynikiem statystyki KS. Dany wykres pochodzi z pracy Poliszczuk i in. (2021).

Ocena wydajności eksperymentu ekstrapolacji jest przedstawiona w formie wizualnej na rysunku 5.12. Szczegółowe wartości metryk są dostępne w Dodatku B w tabelach B.3 i B.4. Tutaj widać, że tylko niektóre modele były w stanie nauczyć się zadania klasyfikacji. Zarówno niezbalansowany las losowy, jak i SVM wykazały wartości metryk w zakresie kasyfikatora naiwnego. Jeszcze gorsze wyniki otrzymano w przypadku niezbalansowanych i zbalansowanych klasowo wersji algorytmu wyjątkowo losowych drzew. Dwa z pozostałych modeli, zbalansowany klasowo las losowy i niezbalansowana regresja logistyczna, były w stanie uzyskać znaczny wzrost wartości metryki precision, ale nie mogły przezwyciężyć problemów z niskimi wartościami metryk F1 i recall. Ze względu na ten problem, niewielki podzbiór najlepszych modeli: zrównoważony klasowo SVM, regresja logistyczna i XGBoost, został wykorzystany do zbudowania dodatkowego klasyfikatora systemu twardego głosowania. Ze względu na wyższe ryzyko nadmiernego dopasowania spowodowanego bardzo niewielką ilością danych, inna forma systemu głosującego w postaci klasyfikatora stosowego nie została wykorzystana w eksperymencie ekstrapolacji. Warto zauważyć, że w eksperymencie ekstrapolacyjnym wszystkie najlepsze modele wykorzystują wagi klasowe. Jest to tendencja odwrotna do tej, którą obserwowano w klasyfikacji głównej. Porównanie stosunku klas AGN-ów do galaktyk w próbce



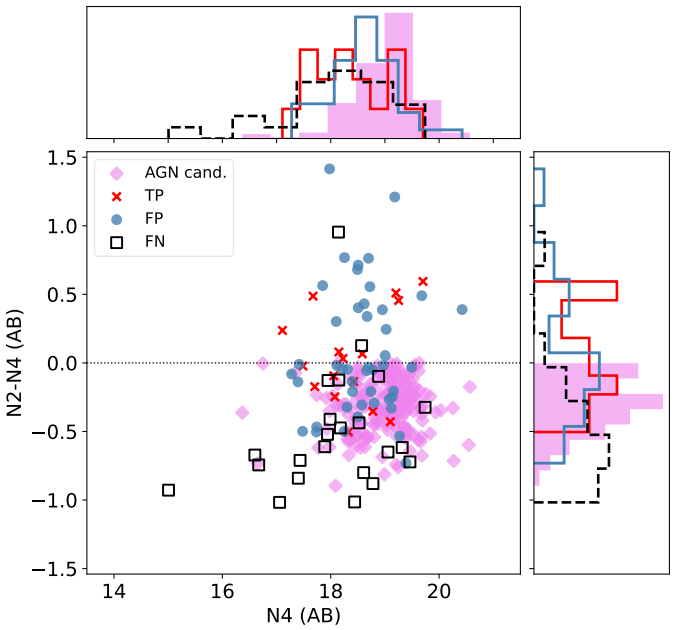
(A) (B)

RYSUNEK 5.12: Ocena jakości predykcji dla różnych modeli klasyfikacyjnych w eksperymencie ekstrapolacyjnym. Przedstawione są tylko modele bez zastosowania logiki rozmytej i system twardego głosowania. *Panel A:* Metryki oceny dla różnych modeli w porównaniu z klasyfikatorem naiwnym. *Panel B:* Legenda. Wykres pochodzi z pracy Poliszczuk i in. (2021).

treningowej w głównej klasyfikacji i w eksperymencie ekstrapolacyjnym pokazuje nam zmianę z $\sim$ 15% do $\sim$ 5%. Jest więc prawdopodobne, że wagi klas w tym zadaniu klasyfikacyjnym stają się ważną częścią modelu w przypadku bardzo silnej nierównowagi klas (tzn. gdy jedna z klas stanowi tylko kilka procent wielkości drugiej).

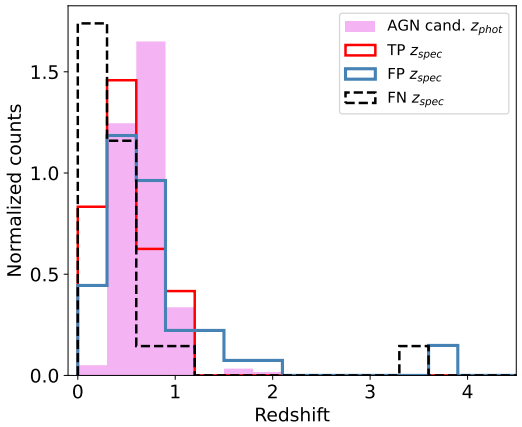
Ze względu na bardzo silne zanieczyszczenie katalogów AGN produkowanych przez wszystkie pozostałe modele, wysoki wynik metryki precision stał się głównym kryterium wyboru najlepszego modelu. Najwyższą wartość precyzji uzyskał zrównoważony klasowo model XGBoost i zrównoważony klasowo las losowy. Spośród tych dwóch modeli, XGBoost cechował się znacznie lepszymi wynikami w zakresie wyników F1, recall i bACC. Zrównoważony klasowo XGBoost, charakteryzujący się 0,25±0,11 czystością katalogu AGN-ów i 0,37±0,16 kompletnością katalogu AGN-ów, został wybrany jako ostateczny klasyfikator.

Wstępna generalizacja przeprowadzona przez ostateczny model wyprodukowała katalog 354 kandydatów na AGN-y. Główną ideą eksperymentu ekstrapolacji było przezwyciężenie problemu selekcji AGN-ów w niebieskim zakresie koloru $N2$–$N4$. Nie narzucono warunku $N2$–$N4 < 0$ na zbiór uogólniający na samym początku budowy eksperymentu ekstrapolacyjnego, ze względu na możliwość bardzo precyzyjnej selekcji w czerwonym zakresie koloru $N2$–$N4$ po dodaniu informacji MIR do danych optycznych i NIR. Rysunek 5.13 pokazuje nam jednak obecność silnego zanieczyszczenia katalogu AGN również w czerwonym zakresie kolorów. Tak więc jedyną użyteczną częścią próbki powstałej podczas procedury generalizacji, która byłaby uzupełnieniem głównego katalogu klasyfikacyjnego, były obiekty charakteryzujące się kolorem $N2$–$N4 < 0$. W ten sposób otrzymaliśmy katalog 198 obiektów ($\sim$9% próbki generalizacyjnej eksperymentu ekstrapolacyjnego). Oprócz dużego zanieczyszczenia katalogu AGN-ów w czerwonym zakresie kolorów $N2$–$N4$, nadal możemy zaobserwować ten sam problem z utratą AGN w niebieskiej części tego koloru. Są to, ponownie, głównie wyselekcjonowane w pasmie rentgenowskim AGN-y znajdujące się na niskich przesunięciach ku czerwieni (patrz rysunek 5.14). Porównanie rysunków 5.13 i 3.1a daje nam pewne pojęcie o możliwościach tego



RYSUNEK 5.13: Wykres kolor-magnitudo $N2$–$N4$ vs $N4$, wraz z odpowiednimi histogramami gęstości. Na wykresie przedstawiono przewidywania modelu eksperymentu ekstrapolacyjnego na zbiorze oznaczonym i generalizacyjnym. True Postive (TP, czerwone krzyżyki) odnosi się do prawidłowo sklasyfikowanych AGN-ów w zbiorze danych oznaczonych. False Positive (FP, niebieskie kropki) odnosi się do galaktyk błędnie sklasyfikowanych jako AGN-y. False Negative (FN, czarne kwadraty) odnosi się do AGN-ów błędnie zaklasyfikowanych jako galaktyki. Kandydaci na AGN-y, oznaczeni fioletowymi rombami, odnoszą się do obserwacji z próbki generalizacyjnej, zaklasyfikowanych jako AGN-y. Kolory na histogramach odpowiadają kolorom na wykresie kolor-magnitudo. Wykres jest zmodyfikowaną wersją wykresu opublikowanego w pracy Poliszczuk i in. (2021).

modelu klasyfikacyjnego. Jest on w stanie odzyskać niektóre AGN-y znajdujące się w obszarze zajmowanym przez galaktyki, czyli prawdopodobnie obiekty ze znaczącym wkładem emisji od galaktyki-gospodarza, ale kosztem znaczącego zanieczyszczenia katalogu AGN-ów. Jednak nawet przy bardzo dużym zanieczyszczeniu katalogu AGN-ów, klasyfikator stworzony w eksperymencie ekstrapolacji nie może odzyskać AGN-ów wyselekcjonowanych w pasmie rentgenowskim. Dowodzi to po raz kolejny zasadniczej różnicy między selekcją AGN-ów w zakresie rentgenowskim i optyczno-IR, wynikającą z różnicy fizycznych własności tak selekcjonowanych populacji AGN. Ze względu na bardzo wysoki poziom zanieczyszczenia katalogu AGN-ów (tj. niską wartość metryki precision), jego niską kompletność (tj. niski wartość metryki recall) i wspomnianą wyżej niemożność odzyskania XAGN-ów, eksperyment ekstrapolacyjny jest traktowany w tej pracy jedynie jako sposób na określenie granic możliwości selekcji AGN-ów opartej na uczeniu maszynowym. Dlatego też kandydaci na AGN-y wyselekcjonowani podczas tego eksperymentu nie zostali włączeni do ostatecznego katalogu.



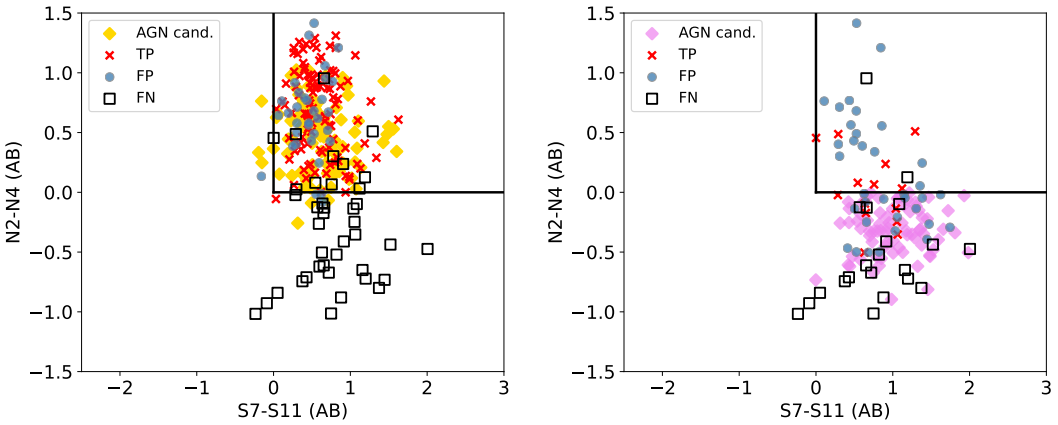
RYSUNEK 5.14: Histogram rozkładu przesunięcia ku czerwieni w odniesieniu do wyników predykcji modelu eksperymentu ekstrapolacji na zbiorze oznaczonym i próbce generalizacyjnej. True Postive (TP, kolor czerwony) odnosi się do prawidłowo sklasyfikowanych AGN-ów. False Positive (FP, kolor niebieski) odnosi się do galaktyk błędnie zaklasyfikowanych jako AGN-y. False Negative (FN, kolor czarny) odnosi się do AGN-ów błędnie zaklasyfikowanych jako galaktyki. Kandydaci na AGN-y, oznaczeni kolorem fioletowym, odnoszą się do obserwacji z próbki generalizacyjnej, zaklasyfikowanych jako AGN-y.

5.4.5 Porównanie z selekcją aktywnych jąder galaktyk w średniej podczerwieni

W tym rozdziale nowa metoda selekcji AGN-ów oparta na uczeniu maszynowym zostanie porównana z oryginalną techniką opartą na selekcji w MIR. Zakładamy pewne wspólne właściwości pomiędzy dwiema metodami selekcji ze względu na konstrukcję próbki treningowej, stosowanej w tej pracy. Wstępna selekcja kandydatów na AGN-y do głównej kampanii obserwacji spektroskopowych (Shim i in., 2013), które zostały włączone do próbki treningowej, opierała się na selekcji źródeł AKARI MIR zastosowanej w Lee i in. (2007). Te obiekty AGN stanowią większość AGN-ów obecnych w danych treningowych (dodatkowe AGN-y pochodzą z próbki wyselekcjonowanej pod kątem promieniowania rentgenowskiego i nielicznych obserwacji spektroskopowych pochodzących z innych optycznych obserwacji uzupełniających, tj. z pomiarów spektroskopowych obiektów, które okazały się AGN-ami, ale do obserwacji spektroskopowych zostały wytypowane z innych powodów).

W oryginalnej metodzie MIR, selekcja AGN-ów w przestrzeni kolorów została przeprowadzona z dodatkowym limitem $S11 < 18,5$mag nałożonym na próbkę. Limit ten został wprowadzony w celu dopasowania metody do charakterystyki katalogu AKARI NEP-Deep. W przypadku danych wykorzystanych w niniejszej pracy tylko bardzo niewielka liczba obiektów przekracza tę granicę. Są to dwa rentgenowskie AGN-y i cztery AGN-y typu I w próbce treningowej oraz dziewiętnaście kandydatów na AGN-y w katalogu końcowym. Ze względu na niewielką liczbę tych obiektów oraz fakt, że niniejsza praca jest wykonywana na katalogu AKARI NEP-Wide (a nie na katalogu AKARI NEP-Deep), a także po to by mieć bardziej bezpośrednie porównanie dwóch metod selekcji, zdecydowano się pominąć ograniczenie w pasmie $S11$.

W celu porównania metody opartej na uczeniu maszynowym z selekcją w przestrzeni kolorów MIR, próbka treningowa i uzyskany katalog kandydatów na AGN-y zostały ograniczone do obiektów wykrytych w pasmach $S7$ i $S11$. Ten dodatkowy



(A) (B)

RYSUNEK 5.15: Wykres kolorów w zakresie NIR-MIR używany do selekcji AGN-ów opisanej w (Lee i in., 2007). Kryteria selekcji tej metody są zaznaczone w prawym górnym kwadracie czarnymi liniami. Punkty obecne na wykresach odnoszą się do predykcji modelu ML na danych oznaczonych i nieoznaczonych. True Postive (TP, czerwone krzyżyki) odnosi się do prawidłowo sklasyfikowanych AGN-ów. False Positive (FP, niebieskie kropki) odnosi się do galaktyk błędnie zaklasyfikowanych jako AGN-y. False Negative (FN, czarne kwadraty) odnosi się do AGN-ów błędnie zaklasyfikowanych jako galaktyki. Kandydaci na AGN-y, oznaczeni żółtymi (klasyfikacja główna) i fioletowymi (eksperyment ekstrapolacyjny) rombami, odnoszą się do obserwacji z próby generalizacyjnej, zaklasyfikowanych jako AGN-y. *Panel A:* Klasyfikacja główna. *Panel B:* Eksperyment ektrapolacyjny. Wykresy pochodzą z pracy Poliszczuk i in. (2021).

warunek spowodował ograniczenie próbki treningowej i katalogu wynikowego odpowiednio do 815 (z 1547 początkowych obserwacji treningowych) i 113 (z 465 początkowych kandydatów na AGN-y) obiektów. W przypadku eksperymentu ekstrapolacji, nie wprowadzono dodatkowych modyfikacji danych, ponieważ warunki detekcji $S7$ i $S11$ zostały już uprzednio narzucone podczas tworzenia próbek treningowych i generalizacyjnych.

Rysunek 5.15 przedstawia wizualne porównanie metody selekcji AGN-ów opartej na kolorach MIR i opartych na technikach ML z klasyfikacji głównej (rysunek 5.15a) i eksperymentu ekstrapolacyjnego (rysunek 5.15b). Metoda ograniczeń w przestrzeni kolorów MIR opiera się na wykładniczym kształcie widmowego rozkładu energii AGN-ów w zakresie NIR i MIR. Ta własność daje AGN-om czerwone kolory w zakresach NIR i MIR, umieszczając obszar zajmowany przez AGN-y w kwadracie $N2–N4 > 0$ i $S7–S11 > 0$ na wykresie kolor-kolor. Rysunek 5.15a porównuje metodę kolorów opartą na MIR z predykcjami na danych treningowych i generalizacyjnych wykonaną przez końcowy model z głównej klasyfikacji. Widzimy znaczne podobieństwo między tymi dwiema metodami. Zdecydowana większość kandydatów na AGN-y wybranych w głównej klasyfikacji zajmuje prawy górny kwadrat (czerwone kolory $N2–N4$ i $S7–S11$), który jest wykorzystywany przez metodę opartą na MIR. Analiza klasyfikacji przeprowadzonej na oznaczonych danych pokazuje nam potwierdzenie tendencji omówionych już w rozdziale 5.4.3. Po pierwsze, widzimy nieuniknione zanieczyszczenie katalogu AGN-ów wyselekcjonowanych w podczerwieni przez składnik SFG, przedstawiony tutaj jako próbka False Positive. To zanieczyszczenie, obecne w nowej metodzie ML, jest również wyraźnie widoczne

Metoda	F1	Precision	Recall	bACC
$\begin{cases} N2-N4>0 \\ S7-S11>0 \end{cases}$	0.76	0.73	0.80	0.84
hard voter	0.75	0.77	0.74	0.86

TABLICA 5.3: Porównanie własności ostatecznego modelu głównej klasyfikacji z metodą selekcji w MIR opartą na ograniczeniach w przestrzeni kolorów (Lee i in., 2007).. Do policzenia wartości metryk wykorzystano obiekty wykryte w pasmach $S7$ i $S11$.

w selekcji AGN opartej na MIR. Po drugie, wizualizacja selekcji AGN-ów opartej na danych MIR potwierdza zasadniczą rozbieżność między selekcją w zakresie promieniowania rentgenowskiego i podczerwonego. Niebieski kolor $N2$–$N4$, który charakteryzuje większość próbki XAGN, silnie oddziela populację False Negative od głównego położenia klasy AGN. Rysunek 5.15b pokazuje nam wyniki eksperymentu ekstrapolacji na tym samym wykresie kolorów. Widzimy tutaj, że model klasyfikacyjny jest w stanie przeprowadzić selekcję AGN-ów poza kwadratem NIR-MIR stosowanym w tradycyjnej metodzie ograniczeń kolorów. Nawet przy wyjściu poza czerwony obszar kolorów, model nie jest w stanie odzyskać próbki XAGN. Można więc stwierdzić, że podejście oparte na ML nie jest w stanie przezwyciężyć podstawowych problemów selekcji AGN-ów w podczerwieni i połączyć metody selekcji w zakresie promieniowania rentgenowskiego i podczerwieni, wykorzystując jedynie dane optyczne i podczerwone.

Tabela 5.3 pokazuje porównanie metryk pomiędzy metodą MIR i metodą opartą na ML, stworzoną podczas głównej klasyfikacji. Obie metody mają bardzo podobny wynik F1 i zbliżone wartości zrównoważonej dokładności bACC. Model oparty na ML wykazuje wyższą wartość metryki precision i niższą wartość recall w porównaniu do metody MIR. Różnica ta może być częściowo spowodowana naturą końcowego wyboru modelu, w którym wyższa wartość precision została uznana za ważniejszą właściwość modelu. Ogólnie wartości metryczne wykazują dobrą zgodność pomiędzy tymi dwoma metodami, przy czym model oparty na metodzie ML jest bardziej konserwatywny, a wyjściowy katalog AGN-ów charakteryzuje się większą czystością i mniejszą kompletnością. Jednocześnie metoda oparta na ML okazuje się bardziej uniwersalna ze względu na brak warunku detekcji MIR: tylko 24% kandydatów na AGN obecnych w naszym katalogu było zaobserwowanych w pasmach MIR $S7$ i $S11$. Przeanalizujmy teraz kompromis między precyzją a kompletnością, jaki wykazuje ostateczny model oparty na ML w głównej klasyfikacji i w przypadku, gdy nałożono dodatkowy warunek detekcji MIR. Dodatkowy warunek detekcji MIR nie ma silnego wpływu na czystość katalogu AGN - wartość metryki precision wzrasta z 0,73 do 0,77. Jednocześnie widzimy znaczny wzrost kompletności katalogu AGN-ów, gdy rozpatrujemy dane z detekcją MIR - wartość metryki recall wzrasta z 0,64 do 0,74. Tak więc włączenie informacji MIR nie pomaga znacząco zmniejszyć zanieczyszczenia z próbki SFG. Pomaga natomiast zwiększyć kompletność katalogu AGN. Takie zachowanie można wyjaśnić w następujący sposób. AGN-y, które zostały wykryte w pasmach optycznych i NIR, ale nie miały pomiarów w pasmach MIR, mogą być AGN-ami o niskiej aktywności i/lub małym składniku pyłowym. W obu przypadkach obiekty te są bardzo trudne do odzyskania przy użyciu tylko informacji optycznych i NIR ze względu na ich podobieństwo do galaktyk w tym zakresie widmowym.

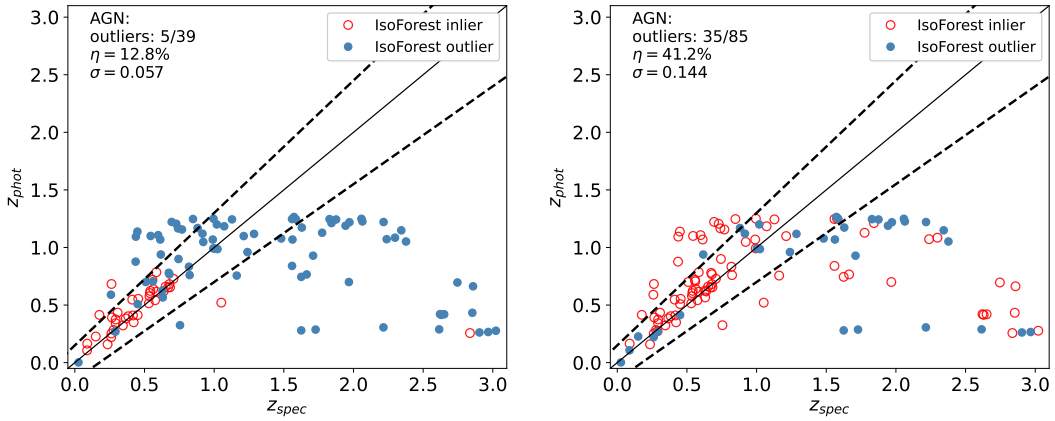
5.5 Wykrywanie obserwacji odstających

Uzyskany w tej pracy katalog AGN-ów może być wykorzystany w astrofizyce do różnych celów, takich jak stworzenie próbki docelowej do obserwacji spektroskopowych lub astrofizycznych badań własności AGN-ów. Jednak w wielu przypadkach analiza statystyczna takiego katalogu może być bardzo wrażliwa na błędy i zanieczyszczenia występujące w danych. Redukcja tych problemów jest szczególnie ważna w badaniach ewolucji galaktyk i kosmologii obserwacyjnej, takich jak to zostało omówione w rozdziale 2.3.

W tym rozdziale zbadano możliwość usunięcia części takich problemów za pomocą technik wykrywania obserwacji odstających. W szczególności, zbadano dwa zastosowania takich metod. Jednym z nich jest możliwość usunięcia katastrofalnych błędów fotometrycznego oszacowania przesunięcia ku czerwieni z katalogu wynikowego. Duże błędy fotometrycznych przesunięć ku czerwieni to poważny problem wielu obecnie istniejących katalogów fotometrycznych. To sprawia, że nie nadają się one do wielu badań kosmologicznych, w których wykorzystuje się właściwości grupowania obiektów. Dlatego usunięcie obiektów o niepoprawnie ustalonych fotometrycznych przesunięciach ku czerwieni jest niezbędne, aby taki katalog był użyteczny. Drugie zastosowanie metody wykrywania obserwacji odstających koncentruje się na usuwaniu zanieczyszczeń z katalogu kandydatów na AGN-y. W tym przypadku wykorzystuje się kombinację wykrywania obserwacji odstających i wizualizacji wielowymiarowej do identyfikacji podejrzanych obiektów i łączenia ich z określonymi źródłami zanieczyszczeń obecnymi w próbce treningowej, takimi jak SFG o wysokim przesunięciu ku czerwieni, galaktyki pyłowe lub AGN-y o niskiej aktywności. Obecność tych obiektów w katalogu wynikowym czyni go mało wiarygodnym dla badań populacji AGN-ów. Dlatego tego typu czyszczenie katalogu jest również bardzo ważne.

5.5.1 Wykrywanie błędnych fotometrycznych przesunięć ku czerwieni

Pierwszym rodzajem detekcji wartości odstających przeprowadzonym w tej pracy była detekcja wartości odstających na podstawie przesunięcia ku czerwieni za pomocą algorytmu Isolation Forest. Głównym celem tej metody było wykrycie obiektów w katalogu wynikowym z niewłaściwie oszacowanymi fotometrycznymi przesunięciami ku czerwieni. Metoda została stworzona w następujący sposób. Po pierwsze, wybrano odpowiedni limit przesunięcia ku czerwieni, aby zmniejszyć prawdopodobieństwo katastrofalnego błędu oszacowania przesunięcia ku czerwieni. Następnie dopasowano model Isolation Forest do danych treningowych w przestrzeni cech utworzonej z cech używanych w głównej klasyfikacji wraz ze spektroskopowym przesunięciem ku czerwieni. Po dopasowaniu modelu do danych, został on wykorzystany do predykcji zarówno na danych oznaczonych, jak i w katalogu wynikowym. Podczas przewidywania zmodyfikowano zestaw cech: zamiast spektroskopowego przesunięcia ku czerwieni wprowadzono jego fotometryczną estymację z pracy Ho i in. (2021). Podczas treningu określona wartość przesunięcia ku czerwieni spektroskopowej jest dopasowywana do pozycji obserwacji w wielowymiarowej przestrzeni kolorów. W następnym kroku model dopasowany do rzeczywistych wartości spektroskopowego przesunięcia ku czerwieni będzie postrzegał wartość fotometrycznego przesunięcia ku czerwieni, która silnie różni się od spektroskopowej, jako znaczące przesunięcie pozycji obserwacji w przestrzeni cech i zaklasyfikuje taką obserwację jako odstającą. W ten sposób można wykryć wątpliwe przypadki fotometrycznego

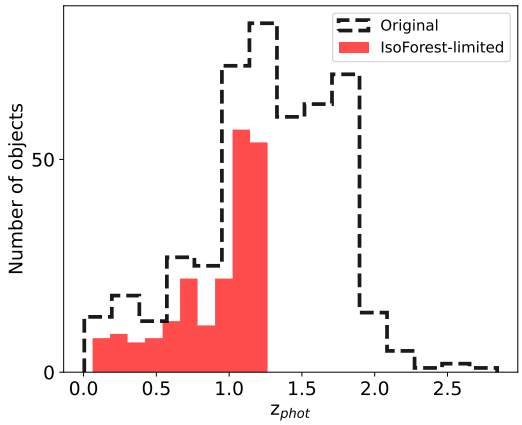


(A) (B)

RYSUNEK 5.16: Porównanie między spektroskopowym przesunięciem ku czerwieni a fotometrycznym oszacowaniem przesunięcia ku czerwieni z pracy Ho i in. (2021), pokazujące wyniki wykrywania wartości odstających za pomocą algorytmu Isolation Forest. Stożki wyznaczone przez linie przerywane, jak również parametry η i s zostały obliczone w taki sam sposób, jak opisano na rys. 3.2. Czerwone kółka i niebieskie kropki odnoszą się do obiektów zidentyfikowanych przez model Isolation Forest odpowiednio jako obiekt typowy (*inlier*) i obiekt odstający (*outlier*). *Panel A:* Predykcje modelu Isolation Forest wytrenowanego na połączonych danych treningowych (galaktyki i AGN-y). *Panel B:* Predykcje modelu Isolation Forest wytrenowanego tylko na danych AGN.

oszacowania przesunięcia ku czerwieni i uzyskać czysty katalog wynikowy do konkretnych zastosowań w kosmologii obserwacyjnej i badaniach grupowania.

Pierwszym ważnym krokiem w konstrukcji tego mechanizmu wykrywania wartości odstających jest wybór jednolitej górnej granicy przesunięcia ku czerwieni zarówno dla treningu modelu, jak i dla predykcji. Ponieważ dla katalogu wynikowego dostępne jest tylko fotometryczne przesunięcie ku czerwieni, zostało ono wykorzystane do stworzenia limitu. Aby znaleźć optymalną wartość, zastosowano dwa kryteria: z jednej strony potrzebne jest zachowanie jak największej liczby kandydatów w katalogu wynikowym. Z drugiej strony, trzeba uniknąć zakresów przesunięć ku czerwieni o znacznym obciążeniu statystycznym, obecnym w fotometrycznych oszacowaniach przesunięć ku czerwieni. Aby znaleźć granicę, która nie ogranicza wielkości katalogu wynikowego w sposób, który sprawia, że ograniczony katalog nie nadaje się już do rozsądnej analizy statystycznej, obliczono kwartyle rozkładu fotometrycznego przesunięcia ku czerwieni. Pierwszy kwartyl, który zawiera 25% całego katalogu wynikowego, został ustalony na $z_{phot} = 1,033$. Drugi i trzeci kwartyl (50% i 75% katalogu) ustalono odpowiednio na $z_{phot} = 1,264$ i $z_{phot} = 1,644$. Analizę tych wartości wraz z właściwościami przesunięcia ku czerwieni próbki treningowej pokazano na rysunku 3.2. Wynika stąd że drugi kwartyl jest rozsądną górną granicą przesunięcia ku czerwieni. Patrząc na wykres różnic między wartością spektroskopową i fotometrycznym oszacowaniem (dolna część rysunku 3.2) można zaobserwować silną zmianę przy $z_{spec} \simeq 1.5$. Przy $z_{spec} \leq 1.5$ możemy zaobserwować małe obciążenie statystyczne z lekką tendencją do przeszacowywania. Na wyższych zakresach przesunięcia ku czerwieni $z_{spec} > 1,5$, widoczne jest znaczne odchylenie w kierunku niedoszacowania fotometrycznego przesunięcia ku czerwieni.



RYSUNEK 5.17: Znormalizowany histogram porównujący rozkład fotometrycznych przesunięć ku czerwieni całego katalogu wynikowego uzyskanego w głównej klasyfikacji i podpróbki tego katalogu ograniczonej przez model Isolation Forest.

Główna przyczyna istnienia tego obciążenia statystycznego wynika nie tylko z trudności dokładnego fotometrycznego oszacowania przesunięcia ku czerwieni przy większych odległościach i oszacowania przesunięcia ku czerwieni AGN-ów w szczególności, ale także ze specyficznych celów postawionych w pracy Ho i in. (2021). Praca ta miała na celu efektywne fotometryczne oszacowanie przesunięć ku czerwieni dla galaktyk obecnych w polu AKARI NEP-Wide. Ponieważ próbka galaktyk znajduje się średnio w niższych zakresach przesunięć ku czerwieni niż próbka AGN, podejście zostało zoptymalizowane w kierunku niższego zakresu tej wielkości. Ponadto autorzy tej pracy wykazali, że uwzględnienie szablonów AGN zmniejsza skuteczność oszacowania przesunięcia ku czerwieni próbki galaktyk. Dlatego w rezultacie szablony AGN nie zostały wykorzystane podczas ostatecznej estymacji fotometrycznych przesunięć ku czerwieni. Wśród planów na przyszłość współpracy AKARI-NEP jest rozwinięcie tej pracy poprzez odrębne szacowanie przesunięć ku czerwieni galaktyk i AGN-ów, po zastosowaniu metod klasyfikacji przedstawionych w prezentowanej rozprawie doktorskiej; takie podejście powinno pozwolić na poprawę jakości fotometrycznych przesunięć ku czerwieni całego katalogu. Z perspektywy tej pracy widzimy, że fotometryczne przesunięcia ku czerwieni dla odległych AGN-ów mogą być błędne z dwóch powodów: niezdolności modelu estymacji przesunięcia ku czerwieni do pracy z obiektami o dużym przesunięciu ku czerwieni i do pracy z AGN-ami w szczególności. Dlatego sięganie daleko poza zakres rozkładu przesunięć ku czerwieni dostępnych do analizy galaktyk wymaga daleko posuniętej ostrożności. Szacowanie przesunięcia ku czerwieni AGN-ów wydaje się być wystarczająco dokładne w zakresie przesunięcia ku czerwieni obecnego w próbce galaktyk, z niewielką liczbą katastrofalnych błędów szacowania. Dodatkowy argument pochodzi z czysto statystycznych właściwości. Wykorzystując dane zarówno dotyczące galaktyk, jak AGN-ów do treningu modelu Isolation Forest, mamy większe prawdopodobieństwo, że obiekty o wysokich przesunięciach ku czerwieni zostaną zaklasyfikowane jako obiekty odstające ze względu na ich niewielką liczbę w próbce. Ponadto duża rozpiętość zakresu przesunięć ku czerwieni utrudni prawidłowe zlokalizowanie pozycji obiektów odstających w wielowymiarowej przestrzeni cech. Z tego względu poszukiwanie AGN-ów o wysokim przesunięciu ku czerwieni może w przyszłości stanowić odrębny projekt z zakresu uczenia maszynowego, wymagający innego doboru zarówno próbek treningowych, jak i metod.

W celu sprawdzenia właściwości wykrywania obserwacji odstających, wytrenowano model Isolation Forest na zestawach danych ograniczonych do pierwszych trzech kwartyli i dokonano przewidywań na dostępnych zbiorach danych. Aby przeanalizować właściwości każdego zestawu danych, sprawdzano, jak AGN-y treningowe, które zostały wybrane przez model jako obiekty typowe, pasują do stożków przesunięć ku czerwieni pokazanych na rysunku. 3.2. Widoczny jest silny spadek dokładności oszacowania przesunięcia ku czerwieni wraz ze wzrostem limitu przesunięcia ku czerwieni. Taki spadek występuje w obu przypadkach: gdy próbka treningowa jest zbudowana zarówno z AGN-ów jak i galaktyk oraz gdy próbka jest złożona jedynie z AGN-ów. W przypadku połączonych danych treningowych (galaktyki i AGN-y) otrzymujemy $\eta_{Q1} = 13,2\%$, $\eta_{Q2} = 12,8\%$ i $\eta_{Q3} = 21,3\%$ odpowiednio dla pierwszego, drugiego i trzeciego kwartyla. W przypadku modelu dopasowanego tylko do danych AGN otrzymujemy $\eta_{Q1} = 29,6$, $\eta_{Q2} = 41,2$ i $\eta_{Q3} = 47,0$ dla pierwszych trzech kwartyli. Widzimy tutaj znacznie lepsze wyniki modelu wykrywania obserwacji odstających, wytrenowanego na połączonych danych galaktyk i AGN-ów we wszystkich zakresach przesunięć ku czerwieni. Słaba wydajność modelu wytrenowanego na danych złożonych wyłącznie z AGN-ów wynika prawdopodobnie z dużego rozproszenia danych znajdujących się w wielowymiarowej przestrzeni cech. Innymi słowy, niewielka ilość danych treningowych z dużymi odległościami między punktami w przestrzeni wielowymiarowej utrudnia modelowi znalezienie pozycji prawdziwych obserwacji odstających. Analizując wydajność modelu wytrenowanego na połączonych próbkach galaktyk i AGN-ów, widzimy, że może on ustabilizować poziom zanieczyszczenia fotometrycznego katalogu przesunięć ku czerwieni w pewnym zakresie przesunięć ku czerwieni. Procent błędów katastroficznych dla modeli działających w pierwszym i drugim kwartylu jest bardzo podobny. Ze względu na te obserwacje, drugi kwartyl ($z_{phot} = 1,264$) został wybrany jako górna granica przesunięcia ku czerwieni dla ostatecznego modelu wykrywania obserwacji odstających, a kombinacja próbek galaktyk i AGN została użyta jako dane treningowe.

Na rysunku 5.16 pokazano wyniki wykrywania wartości odstających dla danych ograniczonych do drugiego kwartyla przy użyciu próbki treningowej złożonej z AGN-ów i galaktyk (rys. 5.16a) oraz tylko AGN-ów (rys. 5.16b). Widać tutaj tendencje, które zostały opisane przy analizie wpływu limitu przesunięcia ku czerwieni na wydajność modelu. Model wytrenowany na danych złożonych z AGN-ów i galaktyk wykrywa dwa rodzaje obserwacji odstających. Pierwszy rodzaj to obiekty znajdujące się ogólnie na wyższych przesunięciach ku czerwieni. Te obserwacje zostały wybrane jako odstające, ponieważ większość galaktyk znajduje się na znacznie niższych przesunięciach ku czerwieni. Próbka AGN-ów, przesunięta w kierunku wyższych przesunięć ku czerwieni, jest zbyt mała, aby zmienić zachowanie modelu. Drugi typ wartości odstających to najbardziej problematyczne AGN-y o wysokim przesunięciu ku czerwieni, które mają błędne fotometryczne oszacowania przesunięcia ku czerwieni. Model wytrenowany na połączonych danych AGN-ów i galaktyk jest bardzo dobry w wykrywaniu takich obiektów. Prawie wszystkie te problematyczne obserwacje z katastrofalnymi błędami przesunięcia ku czerwieni zostały znalezione i usunięte z katalogu. Model wytrenowany na danych złożonych wyłącznie z AGN-ów wykazuje inne tendencje. Tutaj nie widzimy systematycznego odrzucania obiektów o wysokim przesunięciu ku czerwieni. Zamiast tego model może operować na całym zakresie przesunięć ku czerwieni określonym przez drugi kwartyl. Ta elastyczność idzie jednak w parze ze znacznie większym zanieczyszczeniem katalogu końcowego. Widzimy, że model nie jest w stanie wykryć większości obiektów odstających, wykazujących zarówno przeszacowanie, jak i niedoszacowanie fotometrycznego przesunięcia ku czerwieni.

band	median	MAD	min.	max.
z_{phot}	1.034	0.263	0.062	1.264
g	22.138	1.079	19.061	25.826
r	21.557	0.895	18.764	24.516
i	21.106	0.703	18.695	23.569
z	20.832	0.593	18.165	22.723
Y	20.719	0.556	18.283	22.417
N2	19.879	0.451	17.341	20.829
N3	19.690	0.469	17.289	20.648
N4	19.516	0.443	17.299	20.355

TABLICA 5.4: Własności statystyczne katalogu kandydatów na AGN-y bez obiektów ze źle oszacowanym fotometrycznym przesunięciem ku czerwieni. Przedstawione są wartości mediany, mediany odchylenia bezwzględnego (ang. *median absolute deviation*, MAD), minimalne i maksymalne wartości przesunięcia ku czerwieni i wielkości gwiazdowych w pasmach optycznych i NIR.

Wykorzystując ostateczny model lasu izolacyjnego wytrenowanego na połączonych danych AGN-ów i galaktyk i ograniczonego do drugiego kwartyla w rozkładzie fotometrycznych przesunięć ku czerwieni, przeprowadzono detekcję obiektów odstających w katalogu wynikowym. W wyniku tego otrzymano katalog 210 obiektów odstających. Porównanie rozkładu fotometrycznych przesunięć ku czerwieni dla oryginalnego katalogu wynikowego i wybranych obiektów odstających pokazano na rysunku 5.17. Widać podobieństwo kształtu tych dwóch rozkładów. Czyszczenie katalogu za pomocą algorytmu Isolation Forest nie wprowadziło większych zmian w rozkładzie przesunięć ku czerwieni w zakresie drugiego kwartyla. Właściwości katalogu kandydatów na AGN-y z usuniętymi dużymi błędami fotometrycznych przesunięć ku czerwieni są przedstawione w Tabeli 5.4.

5.5.2 Wykrywanie obserwacji odstających w oparciu o klasy obiektów

Drugi rodzaj wykrywania obserwacji odstających, zbadany w tej pracy, rozwiązuje problem zanieczyszczenia katalogu kandydatów na AGN i tworzy grunt do poszukiwania obiektów nietypowych. Ten problem oparty na klasach jest bardziej subtelny niż poprzednio opisane wykrywanie obserwacji odstających w przestrzeni przesunięć ku czerwieni. Wiele rodzajów obiektów odstających, które chcemy wykryć, wraz z mniej wyraźnymi różnicami między nimi, powoduje konieczność zmodyfikowania metody ich wykrywania. Ten problem wykrywania obserwacji odstających będziemy nazywać *klasową detekcją obserwacji odstających*, natomiast metodę opisaną w poprzednim rozdziale będziemy nazywać *regresyjną detekcją obserwacji odstających*.

Wykrywanie obiektów odstających w oparciu o klasy zostało stworzone w następujący sposób. Po pierwsze, stworzono dwa oddzielne modele Isolation Forest. Jeden został dopasowany do próbki treningowej galaktyk (będziemy go nazywać *Galaxy Isolation Forest*), a drugi do próbki treningowej AGN (będziemy go nazywać *AGN Isolation Forest*). W obu przypadkach zastosowaliśmy ten sam zestaw cech, co w klasyfikacji głównej. W przypadku wykrywania obiektów odstających interesują nas dwa rodzaje obiektów. Jednym z nich są obiekty odstające wykryte przez AGN Isolation Forest. Mogą to być ewentualne AGN-y, które nie są typowe dla populacji AGN-ów z próbki treningowej, np. AGN-y o dużym przesunięciu ku czerwieni lub AGN-y typu II. Obiekty odstające wykryte przez AGN Isolation Forest mogą być

również zanieczyszczeniami katalogu AGN, np. SFG, które mogą nie pasować do populacji AGN-ów w przestrzeni cech kolorystycznych. Drugim rodzajem obiektów są obiekty typowe (ang. *inlier*) dla Galaxy Isolation Forest. Takie obietky znalezione w katalogu wynikowym mogą być różnymi zanieczyszczeniami, takimi jak SFG, AGN-y o niskiej aktywności lub galaktyki o silnej składowej pyłowej.

Do analizy predykcji modeli Isolation Forest wykorzystamy wcześniej wymienione podklasy obecne w danych treningowych, tj. AGN-y z próbki AGN1 (163 obiekty) i XAGN (34 obiekty) oraz galaktyki wraz z podpróbą SFG o wysokim przesunięciu ku czerwieni (HzSFG, 17 obiektów). Grupa HzSFG składała się z galaktyk charakteryzujących się kolorem $N2–N4 > 0$ i $z_{spec} \geq 1$. Wykorzystaliśmy tutaj tylko obserwacje, w których spektroskopowy redshift został zmierzony przy użyciu co najmniej dwóch linii emisyjnych.

Teraz przeanalizujmy przewidywania modeli Isolation Forest na oznaczonych danych. Rysunki 5.18b and 5.18a pokazują wykres dla przewidywań modeli Isolation Forest na oznaczonych próbkach AGN-ów i galaktyk. Przeanalizujmy najpierw, jak modele Isolation Forest interpretują rozkład próbki AGN-ów. Obserwacje traktowane przez Isolation Forest jako obserwacje odstające to głównie obiekty znajdujące się w obszarze galaktyk. To samo dotyczy obserwacji wybranych jako obiekty typowe przez model Galaxy Isolation Forest. Warto zauważyć, że obie te próbki obserwacji mają tylko niewielkie przecięcie kilku obiektów. Dlatego te dwa modele Isolation Forest są wrażliwe na różne typy obiektów. W przypadku próbek galaktyk i HzSFG widzimy, że zdecydowana większość HzSFG może być sklasyfikowana jako obserwacje odstające przez AGN Isolation Forest. Rysunek pokazuje nam ten sam wykres kolor-magnitudo z lokalizacją kandydatów na AGN-y wybranych jako obiekty odstające przez AGN Isolation Forest lub jako obiekty typowe przez Galaxy Isolation Forest. Położenie tych dwóch podpróbek jest bardzo różne. Obietky odstające wykryte przez AGN Isolation Forest, które stanowią próbkę 21 obiektów, zajmują ten sam region, co HzSFG z oznaczonej próbki. Obietky typowe wykryte przez Galaxy Isolation Forest znajdują się w regionie, który zawiera wiele galaktyk treningowych i jest niedoreprezentowany przez treningowe AGN-y. Jak wspomniano w poprzednich rozdziałach, obiekty te mogą być AGN-ami o niskiej aktywności lub galaktykami pyłowymi. Moogą to być również AGN o wysokim przesunięciu ku czerwieni. Dlatego usuwając te nietypowe obiekty z katalogu kandydatów na AGN, zmniejszamy zanieczyszczenie katalogu końcowego kosztem możliwego wykluczenia AGN-ów o dużym przesunięciu ku czerwieni.

Analiza wykresu kolor-magnitudo może nie wystarczyć do właściwego zbadania właściwości obserwacji odstających i typowych wykrytych przez różne modele Isolation Forest. Aby pogłębić to badanie, dodano kolejny krok do metody wykrywania obserwacji odstających w postaci wizualizacji algorytmu tSNE. Wykorzystano algorytm tSNE do uzyskania dwuwymiarowej reprezentacji wielowymiarowej przestrzeni cech i dokładniejszej wizualizacji względnych odległości i połączeń różnych grup obiektów. Głównym parametrem dostrajania w algorytmie tSNE jest parametr *perplexity*, który definiuje najważniejszą skalę, która ma być zachowana w wizualizacji. Innymi słowy, duża wartość parametru perplexity daje nam więcej informacji o subtelnych, małoskalowych powiązaniach między obserwacjami, podczas gdy mała wartość tego parametru podkreśla ogólne, wielkoskalowe właściwości danych. Aby znaleźć najbardziej odpowiednią wartość perplexity, sprawdzono, jak ten parametr wpływa na rozkład różnych klas obiektów w danych treningowych. Przetestowano zestaw wartości parametru perplexity: 30, 50, 80 i 150. Okazało się, że podczas gdy ogólny kształt rozkładu galaktyk i AGN-ów pozostaje niezmieniony w dwuwymiarowej wizualizacji tSNE, próbka HzSFG jest szczególnie wrażliwa na zmiany wartości

tego parametru. Ponieważ ta klasa obiektów jest znaczącym źródłem zanieczyszczenia katalogu AGN-ów, ważne było, aby mieć dobrze zlokalizowane skupisko HzSFG w wizualizacji tSNE. Takie silne zgrupowanie uzyskano przy wartości parametru perplexity równej 80 i ta wartość została wykorzystana w dalszej analizie.

Rysunek 5.19a pokazuje nam dwuwymiarową wizualizację rozkładu danych treningowych. Sztuczne wymiary tSNE będziemy określać jako *tSNE1* i *tSNE2*. Podobnie jak w przypadku analizowanego wcześniej wykresu kolor-magnitudo, tutaj widzimy bardzo wyraźny podział na dwie klasy. Jeden duży region jest silnie zdominowany przez galaktyki z niewielkim udziałem AGN-ów typu I i AGN-ów rentgenowskich. W drugim dużym regionie przeważają AGN-y. W przypadku regionów zdominowanych przez AGN-y, widzimy dwa główne regiony zanieczyszczeń. Jednym z nich jest region położony w zakresie tSNE1$\in [0, 20]$ and tSNE2$\in [-10, 0]$. Są to głównie galaktyki o niskim przesunięciu ku czerwieni, podobne do wyglądu AGN. Drugi region zanieczyszczenia znajduje się w zakresie $\in [20, 40]$ and tSNE2$\in [0, 10]$ i jest zajęty głównie przez HzSFG. Rysunek 5.19b przedstawia rozkład katalogu kandydatów na AGN w odniesieniu do danych treningowych. Widzimy tutaj bardzo dobrą zgodność pomiędzy lokalizacją katalogu kandydatów na AGN-y a lokalizacją danych treningowych AGN. Z wyjątkiem dwóch obserwacji, wszyscy kandydaci z katalogu wynikowego znajdują się dokładnie w regionie zdominowanym przez AGN-y. Rozmieszczenie kandydatów na AGN-y, którzy zostali wybrani jako obiekty odstające lub typowe przez modele AGN i Galaxy Isolation Forest, jest pokazane na Rys. 5.19c. Widzimy tutaj, że obie grupy znajdują się w strefach zanieczyszczenia regionów próbki AGN-ów. Kandydaci na AGN-y, zidentyfikowani jako obiekty typowe przez Galaxy Isolation Forest, znajdują się w pierwszym regionie zanieczyszczenia zajętym przez galaktyki o niskim przesunięciu ku czerwieni, które mogą również zawierać AGN o niskiej aktywności. Z drugiej strony, kandydaci na AGN-y, którzy są zidentyfikowani jako obiekty odstające przez AGN Isolation Forest, znajdują się w regionie HzSFG. W ten sposób mamy spójne wyniki z dwóch oddzielnych metod, wykrywania obiektów odstających przez Isolation Forest i wizualizacji tSNE. Łącząc obie te metody, aby lepiej zrozumieć naturę katalogowych kandydatów na AGN, widzimy, że taka połączona metoda pozwala nam wykryć i odrzucić dwa możliwe źródła zanieczyszczeń. Jednym z nich są galaktyki o niskim przesunięciu ku czerwieni i ewentualne AGN-y z silnym składnikiem gospodarza. Drugim jest najbardziej problematyczna (albo, z punktu widzenia potencjału odkrywania rzadkich obiektów, najbardziej obiecująca) grupa HzSFG. Właściwości ograniczonego katalogu wynikowego, oczyszczonego z tych zanieczyszczeń (390 obiektów), jak również najczystszej wersji tego katalogu, w którym usunięto zarówno zanieczyszczenia galaktykami, jak i duże błędy fotometrycznego przesunięcia ku czerwieni (157 obiektów), są przedstawione w Tabeli 5.5.

5.6 Wyniki

Główne wyniki pracy przedstawionej w tej rozprawie można podzielić na dwie części. Metodologicznym rezultatem tych badań jest budowa wieloetapowej metody uczenia maszynowego dla fotometrycznej selekcji AGN-ów w katalogach wielozakresowych (ang. *multiwavelenght*) Dana metoda składa się z trzech komponentów, z których każdy okazał się być skuteczny w realizacji przynależnego mu zadania. Pierwszy komponent metody jest związany z przygotowaniem danych. Tutaj, oprócz tradycyjnych elementów uczenia maszynowego, takich jak selekcja cech, zastosowano dwie

	median	**MAD**	**min.**	**max.**
katalog bez zanieczyszczeń klasowych				
z_{phot}	1.356	0.353	0.059	2.841
g	22.674	1.077	18.768	26.024
r	22.128	0.904	18.888	24.529
i	21.661	0.76	18.532	23.819
z	21.376	0.677	18.515	23.428
Y	21.205	0.631	18.468	23.097
N2	19.974	0.413	17.858	20.846
N3	19.67	0.416	17.53	20.648
N4	19.487	0.402	17.041	20.546
katalog bez zanieczyszczeń klasowych i w przesunięciu ku czerwieni.				
z_{phot}	1.057	0.241	0.062	1.264
g	22.285	1.216	19.379	25.826
r	21.734	0.969	19.162	24.516
i	21.212	0.755	19.272	23.569
z	20.923	0.631	18.799	22.723
Y	20.748	0.583	18.784	22.417
N2	19.875	0.45	18.061	20.829
N3	19.665	0.473	17.854	20.648
N4	19.436	0.431	17.692	20.325

TABLICA 5.5: Statystyczne właściwości katalogu wynikowego oczyszczone z zanieczyszczeń klasowych oraz z łącznych zanieczyszczeń klasowych i w przesunięciu ku czerwieni. Przedstawione są wartości mediany, mediany odchylenia bezwzględnego (ang. *median absolute deviation*, MAD), minimalne i maksymalne wartości przesunięcia ku czerwieni i wielkości gwiazdowych w pasmach optycznych i NIR.

dodatkowe metody kluczowe dla opisywanej procedury. Jedną z nich była konstrukcja próbki treningowej AGN-ów, która opierała się na wstępnej selekcji obiektów w zakresie MIR i pozwalała na pośrednie dostarczenie informacji o selekcji MIR do struktury modelu ML podczas treningu. Drugą metodą było ograniczenie MCD próbki generalizacyjnej. Umożliwiła ona wybór kandydatów na AGN-y dokładnie w obszarze przestrzeni cech zdefiniowanym przez próbkę treningową.

Te dwa elementy umożliwiły skuteczność drugiego komponentu, tj. nadzorowanego modelu klasyfikacyjnego. Testując różne typy modeli klasyfikacyjnych z różnymi strategiami ważenia, w tym opartymi na logice rozmytej, można było wybrać próbkę najlepszych klasyfikatorów i stworzyć ostateczny klasyfikator z głosowaniem większościowym. Porównanie właściwości tego najlepszego klasyfikatora z metodą selekcji opartą na kolorach MIR na podpróbce danych treningowych z istniejącymi pomiarami MIR wykazało wyraźne podobieństwa między tymi dwiema metodami. Wskazuje to na to, że model był w stanie nauczyć się właściwości selekcji MIR poprzez próbkę treningową i dostosować je do danych optycznych i NIR. W ten sposób możliwe jest odtworzenie cech selekcji AGN-ów opartej na MIR dla próbki bez danych MIR, co wcześniej nie było możliwe.

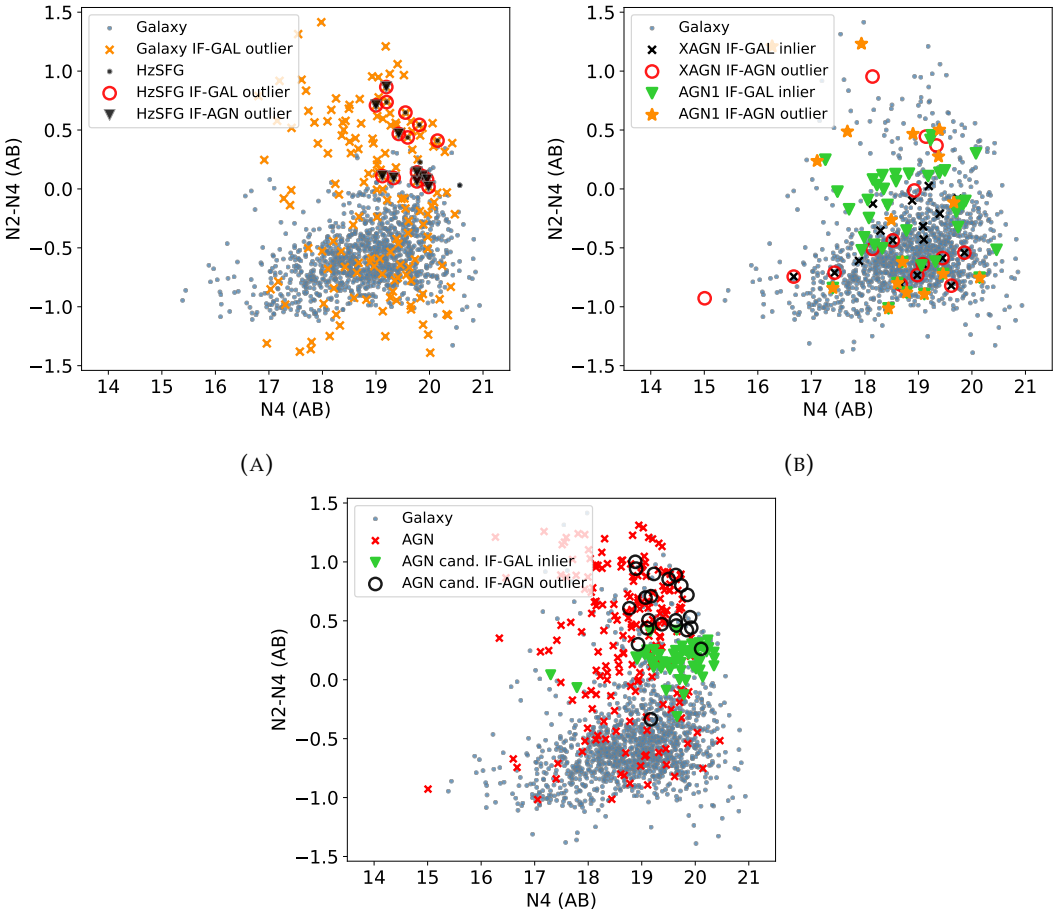
Trzeci komponent metody ML opierał się na metodach wykrywania obserwacji

odstających. Metody te zostały zastosowane do katalogu wynikowego, utworzonego w poprzednich krokach, w celu zwiększenia jego czystości. W tym przypadku zarówno metody wykrywania obserwacji odstających oparte na przesunięciach ku czerwieni, jak i metody oparte na klasach dały bardzo dobre wyniki. Metody oparte na klasach były w stanie wykryć dwa główne źródła zanieczyszczeń obecnych w katalogu AGN-ów. Jedną klasę obiektów odstających stanowiły obiekty znajdujące się pomiędzy klasami galaktyk i AGN-ów, które mogą być mało aktywnymi AGN-ami lub galaktykami pyłowymi o niskim przesunięciu ku czerwieni. Drugim rodzajem obiektów byli kandydaci na SFG o wysokim przesunięciu ku czerwieni, którzy są schowani wewnątrz obszaru w przestrzeni cech zajmowanego przez AGN-y i stanowią najbardziej problematyczne źródło zanieczyszczeń. Połączenie wykrywania obserwacji odstających za pomocą algorytmu Isolation Forest z wizualizacją tSNE pozwoliło nam zidentyfikować specyficzne właściwości SFG o dużym przesunięciu ku czerwieni, co umożliwiło wykrycie zestawu obiektów o podobnych właściwościach w katalogu kandydatów na AGN-y. Metoda wykrywania wartości odstających Isolation Forest okazała się również bardzo skuteczna w identyfikacji błędów nieprawidłowego oszacowania fotometrycznego przesunięcia ku czerwieni. Połączenie tych metod sprawia, że metoda jako całość jest bardzo skutecznym narzędziem do selekcji AGN-ów. Można ją łatwo zaadaptować do innych typów katalogów i danych. Fakt, że wszystkie elementy tej metody można zmodyfikować lub przekwalifikować i dostosować do innych potrzeb astrofizycznych, otwiera wiele możliwości jej zastosowania.

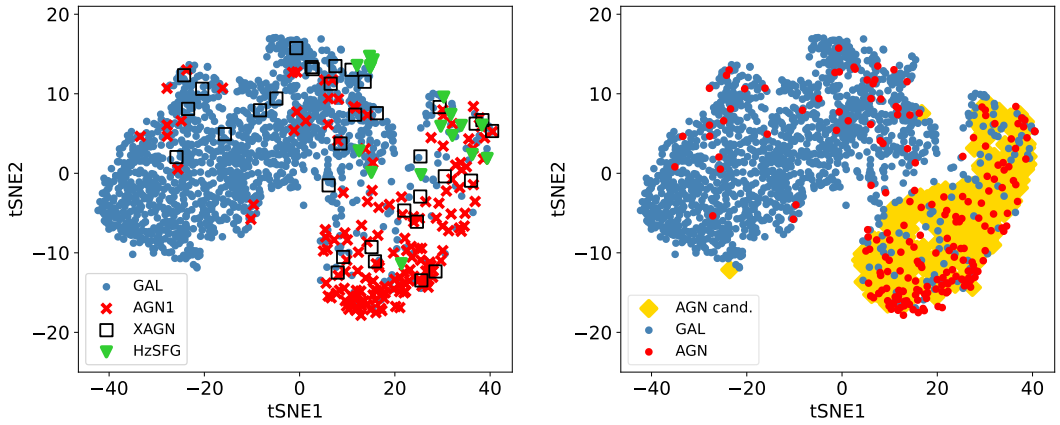
Drugim, fizycznym rezultatem tej pracy jest sam katalog AGN-ów w polu AKARI-NEP. Katalog główny składa się z 465 kandydatów na AGN-y; jego właściwości są podsumowane w Tabeli 5.2. Charakteryzuje się on 73% czystością i 64% kompletnością. Badanie właściwości MIR podpróbki kandydatów na AGN-y z detekcją w pasmie MIR pokazuje, że obiekty z tego katalogu wykazują właściwości typowe dla jasnych w MIR AGN-ów. Część próbki treningowej AGN-ów, która przyczyniła się do stworzenia katalogu wynikowego, składała się w większości z AGN-ów typu I. Można więc przyjąć rozsądne założenie, że większość obiektów obecnych w katalogu wynikowym to AGN-y typu I. Jak pokazano w Poliszczuk i in. (2021), widmowe rozkłady energii (SED) tych AGN-ów, uzyskane przy użyciu kodu CIGALE (Boquien i in., 2019) w większości potwierdzają, że mają one wysoką frakcję AGN.

Oprócz głównego katalogu kandydatów na AGN-y, utworzono trzy czyste katalogi podpróbek. Jeden z nich to katalog z usuniętymi dużymi błędami fotometrycznej estymacji przesunięcia ku czerwieni. Składa się on z 210 obiektów; jego właściwości są podsumowane w Tabeli 5.4. Ze względu na specyficzną metodę, która została zastosowana do fotometrycznej estymacji przesunięcia ku czerwieni w katalogu AKARI NEP-Wide, wartości tych przesunięć dla AGN-ów o dużym przesunięciu ku czerwieni są obarczone dużymi błędami. Ten podkatalog, pozbawiony niedokładnych oszacowań przesunięć ku czerwieni, jest stworzony do badań grupowania obiektów na małych przesunięciach ku czerwieni oraz badań środowiska w którym występują AGN-y w lokalnym Wszechświecie. Drugi podkatalog kandydatów na AGN został oczyszczony z zanieczyszczeń klasowych. Składa się on z 390 obiektów, a jego właściwości są przedstawione w Tabeli 5.5. Katalog ten dobrze nadaje się do wyboru celów do dalszych obserwacji spektroskopowych. Takie obserwacje oparte na przedstawionym katalogu wynikowym w rzeczywistości już trwają, a kolejne są planowane. Co więcej, obiekty zidentyfikowane jako zanieczyszczenia mogą również stanowić interesującą próbkę obiektów per se. Jedną interesującą podklasą zanieczyszczeń są kandydaci na SFG o wysokim przesunięciu ku czerwieni, a drugą - możliwe galaktyki o niskim przesunięciu ku czerwieni goszczące AGN o niskiej

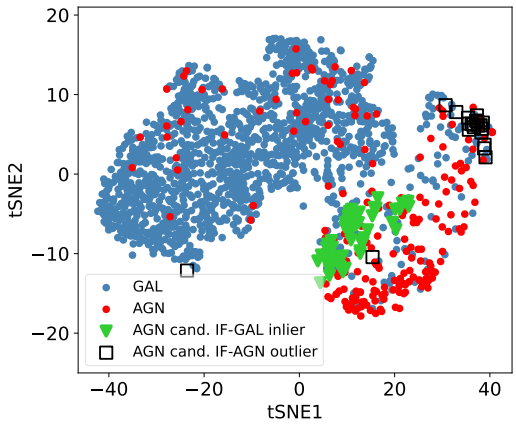
aktywności. Ostateczny podkatalog AGN-ów został uzyskany poprzez kombinację procedur czyszczenia katalogu przedstawionych powyżej. Składa się on z 157 obiektów. Jego właściwości są przedstawione w Tabeli 5.5. Katalog ten jest najczystszą uzyskaną próbką o zredukowanym zanieczyszczeniu klasowym i usuniętych błędach w fotometrycznych przesunięciach ku czerwieni. Jak już wspomniano w rozdziale 2, selekcja IR AGN sonduje wysoki koniec rozkładu współczynnika Eddingtona. Ta najczystsza próbka może być wykorzystana do dalszych badań tej populacji AGN-ów z danymi pochodzącymi z wielu zakresów długości fali, które są dostępne dla pola północnego bieguna ekliptycznego.



RYSUNEK 5.18: Wykres kolor-magnitudo $N2$–$N4$ vs $N4$ przedstawiający wyniki klasowej detekcji obserwacji odstających. *Panel A:* Predykcje wykonane na próbce galaktyk. Szare kropki - ogólny rozkład galaktyk, czarne kropki - SFG o wysokim przesunięciu ku czerwieni (HzSFG). Pomarańczowe krzyżyki - obiekty odstające dla Galaxy Isolation Forest. Czerwone kółka - HzSFG zaklasyfikowane jako obiekty odstające przez Galaxy Isolation Forest. Czarne trójkąty - HzSFG zaklasyfikowane jako obiekty odstające przez AGN Isolation Forest. *Panel B:* Predykcje wykonane na próbce AGN-ów. Czarne krzyżyki i czerwone kółka pokazują XAGN zidentyfikowane jako typowe obiekty przez Galaxy Isolation Forest i odstające przez AGN Isolation Forest. Zielone trójkąty i pomarańczowe gwiazdy przedstawiają AGN1, zidentyfikowane jako typowe obiekty przez Galaxy Isolation Forest i odstające przez AGN Isolation Forest. *Panel C:* Predykcje wykonane na katalogu wynikowym. Zielone trójkąty odnoszą się do obiektów zidentyfikowanych jako typowe przez Galaxy Isolation Forest. Czarne kółka odnoszą się do obiektów zidentyfikowanych jako odstające przez AGN Isolation Forest.



(A) (B)



(C)

RYSUNEK 5.19: Dwuwymiarowa wizualizacja tSNE danych treningowych, katalogu wynikowego i wyników wykrywania obserwacji odstających metodą Isolation Forest. *Panel A:* Wizualizacja tSNE danych treningowych składających się z galaktyk (niebieskie kropki), SFG o wysokim przesunięciu ku czerwieni (zielone trójkąty), AGN1 (czerwone krzyże) i XAGN (czarne kwadraty). *Panel B:* Wizualizacja tSNE danych galaktyk treningowych (niebieskie kropki) i AGN (czerwone kropki) w porównaniu z rozmieszczeniem katalogu wynikowego (żółte romby). *Panel C:* Wizualizacja tSNE danych galaktyk treningowych (niebieskie kropki) i AGN (czerwone kropki) w porównaniu z rozmieszczeniem kandydatów na AGN wybranych jako obiekty typowe (zielone trójkąty) i obiekty odstające (czarne kwadraty) odpowiednio przez modele Galaxy i AGN Isolation Forest.

6

Podsumowanie

W przedstawionej rozprawie udało się osiągnąć kilka znaczących wyników. W tym rozdziale Czytelnik znajdzie podsumowanie uzyskanych wniosków oraz opis możliwych przyszłych badań.

Po pierwsze, ograniczenie nałożone na próbkę generalizacyjną za pomocą algorytmu MCD okazało się być bardzo skuteczną techniką unikania ekstrapolacji. Analiza wyników klasyfikacji na próbkach treningowych i generalizacyjnych w oparciu o rozmieszczenie obiektów z próbce treningowej oraz kandydatów na AGN-y na wykresach kolorów NIR i MIR nie daje żadnych wskaźników istnienia znaczącej ekstrapolacji podczas generalizacji. Limit MCD nałożony na zbiór generalizacyjny wydaje się skutecznie zbliżać wyniki generalizacji do wyników predykcji modelu na zbiorze treningowym. Analiza wpływu ograniczenia MCD na właściwości nieoznakowanej próbki danych pokazuje, że dane ograniczenie wpływa najmocniej na pasma optyczne, odcinając końce rozkładu zawierające słabe obiekty. Efekt ten jest spowodowany wymaganiami doboru obiektów do pomiarów spektroskopowych, które zostały nałożone na próbkę treningową. Ponadto widoczna jest również znaczna redukcja obiektów charakteryzujących się kolorem $N2–N4 = 0$. Usunięcie tych obiektów przez algorytm MCD sugeruje, że nie były one dobrze reprezentowane w zbiorze treningowym i mogły powodować trudności podczas generalizacji. Jest to szczególnie ważne, ponieważ region $N2–N4 = 0$ został uznany podczas analizy metod wykrywania obserwacji odstających za problematyczny obszar z możliwym znacznym zanieczyszczeniem ze strony galaktyk o niskim przesunięciu ku czerwieni lub AGN-ów o niskiej aktywności. Skuteczność metody ograniczenia próbki generalizacjnej w oparciu o algorytm MCD sugeruje jego dużą przydatność w astronomii. Może być ona skutecznie stosowana w różnych przypadkach, gdy próbka treningowa może być niereprezentatywna dla nieoznakowanych danych. Są to w szczególności sytuacje, gdy model nadzorowany jest trenowany na danych zawierających klasy spektroskopowe lub przesunięcia ku czerwieni, jak również sytuacje, gdy model jest trenowany na danych uzyskanych z symulacji lub szablonów. Ponadto, ze względu na względną prostotę metody, powinno być możliwe odzyskanie funkcji selekcji utworzonej przez granicę MCD. Ten problem wymaga dalszych badań, aby jeszcze bardziej zwiększyć użyteczność tej metody.

Następnie badania zastosowania wag klasowych i wag poszczególnych obiektów (logiki rozmytej) na różnych rodzajach algorytmów uczenia maszynowego doprowadziły do kilku ważnych wyników. Po pierwsze, można zauważyć, że wpływ strategii ważenia na model jest silnie zależny od rodzaju algorytmu. Gdy niektóre modele są podatne na zastosowanie technik ważenia (algorytmy regresji logistycznej, SVM i

XGBoost), inne nie wykazują znaczących zmian w funkcjonowaniu (algorytmy lasów losowych i wyjątkowo losowych drzew). Po drugie, zastosowanie wag klasowych wykazało większy wpływ na działanie modelu niż wagi oparte na logice rozmytej. Ogólnie rzecz biorąc, zastosowanie wag klasowych przesuwa granicę separacji między klasami dalej od mniejszej klasy (w tym przypadku klasy AGN-ów), zwiększając kompletność katalogu AGN-ów i zmniejszając jego czystość. Tę tendencję wykazywała większość modeli z wyjątkiem algorytmów opartych na zespołach drzew decyzyjnych (las losowy i algorytm wyjątkowo losowych drzew). Takie zachowanie struktur drzewowych wydaje się być zachowaniem specyficznym dla wybranego zestawu danych. Logika rozmyta (lub ważenie związane z poszczególnymi obiektami) miała mniej bezpośredni wpływ na klasyfikację, spowodowany złożeniem kilku efektów powstałych w wyniku ważenia. Największa zmiana w działaniu modelu w przypadku zastosowania logiki rozmytej pochodzi od ważenia opartego na odległości od środka klasy. Co więcej, ten rodzaj logiki rozmytej działał najskuteczniej w modelach bez dodatkowych wag klasowych. W takim przypadku granica decyzyjna leży blisko mniejszej klasy, a waga oparta na odległości od środka klasy może wywołać efekt kumulacyjny, powodując wzrost kompletności katalogu AGN-ów. Modele z ważeniem klasowym mają granice decyzyjne odsunięte od mniejszej klasy AGN-ów, a wagi oparte na odległości działają przede wszystkim w celu zmniejszenia wpływu obserwacji odstających leżących w obszarze przestrzeni cech zajmowanym przez galaktyki. Co ciekawe, ten wzrost kompletności katalogu AGN-ów dla modeli bez ważenia klasowego nie zmniejsza czystości katalogu, jak to zaobserwowano przy zastosowaniu wag opartych na klasach. Wynik ten wynika z przeciwstawnych tendencji przy zastosowaniu wag opartych na odległości od środka klasy, powodujących wzrost czystości w jednym obszarze przestrzeni cech, i jej spadek w innym. Logika rozmyta oparta na błędach pomiarowych, pomimo swojej fizycznej motywacji, nie wykazała znaczącego wpływu na poprawność klasyfikacji zarówno w modelach zbalansowanych, jak i niezbalansowanych klasowo. Z tego powodu techniki logiki rozmytej należy stosować jako końcowe dopracowanie modelu po ustaleniu wyników otrzymanych za pomocą wag klasowych.

Porównanie selekcji AGN-ów opartej na kolorach MIR i selekcji opartej na technikach uczenia maszynowego potwierdza prawdziwość głównej tezy tej pracy, tj. założenia, że możliwe jest stworzenie i zastosowanie metody selekcji AGN-ów opartej na technikach uczenia maszynowego, która naśladuje właściwości metody wykorzystującej dane MIR, przy użyciu jedynie szerokopasmowej fotometrii z zakresu optycznego i NIR. Skuteczne działanie modelu wynika przede wszystkim z zastosowania dwóch kluczowych narzędzi jeszcze przed początkiem treningu modeli. Po pierwsze, próbka treningowa AGN-ów była oparta na próbce spektroskopowej, w której obiekty były wybrane na podstwie ich właściwości w zakresie MIR. W ten sposób informacje o selekcji AGN-ów opartej na danych MIR zostały wdrukowane w strukturę próbki treningowej i przekazane do konstrukcji modelu. Model efektywnie wykorzystał te informacje i był w stanie odzyskać próbkę AGN-ów dzięki specyficznemu wykładniczemu kształtowi widma w zakresie NIR-MIR. Ta informacja została przełożona na właściwości AGN-ów w przestrzeni kolorów AKARI NIR. Pokazuje to również, jak skuteczna jest kombinacja danych pochodzących z połączonych naziemnych obserwacji optycznych i pomiarów AKARI NIR. Drugim narzędziem, kluczowym dla tego celu, było wykorzystanie algorytmu MCD do ograniczenia rozkładu próbki generalizacyjnej do kształtu próbki treningowej. Ta właściwość dała możliwość efektywnego wytrenowania modelu i bezpiecznego zastosowania go do danych nieoznaczonych.

Eksperyment ekstrapolacji, który miał na celu dalsze zwiększenie skuteczności

klasyfikatora i przezwyciężenie ograniczeń selekcji MIR, wykazał niezadowalające wyniki. Nie udało się odzyskać obiektów znajdujących się w obszarze zajmowanym przez klase galaktyk. Były to przede wszystkim AGN-y wyselekcjonowane w zakresie rentgenowskim. Niemożność odzyskania próbki AGN-ów w zakresie rentgenowskim za pomocą technik opartych na ML i MIR potwierdza zasadnicze różnice między metodami selekcji opartymi na promieniowaniu rentgenowskim i na podczerwieni. Jednak brak sukcesu metody opartej na technikach uczenia maszynowego może być częściowo spowodowany niewielkim rozmiarem danych treningowych próbki AGN-ów użytych w eksperymencie ekstrapolacji. Ten problem może być rozwiązany w przyszłości poprzez trening modelu na dodatkowej próbce AGN-ów opartej na symulacjach. Połączenie informacji z takich danych z dodatkowym ograniczeniem MCD, który uwzględnia właściwości sztucznych danych, może działać podobnie jak w przypadku MIR i przełożyć część informacji z promieniowania rentgenowskiego do modelu.

Badania nad metodami wykrywania obserwacji odstających doprowadziły do obiecujących wyników. Okazało się, że można wykryć większość katastrofalnych błędów fotometrycznych estymacji przesunięcia ku czerwieni za pomocą algorytmu Isolation Forest. Takie podejście działa bardzo dobrze, o ile fotometryczne przesunięcia ku czerwieni znajdują się w zakresie spektroskopowych przesunięć ku czerwieni obecnych w próbce treningowej. Takie podejście pozwala na stworzenie katalogu nadającego się do dalszych badań nad grupowaniem galaktyk i jego dalszych zastosowań w kosmologii obserwacyjnej. Wykrywanie obserwacji odstających w kontekście klas obiektów zostało przeprowadzone poprzez połączenie metody Isolation Forest z wizualizacją za pomocą algorytmu tSNE. Takie połączenie wykazało, że algorytm Isolation Forest wytrenowany na próbce AGN-ów wykrywa obserwacje odstające o właściwościach podobnych do galaktyk gwiazdotwórczych (SFG) o wysokim przesunięciu ku czerwieni. Ten wynik jest szczególnie ważny, ponieważ SFG o wysokim przesunięciu ku czerwieni są głównym źródłem zanieczyszczenia katalogów AGN-ów selekcjonowanych w MIR. Z drugiej strony, Isolation Forest wytrenowany na próbce galaktyk ma tendencję do znajdowania obiektów będących normalnymi galaktykami w próbce kandydatów na AGN-y, które znajdują się w problematycznym regionie $N2–N4 \simeq 0$. Mogą to być galaktyki pyłowe o niskim przesunięciu ku czerwieni, jak również AGN o niskiej aktywności. Z drugiej strony, dany obszar w przestrzeni kolorów może być dodatkowo zajmowany przez specyficzną klasę AGN-ów o dużym przesunięciu ku czerwieni, które są słabo reprezentowane w danych treningowych. Usunięcie tych obserwacji odstających pozwala na uzyskanie katalogu kandydatów na AGN-y o wysokiej czystości, odpowiedniego do różnych badań ewolucyjnych i środowiskowych.

Podsumowując, w przedstawionej rozprawie opracowano skuteczną, złożoną procedurę selekcji AGN-ów opartą na metodach uczenia maszynowego. Pozwala ona przezwyciężyć ograniczenia instrumentalne teleskopów w zakresie średniej podczerwieni i znacząco zwiększyć rozmiar katalogu AGN-ów w porównaniu z rozmiarami danych, które są dostępne tradycyjnym technikom selekcji opartym na średniej podczerwieni. Ponadto, po uzyskaniu katalogu AGN-ów, można wykorzystać opracowane metody wykrywania obserwacji odstających, aby znaleźć odpowiedni dla planowanych zastosowań kompromis pomiędzy czystością i kompletnością katalogu. Można to zrobić poprzez kontrolę zanieczyszczenia różnych właściwości katalogu, takich jak dokładność fotometrycznego oszacowania przesunięcia ku czerwieni lub obecność błędnie sklasyfikowanych obiektów. Jak w przypadku większości metod uczenia maszynowego, dana metoda (lub poszczególne jej części) może być modyfikowana i stosowana do różnych zadań. Niektóre z tych modyfikacji mogą polegać

na przekwalifikowaniu klasyfikatora do pracy na innych klasach obiektów lub naśladowaniu innych właściwości obiektów (niż właściwości AGN-ów w MIR) spoza dostępnego zakresu widma. Innym rodzajem modyfikacji może być stworzenie wag logiki rozmytej odpowiednich do konkretnego zadania. Wreszcie elastyczność stworzonej metody wykrywania obserwacji odstających pozwala na skonstruowanie bardziej złożonych i subtelnych sposobów kontroli właściwości otrzymanego katalogu. Techniki przedstawione w tej pracy mogą znaleźć szerokie zastosowanie w nowoczesnej wielozakresowej astronomii *big data*, jak również w badaniach łączących symulacje z rzeczywistymi danymi jak ma to powszechnie miejsce w kosmologii obserwacyjnej.

Bibliografia

Allamandola, L. J., A. G. G. M. Tielens i J. R. Barker (mar. 1985). Polycyclic aromatic hydrocarbons and the unidentified infrared emission bands: auto exhaust along the milky way. 290, s. L25–L28. DOI: 10.1086/184435.

— (grud. 1989). Interstellar Polycyclic Aromatic Hydrocarbons: The Infrared Emission Bands, the Excitation/Emission Mechanism, and the Astrophysical Implications. 71, s. 733. DOI: 10.1086/191396.

Alonso-Herrero, A. i in. (mar. 2006a). Infrared Power-Law Galaxies in the Chandra Deep Field-South: Active Galactic Nuclei and Ultraluminous Infrared Galaxies. 640.1, s. 167–184. DOI: 10.1086/499800. arXiv: astro-ph/0511507 [astro-ph].

Alonso-Herrero, Almudena i in. (mar. 2001). The Nonstellar Infrared Continuum of Seyfert Galaxies. 121.3, s. 1369–1384. DOI: 10.1086/319410. arXiv: astro-ph/0012096 [astro-ph].

Alonso-Herrero, Almudena i in. (paź. 2006b). Near-Infrared and Star-forming Properties of Local Luminous Infrared Galaxies. 650.2, s. 835–849. DOI: 10.1086/506958. arXiv: astro-ph/0606186 [astro-ph].

Antonucci, Robert (sty. 1993). Unified models for active galactic nuclei and quasars. 31, s. 473–521. DOI: 10.1146/annurev.aa.31.090193.002353.

Arnouts, S. i in. (grud. 1999). Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North. 310.2, s. 540–556. DOI: 10.1046/j.1365-8711.1999.02978.x. arXiv: astro-ph/9902290 [astro-ph].

Ashby, Matthew, J. R. Houck i Perry B. Hacking (wrz. 1992). Deep Infrared Galaxies. 104, s. 980. DOI: 10.1086/116291.

Assef, R. J. i in. (lip. 2013). 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. 772.1, 26, s. 26. DOI: 10.1088/0004-637X/772/1/26. arXiv: 1209.6055 [astro-ph.CO].

Bañados, E. i in. (list. 2016). The Pan-STARRS1 Distant z > 5.6 Quasar Survey: More than 100 Quasars within the First Gyr of the Universe. 227.1, 11, s. 11. DOI: 10.3847/0067-0049/227/1/11. arXiv: 1608.03279 [astro-ph.GA].

Barden, S. C. i in. (1993). Hydra – KittPeak multi-object spectroscopic system. *ASPCS* 37, s. 185–202.

Barrufet, L. i in. (wrz. 2020). A high redshift population of galaxies at the North Ecliptic Pole. Unveiling the main sequence of dusty galaxies. 641, A129, A129. DOI: 10.1051/0004-6361/202037838. arXiv: 2007.07992 [astro-ph.GA].

Barrufet de Soto, Laia i in. (mar. 2017). The AGN Population in the Akari NEP Deep Field. *Publication of Korean Astronomical Society* 32.1, s. 271–273. DOI: 10.5303/PKAS.2017.32.1.271.

Beckert, T. i in. (sierp. 2008). Probing the dusty environment of the Seyfert 1 nucleus in NGC 3783 with MIDI/VLTI interferometry. 486.3, s. L17–L20. DOI: 10.1051/0004-6361:20078881. arXiv: 0806.0531 [astro-ph].

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Blanton, Michael R. i in. (wrz. 2003). The Broadband Optical Properties of Galaxies with Redshifts 0.02<z<0.22. 594.1, s. 186–207. DOI: 10.1086/375528. arXiv: astro-ph/0209479 [astro-ph].

Bock, J. i in. (sierp. 2013). The Cosmic Infrared Background Experiment (CIBER): The Wide-field Imagers. 207.2, 32, s. 32. DOI: 10.1088/0067-0049/207/2/32. arXiv: 1206.4702 [astro-ph.IM].

Boquien, M. i in. (lut. 2019). CIGALE: a python Code Investigating GALaxy Emission. 622, A103, A103. DOI: 10.1051/0004-6361/201834156. arXiv: 1811.03094 [astro-ph.GA].

Bosch, James i in. (sty. 2018). The Hyper Suprime-Cam software pipeline. 70, S5, S5. DOI: 10.1093/pasj/psx080. arXiv: 1705.06766 [astro-ph.IM].

Branchesi, M. i in. (sty. 2006). The radio luminosity function of the NEP distant cluster radio galaxies. 446.1, s. 97–111. DOI: 10.1051/0004-6361:20053767. arXiv: astro-ph/0509138 [astro-ph].

Breiman, L. i in. (1984). *Classification and Regression Trees*. Taylor & Francis. ISBN: 9780412048418. URL: https://books.google.pl/books?id=JwQx-WOmSyQC.

Breiman, Leo (paź. 2001). Random Forests. en. *Machine Learning* 45.1, s. 5–32. ISSN: 0885-6125, 1573-0565. DOI: 10.1023/A:1010933404324.

Brightman, Murray i Kirpal Nandra (lip. 2011). An XMM-Newton spectral survey of 12 μm selected galaxies - II. Implications for AGN selection and unification. 414.4, s. 3084–3104. DOI: 10.1111/j.1365-2966.2011.18612.x. arXiv: 1103.2181 [astro-ph.HE].

Brinchmann, J. i in. (lip. 2004). The physical properties of star-forming galaxies in the low-redshift Universe. 351.4, s. 1151–1179. DOI: 10.1111/j.1365-2966.2004.07881.x. arXiv: astro-ph/0311060 [astro-ph].

Brown, Michael J. I. i in. (sierp. 2008). Red Galaxy Growth and the Halo Occupation Distribution. 682.2, s. 937–963. DOI: 10.1086/589538. arXiv: 0804.2293 [astro-ph].

Bulbul, Esra i in. (paź. 2021). The eROSITA Final Equatorial-Depth Survey (eFEDS): Galaxy Clusters and Groups in Disguise. *arXiv e-prints*, arXiv:2110.09544, arXiv:2110.09544. arXiv: 2110.09544 [astro-ph.GA].

Burgarella, Denis i in. (sty. 2019). AKARI NEP field: Point source catalogs from GALEX and Herschel observations and selection of candidate lensed sub-millimeter galaxies. 71.1, 12, s. 12. DOI: 10.1093/pasj/psy134.

Cackett, Edward M., Misty C. Bentz i Erin Kara (czer. 2021). Reverberation mapping of active galactic nuclei: from X-ray corona to dusty torus. *iScience* 24.6, s. 102557. DOI: 10.1016/j.isci.2021.102557. arXiv: 2105.06926 [astro-ph.GA].

Cappelluti, N. i in. (kw. 2007). The soft X-ray cluster-AGN spatial cross-correlation function in the ROSAT-NEP survey. 465.1, s. 35–40. DOI: 10.1051/0004-6361:20065920. arXiv: astro-ph/0611553 [astro-ph].

Cepa, Jordi i in. (sierp. 2000). OSIRIS tunable imager and spectrograph. *Optical and IR Telescope Instrumentation and Detectors*. Red. Masanori Iye i Alan F. Moorwood. T. 4008. SPIE Conference Series, s. 623–631. DOI: 10.1117/12.395520.

Charlton, Paul J. L. i in. (maj 2019). Gemini Imaging of the Host Galaxies of Changing-look Quasars. 876.1, 75, s. 75. DOI: 10.3847/1538-4357/ab0ec1. arXiv: 1903.08122 [astro-ph.GA].

Chen, Bo Han i in. (mar. 2021). An active galactic nucleus recognition model based on deep neural network. 501.3, s. 3951–3961. DOI: 10.1093/mnras/staa3865. arXiv: 2101.06683 [astro-ph.GA].

Chen, Chao (2004). Using Random Forest to Learn Imbalanced Data.

Chen, Tianqi i Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Chiang, Chia-Ying i in. (kw. 2019). Does AGN fraction depend on redshift or luminosity? An extinction-free test by 18-band near- to mid-infrared SED fitting in the AKARI NEP wide field. 71.2, 31, s. 31. DOI: 10.1093/pasj/psz012. arXiv: 1902.02800 [astro-ph.GA].

Clarke, A. O. i in. (2020). Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra. *A&A* 639, A84. DOI: 10.1051/0004-6361/201936770. URL: https://doi.org/10.1051/0004-6361/201936770.

Clavel, J. i in. (maj 2000). 2.5-11 micron spectroscopy and imaging of AGNs. Implication for unification schemes. 357, s. 839–849. arXiv: astro-ph/0003298 [astro-ph].

Coil, Alison L. i in. (sty. 2008). The DEEP2 Galaxy Redshift Survey: Color and Luminosity Dependence of Galaxy Clustering at z ∼1. 672.1, s. 153–176. DOI: 10.1086/523639. arXiv: 0708.0004 [astro-ph].

Comastri, A. (sierp. 2004). Compton-Thick AGN: The Dark Side of the X-Ray Background. *Supermassive Black Holes in the Distant Universe*. Red. A. J. Barger. T. 308. Astrophysics and Space Science Library, s. 245. DOI: 10.1007/978-1-4020-2471-9_8. arXiv: astro-ph/0403693 [astro-ph].

Comastri, Andrea i Fabrizio Fiore (list. 2004). The Density and Masses of Obscured Black Holes. 294.1-2, s. 63–69. DOI: 10.1007/s10509-004-4023-5. arXiv: astro-ph/0404047 [astro-ph].

Cortes, C. i V. Vapnik (1995). Support-vector networks. *Machine Learning* 20, A39, s. 273–297. DOI: 10.1007/BF00994018.

Croton, Darren J. i in. (sty. 2006). The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. 365.1, s. 11–28. DOI: 10.1111/j.1365-2966.2005.09675.x. arXiv: astro-ph/0508046 [astro-ph].

D'Isanto, A. i K. L. Polsterer (sty. 2018). Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts. 609, A111, A111. DOI: 10.1051/0004-6361/201731326. arXiv: 1706.02467 [astro-ph.IM].

Dodd, Sierra A. i in. (sty. 2021). The Landscape of Galaxies Harboring Changing-look Active Galactic Nuclei in the Local Universe. 907.1, L21, s. L21. DOI: 10.3847/2041-8213/abd852. arXiv: 2010.10527 [astro-ph.GA].

Doi, Yasuo i in. (czer. 2015). The AKARI far-infrared all-sky survey maps. 67.3, 50, s. 50. DOI: 10.1093/pasj/psv022. arXiv: 1503.06421 [astro-ph.GA].

Donley, J. L. i in. (kw. 2012). Identifying Luminous Active Galactic Nuclei in Deep Surveys: Revised IRAC Selection Criteria. 748.2, 142, s. 142. DOI: 10.1088/0004-637X/748/2/142. arXiv: 1201.3899 [astro-ph.CO].

Elvis, Martin i in. (list. 1994). Atlas of Quasar Energy Distributions. 95, s. 1. DOI: 10.1086/192093.

Fabbiano, G. (wrz. 2006). Populations of X-Ray Sources in Galaxies. 44.1, s. 323–366. DOI: 10.1146/annurev.astro.44.051905.092519. arXiv: astro-ph/0511481 [astro-ph].

Faber, S. M. i in. (sierp. 2007). Galaxy Luminosity Functions to z∼1 from DEEP2 and COMBO-17: Implications for Red Galaxy Formation. 665.1, s. 265–294. DOI: 10.1086/519294. arXiv: astro-ph/0506044 [astro-ph].

Faber, Sandra M. i in. (mar. 2003). The DEIMOS spectrograph for the Keck II Telescope: integration and testing. *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*. Red. Masanori Iye i Alan F. M. Moorwood. T. 4841. SPIE Conference Series, s. 1657–1669. DOI: 10.1117/12.460346.

Fabian, A. C. (wrz. 2012). Observational Evidence of Active Galactic Nuclei Feedback. 50, s. 455–489. DOI: 10.1146/annurev-astro-081811-125521. arXiv: 1204.4114 [astro-ph.CO].

Fabricant, Daniel i in. (grud. 2005). Hectospec, the MMT's 300 Optical Fiber-Fed Spectrograph. 117.838, s. 1411–1434. DOI: 10.1086/497385. arXiv: astro-ph/0508554 [astro-ph].

Feltre, A. i in. (paź. 2012). Smooth and clumpy dust distributions in AGN: a direct comparison of two commonly explored infrared emission models. 426.1, s. 120–127. DOI: 10.1111/j.1365-2966.2012.21695.x. arXiv: 1207.2668 [astro-ph.CO].

Fernández, A. i in. (2018). *Learning from Imbalanced Data Sets*. Springer.

Fletcher, Roger (1987). *Practical Methods of Optimization*. Second. New York, NY, USA: John Wiley & Sons.

Fritz, J., A. Franceschini i E. Hatziminaoglou (mar. 2006). Revisiting the infrared spectra of active galactic nuclei with a new torus emission model. 366.3, s. 767–786. DOI: 10.1111/j.1365-2966.2006.09866.x. arXiv: astro-ph/0511428 [astro-ph].

Geach, J. E. i in. (lut. 2017). The SCUBA-2 Cosmology Legacy Survey: 850 μm maps, catalogues and number counts. 465.2, s. 1789–1806. DOI: 10.1093/mnras/stw2721. arXiv: 1607.03904 [astro-ph.GA].

Geurts, P., D. Ernst i Wehenkel L. (2006). Extremely randomized trees. *Machine Learning* 63, s. 42–63. DOI: 10.1007/s10994-006-6226-1.

Gilfanov, Marat i Andrea Merloni (wrz. 2014). Observational Appearance of Black Holes in X-Ray Binaries and AGN. 183.1-4, s. 121–148. DOI: 10.1007/s11214-014-0071-5.

Gilli, R. i in. (lut. 2005). The spatial clustering of X-ray selected AGN and galaxies in the Chandra Deep Field South and North. 430, s. 811–825. DOI: 10.1051/0004-6361:20041375. arXiv: astro-ph/0409759 [astro-ph].

Gioia, I. M. i in. (czer. 2001). Cluster Evolution in the ROSAT North Ecliptic Pole Survey. 553.2, s. L105–L108. DOI: 10.1086/320671. arXiv: astro-ph/0102332 [astro-ph].

Gioia, I. M. i in. (list. 2003). The ROSAT North Ecliptic Pole Survey: the Optical Identifications. 149.1, s. 29–51. DOI: 10.1086/378229. arXiv: astro-ph/0309788 [astro-ph].

Gioia, I. M. i in. (grud. 2004). RX J1821.6+6827: A cool cluster at z = 0.81 from the ROSAT NEP survey. 428, s. 867–875. DOI: 10.1051/0004-6361:20041426. arXiv: astro-ph/0408028 [astro-ph].

González-Martín, Omaira i in. (paź. 2019a). Exploring the Mid-infrared SEDs of Six AGN Dusty Torus Models. I. Synthetic Spectra. 884.1, 10, s. 10. DOI: 10.3847/1538-4357/ab3e6b. arXiv: 1908.11381 [astro-ph.GA].

— (paź. 2019b). Exploring the Mid-infrared SEDs of Six AGN Dusty Torus Models. II. The Data. 884.1, 11, s. 11. DOI: 10.3847/1538-4357/ab3e4f. arXiv: 1908.11389 [astro-ph.GA].

Gorjian, V. i in. (czer. 2008). The Mid-Infrared Properties of X-Ray Sources. 679.2, s. 1040–1046. DOI: 10.1086/587431. arXiv: 0803.0357 [astro-ph].

Goto, Tomotsugu i in. (mar. 2017). Hyper Suprime-Camera Survey of the Akari NEP Wide Field. *Publication of Korean Astronomical Society* 32.1, s. 225–230. DOI: 10.5303/PKAS.2017.32.1.225. arXiv: 1505.00012 [astro-ph.GA].

Goto, Tomotsugu i in. (kw. 2019). Infrared luminosity functions based on 18 mid-infrared bands: revealing cosmic star formation history with AKARI and Hyper Suprime-Cam*. 71.2, 30, s. 30. DOI: 10.1093/pasj/psz009. arXiv: 1902.02801 [astro-ph.GA].

Griffin, M. J. i in. (lip. 2010). The Herschel-SPIRE instrument and its in-flight performance. 518, L3, s. L3. DOI: 10.1051/0004-6361/201014519. arXiv: 1005.5123 [astro-ph.IM].

Gültekin, Kayhan i in. (czer. 2009). The M-σ and M-L Relations in Galactic Bulges, and Determinations of Their Intrinsic Scatter. 698.1, s. 198–221. DOI: 10.1088/0004-637X/698/1/198. arXiv: 0903.4897 [astro-ph.GA].

Haas, M., U. Klaas i S. Bianchi (kw. 2002). The relation of PAH strength with cold dust in galaxies. 385, s. L23–L26. DOI: 10.1051/0004-6361:20020222.

Hacking, P. B. i B. T. Soifer (lut. 1991). The Number Counts and Infrared Backgrounds from Infrared-bright Galaxies. 367, s. L49. DOI: 10.1086/185929.

Hacking, Perry, J. J. Condon i J. R. Houck (maj 1987). A Very Deep IRAS Survey: Constraints on the Evolution of Starburst Galaxies. 316, s. L15. DOI: 10.1086/184883.

Hacking, Perry i J. R. Houck (lut. 1987). A Very Deep IRAS Survey at L = 97 degrees , B = 30. 63, s. 311. DOI: 10.1086/191167.

Hacking, Perry i in. (kw. 1989). A Very Deep IRAS Survey. III. VLA Observations. 339, s. 12. DOI: 10.1086/167272.

Hao, Heng i in. (list. 2010). Hot-dust-poor Type 1 Active Galactic Nuclei in the COSMOS Survey. 724.1, s. L59–L63. DOI: 10.1088/2041-8205/724/1/L59. arXiv: 1009.3276 [astro-ph.CO].

Hao, Lei i in. (czer. 2005). The Detection of Silicate Emission from Quasars at 10 and 18 Microns. 625.2, s. L75–L78. DOI: 10.1086/431227. arXiv: astro-ph/0504423 [astro-ph].

Harris, Charles R. i in. (2020). Array programming with NumPy. *Nature* 585, 357–362. DOI: 10.1038/s41586-020-2649-2.

Harris, D. E. i Henric Krawczynski (wrz. 2006). X-Ray Emission from Extragalactic Jets. 44.1, s. 463–506. DOI: 10.1146/annurev.astro.44.051905.092446. arXiv: astro-ph/0607228 [astro-ph].

Harrison, Fiona A. i in. (czer. 2013). The Nuclear Spectroscopic Telescope Array (NuSTAR) High-energy X-Ray Mission. 770.2, 103, s. 103. DOI: 10.1088/0004-637X/770/2/103. arXiv: 1301.7307 [astro-ph.IM].

Hasinger, G. i in. (sty. 2021). The ROSAT Raster survey in the north ecliptic pole field. X-ray catalogue and optical identifications. 645, A95, A95. DOI: 10.1051/0004-6361/202039476. arXiv: 2011.04718 [astro-ph.CO].

Hastie, Trevor, Robert Tibshirani i Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hatziminaoglou, E. i in. (lip. 2010). HerMES: Far infrared properties of known AGN in the HerMES fields. 518, L33, s. L33. DOI: 10.1051/0004-6361/201014679. arXiv: 1005.2192 [astro-ph.CO].

Heckman, Timothy M. i Philip N. Best (sierp. 2014). The Coevolution of Galaxies and Supermassive Black Holes: Insights from Surveys of the Contemporary Universe. 52, s. 589–660. DOI: 10.1146/annurev-astro-081913-035722. arXiv: 1403.4620 [astro-ph.GA].

Henghes, Ben i in. (sierp. 2021). Benchmarking and scalability of machine-learning methods for photometric redshift estimation. 505.4, s. 4847–4856. DOI: 10.1093/mnras/stab1513. arXiv: 2104.01875 [astro-ph.IM].

Henry, J. P. i in. (czer. 2001). Overview of the ROSAT North Ecliptic Pole Survey. 553.2, s. L109–L113. DOI: 10.1086/320672.

Henry, J. Patrick i in. (lut. 2006). The ROSAT North Ecliptic Pole Survey: The X-Ray Catalog. 162.2, s. 304–328. DOI: 10.1086/498749. arXiv: astro-ph/0511195 [astro-ph].

Hickox, Ryan C. i in. (maj 2009). Host Galaxies, Clustering, Eddington Ratios, and Evolution of Radio, X-Ray, and Infrared-Selected AGNs. 696.1, s. 891–919. DOI: 10.1088/0004-637X/696/1/891. arXiv: 0901.4121 [astro-ph.GA].

Hinton, Geoffrey i Sam Roweis (2002). Stochastic Neighbor Embedding. *Proceedings of the 15th International Conference on Neural Information Processing Systems*. NIPS'02. Cambridge, MA, USA: MIT Press, 857–864.

Ho, L. C. (wrz. 2008). Nuclear activity in nearby galaxies. 46, s. 475–539. DOI: 10.1146/annurev.astro.45.051806.110546. arXiv: 0803.2268 [astro-ph].

Ho, Simon C. C. i in. (mar. 2021). Photometric redshifts in the North Ecliptic Pole Wide field based on a deep optical survey with Hyper Suprime-Cam. 502.1, s. 140–156. DOI: 10.1093/mnras/staa3549. arXiv: 2012.02421 [astro-ph.GA].

Holland, W. S. i in. (mar. 1999). SCUBA: a common-user submillimetre camera operating on the James Clerk Maxwell Telescope. 303.4, s. 659–672. DOI: 10.1046/j.1365-8711.1999.02111.x. arXiv: astro-ph/9809122 [astro-ph].

Horst, H. i in. (paź. 2006). The small dispersion of the mid IR - hard X-ray correlation in active galactic nuclei. 457.2, s. L17–L20. DOI: 10.1051/0004-6361:20065820. arXiv: astro-ph/0608358 [astro-ph].

Huang, Song i in. (sty. 2018). Characterization and photometric performance of the Hyper Suprime-Cam Software Pipeline. 70, S6, S6. DOI: 10.1093/pasj/psx126. arXiv: 1705.01599 [astro-ph.IM].

Huang, Ting-Chi i in. (paź. 2020). CFHT MegaPrime/MegaCam u-band source catalogue of the AKARI North Ecliptic Pole Wide field. 498.1, s. 609–620. DOI: 10.1093/mnras/staa2459. arXiv: 2008.05224 [astro-ph.GA].

Huang, Ting-Chi i in. (paź. 2021). Optically detected galaxy cluster candidates in the AKARI North Ecliptic Pole field based on photometric redshift from the Subaru Hyper Suprime-Cam. 506.4, s. 6063–6080. DOI: 10.1093/mnras/stab2128. arXiv: 2107.10010 [astro-ph.GA].

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* 9.3, s. 90–95. DOI: 10.1109/MCSE.2007.55.

Hwang, Narae i in. (paź. 2007). An Optical Source Catalog of the North Ecliptic Pole Region. 172.2, s. 583–598. DOI: 10.1086/519216. arXiv: 0704.1182 [astro-ph].

Ilbert, O. i in. (paź. 2006). Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. 457.3, s. 841–856. DOI: 10.1051/0004-6361:20065138. arXiv: astro-ph/0603217 [astro-ph].

Ishihara, D. i in. (maj 2010). The AKARI/IRC mid-infrared all-sky survey. 514, A1, A1. DOI: 10.1051/0004-6361/200913811. arXiv: 1003.0270 [astro-ph.IM].

Jaffe, W. i in. (maj 2004). The central dusty torus in the active nucleus of NGC 1068. 429.6987, s. 47–49. DOI: 10.1038/nature02531.

Jarrett, T. H. i in. (lip. 2011). The Spitzer-WISE Survey of the Ecliptic Poles. 735.2, 112, s. 112. DOI: 10.1088/0004-637X/735/2/112.

Jeon, Yiseul i in. (wrz. 2010). Optical Images and Source Catalog of AKARI North Ecliptic Pole Wide Survey Field. 190.1, s. 166–180. DOI: 10.1088/0067-0049/190/1/166. arXiv: 1010.3517 [astro-ph.CO].

John, T. L. (mar. 1988). Continuous absorption by the negative hydrogen ion reconsidered. 193.1-2, s. 189–192.

Jones, Mark H., Robert J. A. Lambourne i Stephen Serjeant (2015). *An Introduction to Galaxies and Cosmology*.

Jović, A., K. Brkić i N. Bogunović (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, s. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.

Kauffmann, Guinevere i in. (maj 2003a). The dependence of star formation history and internal structure on stellar mass for 10^5 low-redshift galaxies. 341.1, s. 54–69. DOI: 10.1046/j.1365-8711.2003.06292.x. arXiv: astro-ph/0205070 [astro-ph].

Kauffmann, Guinevere i in. (grud. 2003b). The host galaxies of active galactic nuclei. 346.4, s. 1055–1077. DOI: 10.1111/j.1365-2966.2003.07154.x. arXiv: astro-ph/0304239 [astro-ph].

Kawada, M. i in. (paź. 2007). The Far-Infrared Surveyor (FIS) for AKARI. 59, S389. DOI: 10.1093/pasj/59.sp2.S389. arXiv: 0708.3004 [astro-ph].

Kessler, M. F. i in. (list. 1996). The Infrared Space Observatory (ISO) mission. 500, s. 493–497.

Kim, Eunbin i in. (list. 2021a). The evolution of merger fraction of galaxies at z < 0.6 depending on the star formation mode in the AKARI NEP-Wide Field. 507.3, s. 3113–3124. DOI: 10.1093/mnras/stab2090. arXiv: 2108.07125 [astro-ph.GA].

Kim, S. J. i in. (grud. 2012). The North Ecliptic Pole Wide survey of AKARI: a near- and mid-infrared source catalog. 548, A29, A29. DOI: 10.1051/0004-6361/201219105. arXiv: 1208.5008 [astro-ph.CO].

Kim, Seong Jin i in. (grud. 2015). Mid-infrared luminosity function of local star-forming galaxies in the North Ecliptic Pole-Wide survey field of AKARI. 454.2, s. 1573–1584. DOI: 10.1093/mnras/stv2006. arXiv: 1509.04384 [astro-ph.GA].

Kim, Seong Jin i in. (sty. 2019). Characteristics of mid-infrared PAH emission from star-forming galaxies selected at 250 μm in the North Ecliptic Pole field. 71.1, 11, s. 11. DOI: 10.1093/pasj/psy121. arXiv: 1902.02883 [astro-ph.GA].

Kim, Seong Jin i in. (sty. 2021b). Identification of AKARI infrared sources by the Deep HSC Optical Survey: construction of a new band-merged catalogue in the North Ecliptic Pole Wide field. 500.3, s. 4078–4094. DOI: 10.1093/mnras/staa3359. arXiv: 2012.00750 [astro-ph.GA].

Kimura, Masahiko i in. (paź. 2010). Fibre Multi-Object Spectrograph (FMOS) for the Subaru Telescope. 62, s. 1135–1147. DOI: 10.1093/pasj/62.5.1135.

Klaas, U. i in. (grud. 2001). Infrared to millimetre photometry of ultra-luminous IR galaxies: New evidence favouring a 3-stage dust model. 379, s. 823–844. DOI: 10.1051/0004-6361:20011377. arXiv: astro-ph/0110213 [astro-ph].

Koenig, X. P. i in. (sty. 2012). Wide-field Infrared Survey Explorer Observations of the Evolution of Massive Star-forming Regions. 744.2, 130, s. 130. DOI: 10.1088/0004-637X/744/2/130.

Kormendy, John i Luis C. Ho (sierp. 2013). Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies. 51.1, s. 511–653. DOI: 10.1146/annurev-astro-082708-101811. arXiv: 1304.7762 [astro-ph.CO].

Krumpe, M. i in. (sty. 2015). Chandra survey in the AKARI North Ecliptic Pole Deep Field - I. X-ray data, point-like source catalogue, sensitivity maps, and number counts. 446.1, s. 911–931. DOI: 10.1093/mnras/stu2010. arXiv: 1409.7697 [astro-ph.HE].

LaMassa, Stephanie M. i in. (sty. 2015a). Discovery of the First Changing-Look Quasar. *American Astronomical Society Meeting Abstracts #225*. T. 225. American Astronomical Society Meeting Abstracts, 204.01, s. 204.01.

LaMassa, Stephanie M. i in. (lut. 2015b). The Discovery of the First "Changing Look" Quasar: New Insights Into the Physics and Phenomenology of Active Galactic Nucleus. 800.2, 144, s. 144. DOI: 10.1088/0004-637X/800/2/144. arXiv: 1412.2136 [astro-ph.GA].

Lee, H. M. i in. (paź. 2007). Nature of Infrared Sources in 11 μm Selected Sample from Early Data of the AKARI North Ecliptic Pole Deep Survey. 59, S529. DOI: 10.1093/pasj/59.sp2.S529. arXiv: 0705.1387 [astro-ph].

Lee, Hyung Mok i in. (lut. 2009). North Ecliptic Pole Wide Field Survey of AKARI: Survey Strategy and Data Characteristics. 61, s. 375. DOI: 10.1093/pasj/61.2.375. arXiv: 0901.3256 [astro-ph.GA].

Leger, A. i J. L. Puget (sierp. 1984). Identification of the "unidentified" IR emission features of interstellar dust ? 500, s. 279–282.

Li, Cheng i in. (grud. 2006). The clustering of narrow-line AGN in the local Universe. 373.2, s. 457–468. DOI: 10.1111/j.1365-2966.2006.11079.x. arXiv: astro-ph/0607492 [astro-ph].

Lilly, Simon J. i in. (sierp. 2013). Gas Regulation of Galaxies: The Evolution of the Cosmic Specific Star Formation Rate, the Metallicity-Mass-Star-formation Rate Relation, and the Stellar Content of Halos. 772.2, 119, s. 119. DOI: 10.1088/0004-637X/772/2/119. arXiv: 1303.5059 [astro-ph.CO].

Lin, Chun-Fu i Sheng-De Wang (2002). Fuzzy support vector machines. *IEEE transactions on neural networks* 13.2, s. 464–471.

Lira, Paulina i in. (lut. 2013). Modeling the Nuclear Infrared Spectral Energy Distribution of Type II Active Galactic Nuclei. 764.2, 159, s. 159. DOI: 10.1088/0004-637X/764/2/159. arXiv: 1301.7049 [astro-ph.CO].

Liu, Fei Tony, Kai Ming Ting i Zhi-Hua Zhou (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, s. 413–422. DOI: 10.1109/ICDM.2008.17.

Lopez-Rodriguez, E. i in. (sierp. 2018). The origin of the mid-infrared nuclear polarization of active galactic nuclei. 478.2, s. 2350–2358. DOI: 10.1093/mnras/sty1197. arXiv: 1805.01899 [astro-ph.GA].

Lusso, E. i G. Risaliti (mar. 2016). The Tight Relation between X-Ray and Ultraviolet Luminosity of Quasars. 819.2, 154, s. 154. DOI: 10.3847/0004-637X/819/2/154. arXiv: 1602.01090 [astro-ph.GA].

Lutz, D. i in. (paź. 2003). ISO spectroscopy of star formation and active nuclei in the luminous infrared galaxy <ASTROBJ>NGC 6240</ASTROBJ>. 409, s. 867–878. DOI: 10.1051/0004-6361:20031165. arXiv: astro-ph/0307552 [astro-ph].

Maaten, Laurens van der i Geoffrey Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, s. 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Magorrian, John i in. (czer. 1998). The Demography of Massive Dark Objects in Galaxy Centers. 115.6, s. 2285–2305. DOI: 10.1086/300353. arXiv: astro-ph/9708072 [astro-ph].

Maiolino, R. i in. (czer. 2007). Dust covering factor, silicate emission, and star formation in luminous QSOs. 468.3, s. 979–992. DOI: 10.1051/0004-6361:20077252. arXiv: 0704.1559 [astro-ph].

Mandelbaum, Rachel i in. (lut. 2009). Halo masses for optically selected and for radio-loud AGN from clustering and galaxy-galaxy lensing. 393.2, s. 377–392. DOI: 10.1111/j.1365-2966.2008.14235.x. arXiv: 0806.4089 [astro-ph].

Marconi, Alessandro i Leslie K. Hunt (maj 2003). The Relation between Black Hole Mass, Bulge Mass, and Near-Infrared Luminosity. 589.1, s. L21–L24. DOI: 10.1086/375804. arXiv: astro-ph/0304274 [astro-ph].

Marin, F. i in. (maj 2018). A near-infrared, optical, and ultraviolet polarimetric and timing investigation of complex equatorial dusty structures. 613, A30, A30. DOI: 10.1051/0004-6361/201732464. arXiv: 1801.08438 [astro-ph.HE].

Martin, D. Christopher i in. (sty. 2005). The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. 619.1, s. L1–L6. DOI: 10.1086/426387. arXiv: astro-ph/0411302 [astro-ph].

Martínez-Paredes, M. i in. (lut. 2020). Modeling the Strongest Silicate Emission Features of Local Type 1 AGNs. 890.2, 152, s. 152. DOI: 10.3847/1538-4357/ab6732. arXiv: 2001.00844 [astro-ph.GA].

Matsuhara, Hideo i in. (sierp. 2006). Deep Extragalactic Surveys around the Ecliptic Poles with AKARI (ASTRO-F). 58, s. 673–694. DOI: 10.1093/pasj/58.4.673. arXiv: astro-ph/0605589 [astro-ph].

McKinney, Wes (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. Red. Stéfan van der Walt i Jarrod Millman, s. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

Mendez, Alexander J. i in. (czer. 2013). PRIMUS: Infrared and X-Ray AGN Selection Techniques at 0.2 < z < 1.2. 770.1, 40, s. 40. DOI: 10.1088/0004-637X/770/1/40. arXiv: 1302.2920 [astro-ph.CO].

Meneux, B. i in. (czer. 2006). The VIMOS-VLT Deep Survey. The evolution of galaxy clustering per spectral type to z 1.5. 452.2, s. 387–395. DOI: 10.1051/0004-6361:20054571. arXiv: astro-ph/0511656 [astro-ph].

Merloni, A. i in. (lut. 2014a). The incidence of obscuration in active galactic nuclei. 437.4, s. 3550–3567. DOI: 10.1093/mnras/stt2149. arXiv: 1311.1305 [astro-ph.CO].

— (lut. 2014b). The incidence of obscuration in active galactic nuclei. 437.4, s. 3550–3567. DOI: 10.1093/mnras/stt2149. arXiv: 1311.1305 [astro-ph.CO].

Meusinger, H. i N. Balafkan (sierp. 2014). A large sample of Kohonen-selected SDSS quasars with weak emission lines: selection effects and statistical properties. 568, A114, A114. DOI: 10.1051/0004-6361/201423810. arXiv: 1407.0193 [astro-ph.GA].

Miyaji, Takamitsu i in. (wrz. 2007). The XMM-Newton Wide-Field Survey in the COSMOS Field. V. Angular Clustering of the X-Ray Point Sources. 172.1, s. 396–405. DOI: 10.1086/516579. arXiv: astro-ph/0612369 [astro-ph].

Miyazaki, Satoshi i in. (2012). Hyper Suprime-Cam. *Ground-based and Airborne Instrumentation for Astronomy IV*. Red. Ian S. McLean, Suzanne K. Ramsay i Hideki Takami. T. 8446. International Society for Optics i Photonics. SPIE, s. 327 –335. DOI: 10.1117/12.926844. URL: https://doi.org/10.1117/12.926844.

Mullaney, J. R. i in. (czer. 2011). Defining the intrinsic AGN infrared spectral energy distribution and measuring its contribution to the infrared output of composite galaxies. 414.2, s. 1082–1110. DOI: 10.1111/j.1365-2966.2011.18448.x. arXiv: 1102.1425 [astro-ph.CO].

Mullis, C. R. i in. (czer. 2001). The North Ecliptic Pole Supercluster. 553.2, s. L115–L118. DOI: 10.1086/320670. arXiv: astro-ph/0103202 [astro-ph].

Murakami, Hiroshi i in. (paź. 2007). The Infrared Astronomical Mission AKARI*. 59, S369–S376. DOI: 10.1093/pasj/59.sp2.S369. arXiv: 0708.1796 [astro-ph].

Narayan, Gautham i in. (maj 2018). Machine-learning-based Brokers for Real-time Classification of the LSST Alert Stream. 236.1, 9, s. 9. DOI: 10.3847/1538-4365/aab781. arXiv: 1801.07323 [astro-ph.IM].

Nayyeri, H. i in. (lut. 2018). Spitzer Observations of the North Ecliptic Pole. 234.2, 38, s. 38. DOI: 10.3847/1538-4365/aaa07e. arXiv: 1712.01290 [astro-ph.GA].

Nenkova, Maia, Željko Ivezić i Moshe Elitzur (maj 2002). Dust Emission from Active Galactic Nuclei. 570.1, s. L9–L12. DOI: 10.1086/340857. arXiv: astro-ph/0202405 [astro-ph].

Nenkova, Maia i in. (wrz. 2008a). AGN Dusty Tori. I. Handling of Clumpy Media. 685.1, s. 147–159. DOI: 10.1086/590482. arXiv: 0806.0511 [astro-ph].

Nenkova, Maia i in. (wrz. 2008b). AGN Dusty Tori. II. Observational Implications of Clumpiness. 685.1, s. 160–180. DOI: 10.1086/590483. arXiv: 0806.0512 [astro-ph].

Netzer, Hagai (sierp. 2015). Revisiting the Unified Model of Active Galactic Nuclei. 53, s. 365–408. DOI: 10.1146/annurev-astro-082214-122302. arXiv: 1505.00811 [astro-ph.GA].

Netzer, Hagai i in. (wrz. 2007). Spitzer Quasar and ULIRG Evolution Study (QUEST). II. The Spectral Energy Distributions of Palomar-Green Quasars. 666.2, s. 806–816. DOI: 10.1086/520716. arXiv: 0706.0818 [astro-ph].

Neugebauer, G. i in. (mar. 1984). The Infrared Astronomical Satellite (IRAS) mission. 278, s. L1–L6. DOI: 10.1086/184209.

Noeske, K. G. i in. (maj 2007). Star Formation in AEGIS Field Galaxies since z=1.1: The Dominance of Gradually Declining Star Formation, and the Main Sequence of Star-forming Galaxies. 660.1, s. L43–L46. DOI: 10.1086/517926. arXiv: astro-ph/0701924 [astro-ph].

Oi, Nagisa i in. (mar. 2017). Properties of Dust Obscured Galaxies in the Nep-Deep Field. *Publication of Korean Astronomical Society* 32.1, s. 245–249. DOI: 10.5303/PKAS.2017.32.1.245.

Oi, Nagisa i in. (sty. 2021). Subaru/HSC deep optical imaging of infrared sources in the AKARI North Ecliptic Pole-Wide field. 500.4, s. 5024–5042. DOI: 10.1093/mnras/staa3080.

Onaka, Takashi i in. (paź. 2004). The infrared camera (IRC) on board the ASTRO-F: laboratory tests and expected performance. *Optical, Infrared, and Millimeter Space Telescopes*. Red. John C. Mather. T. 5487. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, s. 338–349. DOI: 10.1117/12.550857.

Oyabu, S. i in. (maj 2011). AKARI detections of hot dust in luminous infrared galaxies. Search for dusty active galactic nuclei. 529, A122, A122. DOI: 10.1051/0004-6361/201014221.

Padovani, P. i in. (sierp. 2017). Active galactic nuclei: what's in a name? 25.1, 2, s. 2. DOI: 10.1007/s00159-017-0102-9. arXiv: 1707.07134 [astro-ph.GA].

Pan, ShuYang i in. (wrz. 2020). Cosmological parameter estimation from large-scale structure deep learning. *Science China Physics, Mechanics, and Astronomy* 63.11, 110412, s. 110412. DOI: 10.1007/s11433-020-1586-3. arXiv: 1908.10590 [astro-ph.CO].

Pearson, Chris i in. (mar. 2017). Herschel Observations in the Akari NEP Field: Initial Source Counts. *Publication of Korean Astronomical Society* 32.1, s. 219–223. DOI: 10.5303/PKAS.2017.32.1.219.

Pearson, Chris i in. (sty. 2019). The Herschel-PACS North Ecliptic Pole Survey. 71.1, 13, s. 13. DOI: 10.1093/pasj/psy107. arXiv: 1809.03990 [astro-ph.GA].

Pearson, W. J. i in. (lut. 2022). North Ecliptic Pole merging galaxy catalogue. *arXiv e-prints*, arXiv:2202.10780, arXiv:2202.10780. arXiv: 2202.10780 [astro-ph.GA].

Pedregosa, F. i in. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, s. 2825–2830.

Peeters, E., H. W. W. Spoon i A. G. G. M. Tielens (paź. 2004). Polycyclic Aromatic Hydrocarbons as a Tracer of Star Formation? 613.2, s. 986–1003. DOI: 10.1086/423237. arXiv: astro-ph/0406183 [astro-ph].

Peterson, Bradley M. (mar. 1993). Reverberation Mapping of Active Galactic Nuclei. 105, s. 247. DOI: 10.1086/133140.

— (1997). *An Introduction to Active Galactic Nuclei*.

Pier, Edward A. i Julian H. Krolik (grud. 1992). Infrared Spectra of Obscuring Dust Tori around Active Galactic Nuclei. I. Calculational Method and Basic Trends. 401, s. 99. DOI: 10.1086/172042.

Pilbratt, G. L. i in. (lip. 2010). Herschel Space Observatory. An ESA facility for far-infrared and submillimetre astronomy. 518, L1, s. L1. DOI: 10.1051/0004-6361/201014759. arXiv: 1005.5331 [astro-ph.IM].

Platt, John C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, s. 61–74.

Poglitsch, A. i in. (lip. 2010). The Photodetector Array Camera and Spectrometer (PACS) on the Herschel Space Observatory. 518, L2, s. L2. DOI: 10.1051/0004-6361/201014535. arXiv: 1005.1487 [astro-ph.IM].

Poliszczuk, Artem i in. (czer. 2019). Active galactic nucleus selection in the AKARI NEP-Deep field with the fuzzy support vector machine algorithm. 71.3, 65, s. 65. DOI: 10.1093/pasj/psz043. arXiv: 1902.04922 [astro-ph.IM].

Poliszczuk, Artem i in. (lip. 2021). Active galactic nuclei catalog from the AKARI NEP-Wide field. 651, A108, A108. DOI: 10.1051/0004-6361/202040219. arXiv: 2104.13428 [astro-ph.GA].

Pratt, Cameron T. i Joel N. Bregman (lut. 2020). SZ Scaling Relations of Galaxy Groups and Clusters Near the North Ecliptic Pole. 890.2, 156, s. 156. DOI: 10.3847/1538-4357/ab6e6c. arXiv: 2001.07802 [astro-ph.CO].

Predehl, P. i in. (mar. 2021). The eROSITA X-ray telescope on SRG. 647, A1, A1. DOI: 10.1051/0004-6361/202039313. arXiv: 2010.03477 [astro-ph.HE].

Probst, Philipp, Anne-Laure Boulesteix i Bernd Bischl (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *J. Mach. Learn. Res.* 20.1, 1934–1965. ISSN: 1532-4435.

Richards, Gordon T. i in. (czer. 2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. 123.6, s. 2945–2975. DOI: 10.1086/340187. arXiv: astro-ph/0202251 [astro-ph].

Richards, Gordon T. i in. (czer. 2006). The Sloan Digital Sky Survey Quasar Survey: Quasar Luminosity Function from Data Release 3. 131.6, s. 2766–2787. DOI: 10.1086/503559. arXiv: astro-ph/0601434 [astro-ph].

Richards, Gordon T. i in. (kw. 2009). Eight-Dimensional Mid-Infrared/Optical Bayesian Quasar Selection. 137.4, s. 3884–3899. DOI: 10.1088/0004-6256/137/4/3884. arXiv: 0810.3567 [astro-ph].

Rigopoulou, D. i in. (grud. 1999). A Large Mid-Infrared Spectroscopic and Near-Infrared Imaging Survey of Ultraluminous Infrared Galaxies: Their Nature and Evolution. 118.6, s. 2625–2645. DOI: 10.1086/301146. arXiv: astro-ph/9908300 [astro-ph].

Rousseeuw, Peter J. i Katrien Van Driessen (sierp. 1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41.3, 212–223. ISSN: 0040-1706. DOI: 10.2307/1270566. URL: https://doi.org/10.2307/1270566.

Sanders, D. B. i I. F. Mirabel (sty. 1996). Luminous Infrared Galaxies. 34, s. 749. DOI: 10.1146/annurev.astro.34.1.749.

Santos, Daryl Joe D. i in. (paź. 2021). Environmental effects on AGN activity via extinction-free mid-infrared census. 507.2, s. 3070–3088. DOI: 10.1093/mnras/stab2352. arXiv: 2108.06899 [astro-ph.GA].

Sawicki, Marcin (grud. 2002). The 1.6 Micron Bump as a Photometric Redshift Indicator. 124.6, s. 3050–3060. DOI: 10.1086/344682. arXiv: astro-ph/0209437 [astro-ph].

Schweitzer, M. i in. (maj 2008). Extended Silicate Dust Emission in Palomar-Green QSOs. 679.1, s. 101–117. DOI: 10.1086/587097. arXiv: 0801.4637 [astro-ph].

Sen, Snigdha i in. (lut. 2022). Astronomical big data processing using machine learning: A comprehensive review. *Experimental Astronomy* 53.1, s. 1–43. DOI: 10.1007/s10686-021-09827-4.

Seo, Hyunjong i in. (paź. 2019). Clustering of extremely red objects in the AKARI NEP-deep field. 71.5, 96, s. 96. DOI: 10.1093/pasj/psz079.

Shapiro, Stuart L. i Saul A. Teukolsky (1983). *Black holes, white dwarfs, and neutron stars : the physics of compact objects*.

Sheth, Ravi K. i Giuseppe Tormen (wrz. 1999). Large-scale bias and the peak background split. 308.1, s. 119–126. DOI: 10.1046/j.1365-8711.1999.02692.x. arXiv: astro-ph/9901122 [astro-ph].

Shi, Y. i in. (grud. 2006). 9.7 μm Silicate Features in Active Galactic Nuclei: New Insights into Unification Models. 653.1, s. 127–136. DOI: 10.1086/508737. arXiv: astro-ph/0608645 [astro-ph].

Shim, Hyunjin i in. (sierp. 2013). Hectospec and Hydra Spectra of Infrared Luminous Sources in the AKARI North Ecliptic Pole Survey Field. 207.2, 37, s. 37. DOI: 10.1088/0067-0049/207/2/37.

Shim, Hyunjin i in. (list. 2020). NEPSC2, the North Ecliptic Pole SCUBA-2 survey: 850-μm map and catalogue of 850-μm-selected sources over 2 deg^2. 498.4, s. 5065–5079. DOI: 10.1093/mnras/staa2621.

Sirocky, M. M. i in. (maj 2008). Silicates in Ultraluminous Infrared Galaxies. 678.2, s. 729–743. DOI: 10.1086/586727. arXiv: 0801.4776 [astro-ph].

Smith, J. D. T. i in. (lut. 2007). The Mid-Infrared Spectrum of Star-forming Galaxies: Global Properties of Polycyclic Aromatic Hydrocarbon Emission. 656.2, s. 770–791. DOI: 10.1086/510549. arXiv: astro-ph/0610913 [astro-ph].

Sobolewska, Malgorzata A., Aneta Siemiginowska i Piotr T. Zycki (czer. 2004). High-Redshift Radio-quiet Quasars: Exploring the Parameter Space of Accretion Models. I. Hot Semispherical Flow. 608.1, s. 80–94. DOI: 10.1086/392529. arXiv: astro-ph/0410204 [astro-ph].

Solarz, A. i in. (maj 2012). Star-galaxy separation in the AKARI NEP deep field. 541, A50, A50. DOI: 10.1051/0004-6361/201118108. arXiv: 1203.1931 [astro-ph.IM].

Solarz, A. i in. (paź. 2015). Clustering of the AKARI NEP deep field 24 μm selected galaxies. 582, A58, A58. DOI: 10.1051/0004-6361/201423370. arXiv: 1509.00219 [astro-ph.GA].

Spoon, H. W. W. i in. (sty. 2007). Mid-Infrared Galaxy Classification Based on Silicate Obscuration and PAH Equivalent Width. 654.1, s. L49–L52. DOI: 10.1086/511268. arXiv: astro-ph/0611918 [astro-ph].

Stalevski, Marko i in. (mar. 2012). 3D radiative transfer modelling of the dusty tori around active galactic nuclei as a clumpy two-phase medium. 420.4, s. 2756–2772. DOI: 10.1111/j.1365-2966.2011.19775.x. arXiv: 1109.1286 [astro-ph.CO].

Stern, Daniel i in. (wrz. 2005). Mid-Infrared Selection of Active Galaxies. 631.1, s. 163–168. DOI: 10.1086/432523. arXiv: astro-ph/0410523 [astro-ph].

Stern, Daniel i in. (lip. 2007). Mid-Infrared Selection of Brown Dwarfs and High-Redshift Quasars. 663.1, s. 677–685. DOI: 10.1086/516833. arXiv: astro-ph/0608603 [astro-ph].

Stern, Daniel i in. (lip. 2012). Mid-infrared Selection of Active Galactic Nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected Active

Galactic Nuclei in COSMOS. 753.1, 30, s. 30. DOI: 10.1088/0004-637X/753/1/30. arXiv: 1205.0811 [astro-ph.CO].

Stern, Daniel i in. (wrz. 2018). A Mid-IR Selected Changing-look Quasar and Physical Scenarios for Abrupt AGN Fading. 864.1, 27, s. 27. DOI: 10.3847/1538-4357/aac726. arXiv: 1805.06920 [astro-ph.GA].

Takagi, T. i in. (sty. 2012). The AKARI NEP-Deep survey: a mid-infrared source catalogue. 537, A24, A24. DOI: 10.1051/0004-6361/201117759. arXiv: 1201.0797 [astro-ph.CO].

team, The pandas development (lut. 2020). *pandas-dev/pandas: Pandas*. Wer. latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Thompson, G. D. i in. (maj 2009). Dust Emission from Unobscured Active Galactic Nuclei. 697.1, s. 182–193. DOI: 10.1088/0004-637X/697/1/182. arXiv: 0903.2422 [astro-ph.GA].

Trakhtenbrot, Benny i Hagai Netzer (grud. 2012). Black hole growth to z = 2 - I. Improved virial methods for measuring M_{BH} and L/L_{Edd}. 427.4, s. 3081–3102. DOI: 10.1111/j.1365-2966.2012.22056.x. arXiv: 1209.1096 [astro-ph.CO].

Truemper, J. (sty. 1982). The ROSAT mission. *Advances in Space Research* 2.4, s. 241–249. DOI: 10.1016/0273-1177(82)90070-9.

Trump, Jonathan R. i in. (list. 2009). The Nature of Optically Dull Active Galactic Nuclei in COSMOS. 706.1, s. 797–809. DOI: 10.1088/0004-637X/706/1/797. arXiv: 0910.2672 [astro-ph.CO].

Uchida, K. I., K. Sellgren i M. Werner (lut. 1998). Do the Infrared Emission Features Need Ultraviolet Excitation? 493.2, s. L109–L112. DOI: 10.1086/311136. arXiv: astro-ph/9711200 [astro-ph].

Urry, C. Megan i Paolo Padovani (wrz. 1995). Unified Schemes for Radio-Loud Active Galactic Nuclei. 107, s. 803. DOI: 10.1086/133630. arXiv: astro-ph/9506063 [astro-ph].

Vanden Berk, Daniel E. i in. (sierp. 2001). Composite Quasar Spectra from the Sloan Digital Sky Survey. 122.2, s. 549–564. DOI: 10.1086/321167. arXiv: astro-ph/0105231 [astro-ph].

Vestergaard, M. i in. (lut. 2008). Mass Functions of the Active Black Holes in Distant Quasars from the Sloan Digital Sky Survey Data Release 3. 674.1, s. L1. DOI: 10.1086/528981. arXiv: 0801.0243 [astro-ph].

Victoria-Ceballos, César Ivan i in. (lut. 2022). The Complex Infrared Dust Continuum Emission of NGC 1068: Ground-based N- and Q-band Spectroscopy and New Radiative Transfer Models. 926.2, 192, s. 192. DOI: 10.3847/1538-4357/ac441a. arXiv: 2201.11869 [astro-ph.GA].

Virtanen, Pauli i in. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, s. 261–272. DOI: 10.1038/s41592-019-0686-2.

Voges, W. i in. (czer. 2001). The ROSAT North Ecliptic Pole Survey X-Ray Data. 553.2, s. L119–L123. DOI: 10.1086/320673.

Wada, Takehiko i in. (grud. 2008). AKARI/IRC Deep Survey in the North Ecliptic Pole Region. 60, S517. DOI: 10.1093/pasj/60.sp2.S517.

Wake, David A. i in. (lip. 2008). The 2dF-SDSS LRG and QSO Survey: evolution of the clustering of luminous red galaxies since z = 0.6. 387.3, s. 1045–1062. DOI: 10.1111/j.1365-2966.2008.13333.x. arXiv: 0802.4288 [astro-ph].

Wang, Ting-Wen i in. (grud. 2020). Extinction-free Census of AGNs in the AKARI/IRC North Ecliptic Pole Field from 23-band infrared photometry from Space Telescopes. 499.3, s. 4068–4081. DOI: 10.1093/mnras/staa2988. arXiv: 2010.08225 [astro-ph.GA].

Waskom, Michael i the seaborn development team (wrz. 2020). *mwaskom/seaborn*. Wer. latest. DOI: 10.5281/zenodo.592845. URL: https://doi.org/10.5281/zenodo.592845.

Weisskopf, Martin C. i in. (lip. 2000). Chandra X-ray Observatory (CXO): overview. *X-Ray Optics, Instruments, and Missions III*. Red. Joachim E. Truemper i Bernd Aschenbach. T. 4012. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, s. 2–16. DOI: 10.1117/12.391545. arXiv: astro-ph/0004127 [astro-ph].

Werner, M. W. i in. (wrz. 2004). The Spitzer Space Telescope Mission. 154.1, s. 1–9. DOI: 10.1086/422992. arXiv: astro-ph/0406223 [astro-ph].

White, G. J. i in. (lip. 2010). A deep survey of the AKARI north ecliptic pole field . I. WSRT 20 cm radio survey description, observations and data reduction. 517, A54, A54. DOI: 10.1051/0004-6361/200913366. arXiv: 1006.0352 [astro-ph.CO].

Wilms, J., A. Allen i R. McCray (paź. 2000). On the Absorption of X-Rays in the Interstellar Medium. 542.2, s. 914–924. DOI: 10.1086/317016. arXiv: astro-ph/0008425 [astro-ph].

Wolter, A. i in. (grud. 2005). Unobscured QSO 2: a new class of objects? 444.1, s. 165–174. DOI: 10.1051/0004-6361:20053441. arXiv: astro-ph/0510045 [astro-ph].

Wright, Edward L. i in. (grud. 2010). The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. 140.6, s. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM].

Yang, G. i in. (sty. 2020). X-CIGALE: Fitting AGN/galaxy SEDs from X-ray to infrared. 491.1, s. 740–757. DOI: 10.1093/mnras/stz3001. arXiv: 2001.08263 [astro-ph.GA].

Yuan, Feng i Ramesh Narayan (sierp. 2014). Hot Accretion Flows Around Black Holes. 52, s. 529–588. DOI: 10.1146/annurev-astro-082812-141003. arXiv: 1401.0586 [astro-ph.HE].

Zakamska, Nadia L. i in. (list. 2003). Candidate Type II Quasars from the Sloan Digital Sky Survey. I. Selection and Optical Properties of a Sample at 0.3<Z<0.83. 126.5, s. 2125–2144. DOI: 10.1086/378610. arXiv: astro-ph/0309551 [astro-ph].

Zehavi, Idit i in. (czer. 2004). On Departures from a Power Law in the Galaxy Correlation Function. 608.1, s. 16–24. DOI: 10.1086/386535. arXiv: astro-ph/0301280 [astro-ph].

Zehavi, Idit i in. (wrz. 2005). The Luminosity and Color Dependence of the Galaxy Correlation Function. 630.1, s. 1–27. DOI: 10.1086/431891. arXiv: astro-ph/0408569 [astro-ph].

Zehavi, Idit i in. (lip. 2011). Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. 736.1, 59, s. 59. DOI: 10.1088/0004-637X/736/1/59. arXiv: 1005.2413 [astro-ph.CO].

Zemcov, Michael i in. (list. 2014). On the origin of near-infrared extragalactic background light anisotropy. *Science* 346.6210, s. 732–735. DOI: 10.1126/science.1258168. arXiv: 1411.1411 [astro-ph.CO].

Zhao, X. i in. (grud. 2021). The NuSTAR extragalactic survey of the James Webb Space Telescope North Ecliptic Pole time-domain field. 508.4, s. 5176–5195. DOI: 10.1093/mnras/stab2885. arXiv: 2109.13839 [astro-ph.HE].

A

Dodatek: Oprogramowanie

W tej pracy wykorzystano kilka bibliotek napisanych w języku Python 3. Kody algorytmów uczenia maszynowego oraz analiza otrzymanych wyników została wykonana za pomocą pakietów SciPy (Virtanen i in., 2020), NumPy (Harris i in., 2020), Pandas (McKinney, 2010; team, 2020), Scikit-learn (Pedregosa i in., 2011) i XGBoost (Chen i Guestrin, 2016). Wizualizację wyników wykonano za pomocą pakietów Matplotlib (Hunter, 2007) i Seaborn (Waskom i team, 2020). Większość kodów, dane treningowe, katalog AGN-ów, jak również dodatkowe dopasowania SED wykonane przez członków zespołu NEP można znaleźć na stronie GitHub: https://github.com/ArtemPoliszczuk/NEPWide_AGN

B

Dodatek: Wartości metryk

Niniejszy załącznik przedstawia szczegółowe wyniki metryczne uzyskane podczas treningu. Tabela B.1 i B.2 zawierają wyniki uzyskane w trakcie treningu głównej klasyfikacji. Tabela B.3 i B.4 przedstawiają wyniki z eksperymentu ekstrapolacyjnego. Przedstawione tabele pochodzą z pracy Poliszczuk i in. (2021).

Classifier	**F1**	**Precision**	**Recall**	**PR AUC**	**bACC**
dummy classifier	0.12	0.12	0.12	0.20	0.50
		Logistic regression:			
non-balanced normal	0.61±0.06	0.75±0.09	0.52±0.07	0.65±0.08	0.75±0.04
class-balanced normal	0.60±0.05	0.49±0.06	0.78±0.07	0.66±0.07	0.83±0.03
non-balanced fuzzy error	0.61±0.05	0.75±0.07	0.52±0.07	0.66±0.07	0.75±0.03
class-balanced fuzzy error	0.59±0.05	0.48±0.06	0.78±0.07	0.64±0.08	0.83±0.03
non-balanced fuzzy distance	0.64±0.05	0.72±0.07	0.57±0.06	0.65±0.07	0.77±0.03
class-balanced fuzzy distance	0.60±0.06	0.49±0.06	0.78±0.07	0.65±0.08	0.83±0.03
		SVM:			
non-balanced normal	0.65±0.06	0.75±0.06	0.58±0.08	0.61±0.08	0.77±0.04
class-balanced normal	0.67±0.05	0.63±0.07	0.73±0.06	0.65±0.07	0.83±0.03
non-balanced fuzzy error	0.63±0.06	0.74±0.07	0.56±0.08	0.61±0.08	0.76±0.04
class-balanced fuzzy error	0.68±0.05	0.64±0.06	0.74±0.07	0.65±0.08	0.84±0.03
non-balanced fuzzy distance	0.67±0.05	0.75±0.07	0.60±0.07	0.62±0.08	0.79±0.03
class-balanced fuzzy distance	0.66±0.05	0.60±0.06	0.73±0.05	0.64±0.07	0.83±0.03
		Random forest:			
non-balanced normal	0.66±0.06	0.72±0.08	0.61±0.07	0.65±0.09	0.79±0.03
class-balanced normal	0.64±0.06	0.74±0.07	0.57±0.08	0.65±0.08	0.77±0.04
non-balanced fuzzy error	0.66±0.05	0.72±0.06	0.62±0.07	0.65±0.07	0.79±0.03
class-balanced fuzzy error	0.64±0.06	0.74±0.08	0.57±0.07	0.66±0.08	0.77±0.04
non-balanced fuzzy distance	0.66±0.05	0.73±0.07	0.61±0.07	0.65±0.08	0.79±0.03
class-balanced fuzzy distance	0.64±0.06	0.74±0.09	0.57±0.06	0.65±0.08	0.77±0.03
		Extremely randomized trees:			
non-balanced normal	0.66±0.05	0.74±0.07	0.60±0.07	0.67±0.07	0.78±0.03
class-balanced normal	0.65±0.06	0.74±0.07	0.59±0.08	0.66±0.08	0.78±0.04
non-balanced fuzzy error	0.64±0.07	0.73±0.08	0.59±0.08	0.66±0.08	0.78±0.04
class-balanced fuzzy error	0.64±0.06	0.73±0.07	0.58±0.07	0.65±0.08	0.78±0.04
non-balanced fuzzy distance	0.66±0.06	0.75±0.07	0.60±0.07	0.66±0.08	0.79±0.04
class-balanced fuzzy distance	0.65±0.06	0.73±0.08	0.59±0.07	0.65±0.08	0.78±0.03

TABLICA B.1: Metryki dla klasyfikacji głównej. Cześć 1/2.

Classifier	**F1**	**Precision**	**Recall**	**PR AUC**	**bACC**
			XGBoost:		
non-balanced normal	0.67±0.06	0.74±0.07	0.62±0.08	0.68±0.08	0.79±0.04
class-balanced normal	0.68±0.06	0.66±0.08	0.69±0.06	0.67±0.08	0.82±0.03
non-balanced fuzzy error	0.66±0.06	0.74±0.08	0.60±0.06	0.67±0.07	0.78±0.03
class-balanced fuzzy error	0.68±0.06	0.66±0.07	0.70±0.08	0.67±0.08	0.82±0.04
non-balanced fuzzy distance	0.68±0.06	0.74±0.07	0.64±0.08	0.68±0.08	0.80±0.04
class-balanced fuzzy distance	0.68±0.05	0.65±0.07	0.72±0.06	0.66±0.08	0.83±0.03
			Voting schemes:		
stacked classifier	0.66±0.05	0.73±0.08	0.61±0.07	0.68±0.08	0.79±0.03
hard voter	0.68	0.73	0.64	—	0.80

TABLICA B.2: Metryki dla klasyfikacji głównej. Część 2/2.

Classifier	**F1**	**Precision**	**Recall**	**PR AUC**	**bACC**
dummy classifier	0.04	0.04	0.05	0.06	0.50
		Logistic regression:			
non-balanced normal	0.05±0.09	0.18±0.36	0.03±0.06	0.20±0.11	0.51±0.03
class-balanced normal	0.24±0.07	0.14±0.05	0.73±0.18	0.20±0.10	0.75±0.09
non-balanced fuzzy error	0.09±0.09	0.17±0.21	0.07±0.08	0.19±0.10	0.52±0.04
class-balanced fuzzy error	0.17±0.05	0.10±0.03	0.80±0.16	0.17±0.09	0.71±0.07
non-balanced fuzzy distance	0.07±0.12	0.19±0.32	0.05±0.08	0.17±0.09	0.52±0.04
class-balanced fuzzy distance	0.27±0.09	0.17±0.07	0.68±0.19	0.20±0.10	0.74±0.09
		SVM:			
non-balanced normal	0.02±0.06	0.07±0.23	0.01±0.04	0.11±0.07	0.50±0.02
class-balanced normal	0.25±0.08	0.16±0.06	0.67±0.16	0.17±0.09	0.73±0.08
non-balanced fuzzy error	0.06±0.11	0.11±0.19	0.05±0.09	0.14±0.08	0.52±0.05
class-balanced fuzzy error	0.20±0.08	0.12±0.05	0.55±0.18	0.20±0.13	0.67±0.09
non-balanced fuzzy distance	0.00±0.02	0.00±0.02	0.00±0.01	0.08±0.05	0.49±0.006
class-balanced fuzzy distance	0.24±0.07	0.16±0.05	0.55±0.15	0.15±0.07	0.69±0.07
		Random forest:			
non-balanced normal	0.01±0.05	0.02±0.08	0.01±0.04	0.18±0.09	0.50±0.02
class-balanced normal	0.08±0.14	0.23±0.39	0.05±0.09	0.23±0.15	0.52±0.05
non-balanced fuzzy error	0.02±0.07	0.03±0.12	0.01±0.07	0.20±0.11	0.50±0.04
class-balanced fuzzy error	0.08±0.13	0.23±0.36	0.05±0.08	0.24±0.12	0.52±0.04
non-balanced fuzzy distance	0.00±0.03	0.01±0.07	0.00±0.02	0.18±0.10	0.50±0.01
class-balanced fuzzy distance	0.11±0.13	0.32±0.40	0.07±0.09	0.25±0.12	0.53±0.04
		Extremely randomized trees:			
non-balanced normal	0.0±0.0	0.0±0.0	0.0±0.0	0.23±0.12	0.499±0.002
class-balanced normal	0.0±0.02	0.0±0.05	0.0±0.01	0.24±0.13	0.50±0.01
non-balanced fuzzy error	0.0±0.0	0.0±0.0	0.0±0.0	0.24±0.11	0.498±0.003
class-balanced fuzzy error	0.0±0.0	0.0±0.0	0.0±0.0	0.24±0.14	0.499±0.002
non-balanced fuzzy distance	0.0±0.0	0.0±0.0	0.0±0.0	0.24±0.13	0.499±0.002
class-balanced fuzzy distance	0.0±0.0	0.0±0.0	0.0±0.0	0.22±0.11	0.499±0.002

TABLICA B.3: Metriki dla eksperymentu ektrapolacyjnego. Część 1/2.

Classifier	**F1**	**Precision**	**Recall**	**PR AUC**	**bACC**
		XGBoost:			
non-balanced normal	0.06±0.11	0.16±0.31	0.04±0.08	0.20±0.11	0.52±0.04
class-balanced normal	0.26±0.11	0.23±0.11	0.33±0.15	0.23±0.11	0.63±0.07
non-balanced fuzzy error	0.07±0.11	0.15±0.25	0.05±0.08	0.17±0.09	0.52±0.04
class-balanced fuzzy error	0.23±0.11	0.18±0.10	0.34±0.16	0.20±0.10	0.63±0.08
non-balanced fuzzy distance	0.08±0.12	0.26±0.41	0.048±0.08	0.26±0.12	0.52±0.04
class-balanced fuzzy distance	0.29±0.12	0.25±0.11	0.37±0.16	0.24±0.11	0.65±0.08
		Voting Schemes:			
hard voter	0.26	0.16	0.59	—	0.71

TABLICA B.4: Metryki dla eksperymentu ekstrapolacyjnego. Część 2/2.