NATIONAL CENTRE FOR NUCLEAR RESEARCH

DOCTORAL THESIS

Application of machine learning methods in astrophysics and cosmology

Author:
Artem POLISZCZUK

Supervisor:
Agnieszka POLLO
Auxiliary supervisor:
Aleksandra Solarz

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Astrophysics Division
Fundamental Research Department

NATIONAL
CENTRE
FOR NUCLEAR
RESEARCH
ŚWIERK

May 3, 2022

Declaration of Authorship

I, Artem POLISZCZUK, declare that this thesis titled, “Application of machine learning methods in astrophysics and cosmology ” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the National Centre for Nuclear Research.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at the National Centre for Nuclear Research or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

NATIONAL CENTRE FOR NUCLEAR RESEARCH

Abstract

Application of machine learning methods in astrophysics and cosmology

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.

NATIONAL CENTRE FOR NUCLEAR RESEARCH

Streszczenie

Application of machine learning methods in astrophysics and cosmology

Artem POLISZCZUK

Celem przedstawionej pracy 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. Zastosowania takiej metody pozwolą stworzyć wysokiej jakości katalogi aktywnych jąder galaktyk do zastosowań w astronomii pozagalaktycznej i kosmologii obserwacyjnej. Badania przedstawione w pracy 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 nie stosowanych 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 reprezentowanego 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.

Acknowledgements

I would first like to thank my supervisors, Prof. Agnieszka Pollo and Dr. Aleksandra Solarz, who helped me to grow as a scientist and gave me continuous support and patience during my Ph.D. study.

I would also like to acknowledge my colleagues Dr. Katarzyna Małek, Prof. Matthew Malkan, Prof. Tomotsugu Goto, and Dr. Seong Jin Kim and members of the North Ecliptic Pole Team for insightful comments and valuable scientific discussions

Contents

List of Figures

List of Tables

To my Grandfather

1

Introduction

The rapid development of astrophysics and observational cosmology in the twenty-first century led to a major development of new statistical analysis methods in these fields. A significant part of this methodological revolution is connected to the quickly progressing field of big data and machine learning. Now, machine learning (ML) algorithms are applied with great success to various tasks in extragalactic astrophysics and cosmology. These are fundamental classification problems of catalog creation (see, e.g., Clarke, A. O. et al., 2020), estimation of astrophysical and cosmological parameters (see, e.g., D'Isanto and Polsterer, 2018; Henghes et al., 2021; Pan et al., 2020), or low-level ML-based telescope data processing pipelines for signal reconstruction and classification (see, e.g., Narayan et al., 2018), to name a few. The fast emergence of new algorithms and an exponentially growing amount of astronomical data suggest that ML methods are becoming an inherent part of modern astronomy (Sen et al., 2022).

The primary goal of this work was to create an ML-based method for photometric Active Galactic Nuclei (AGN) selection, which mimics the broadband photometric mid-IR selection method using only optical and near-IR broadband photometry. The importance of such a technique results from the trade-off that comes with the mid-IR AGN selection. The mid-IR range of the electromagnetic spectrum contains an essential part of information about the AGN emission, allowing one to obtain AGN catalogs characterized by both high purity and completeness (Padovani et al., 2017). Moreover, the mid-IR-based selection is sensitive to a specific stage of AGN accretion efficiency, which tights these objects to evolutionary processes of host galaxies and their placement in the cosmic web (Hickox et al., 2009). These valuable features of mid-IR AGN selection come with a significant limitation. Mainly, mid-IR telescopes' source catalogs are significantly smaller than optical and near-IR instruments. Two main factors cause this issue. One of them is the low resolution of mid-IR detectors. Second is the nature of mid-IR observations. This part of the electromagnetic spectrum is blocked by the Earth's atmosphere Kim et al., 2012. Thus mid-IR observations are possible only with space-based telescopes. However, it has to be artificially cooled down to not make a mid-IR detector blind by thermal emission from the telescope electronics. Moreover, such a cryogenic phase of the telescope work has limited operation time (see, e.g., Murakami et al., 2007). Thus, the combination of the low resolution of the instrument and limited operation time make obtaining large mid-IR catalogs very challenging.

The fundamental idea that allowed to overcome the problem of mid-IR selection was to indirectly impose mid-IR information into the structure of the classification model. Model constructed this way could effectively search for similar AGN candidates, exploiting the optical and near-IR data information. This approach overcomes the mid-IR detection condition and obtains AGN candidates catalog with similar properties but much larger size. The created method shown in this work is based on the research published in two works Poliszczuk et al. (2019) and Poliszczuk et al. (2021). These publications contain studies on various ML techniques applied to the IR data collected by the space-based AKARI telescope in the North Ecliptic Pole region. The first work (Poliszczuk et al., 2019) was a preliminary study on ML-based methods for the combined near-IR and mid-IR AGN selection. For this purpose, a specific ML algorithm called *support vector machine* was used to test different problems of classification. One of them was application of fuzzy logic to the classification algorithm structure, which allowed one to differentiate the impact of different objects in the training set on the classification based on their specific properties, such as measurement precision. To our best knowledge, this type of physically-motivated modification of the model via fuzzy logic was never used in astronomy before. Another was testing how the extrapolation outside of the region demarcated by the training data affects the classifier performance. These results were treated as a preliminary study and allowed to create a novel method for mimicking mid-IR selection using optical and near-IR data. This approach was described in the second publication (Poliszczuk et al., 2021), which is the basis of this work.

Compared to the mentioned publication, this work was modified and enriched. First, it contains an in-depth discussion of different fuzzy logic strategies and a comparison of their impact on classification with different types of supervised classification algorithms. This part was missing in the original work. It allows to better understand how this new approach can modify model performance and how does it affect the resulting catalog. The second substantial modification is the additional research on unsupervised outlier detection techniques. These methods are used to control AGN catalog properties better and remove different sources of contamination. This way obtained AGN catalog can match the needs of different applications. The part of developed machine learning pipeline which focuses on the outlier detection methods will be the main topic of the publication currently being prepared by the author (Poliszczuk et al., *in prep*).

This work is organized as follows. Chapter 2, contains a description of the Unified Model of AGN, physical processes behind AGN spectral energy distribution, how they affect different AGN selection methods, and some aspects of AGN connection to galaxy evolution and observational cosmology. Chapter 3 contains a description of data. The reader will find a general description of panchromatic data from the North Ecliptic Pole (NEP) in its first part. The second part contains a review of the multi-wavelength catalog of the AKARI NEP-Wide field, used to train ML models and obtain the final catalog of AGN candidates. Finally, this chapter contains information about the preparation and properties of training and generalization samples used in this work. In particular, Sec. 3.3.2 describes how to mitigate extrapolation risk during the prediction on the unlabeled data using the Minimum Covariance Determinant (MCD) algorithm. In Chapter 4 I review machine learning methods used in this work together with performance evaluation metrics, which were employed to train classifiers as well as to compare them to mid-IR selection. Chapter 5 contains a discussion of the obtained results. In particular, studies on the effectiveness of different ML algorithms on the impact of various class-based and fuzzy logic weighting strategies on classification. Furthermore, one will also find a discussion on the properties of the obtained AGN catalog and compare it to the catalog created with mid-IR selection methods. Finally, an additional experiment was done to test the possibility of boosting classification effectiveness even further to overcome problems present in both mid-IR selection and the ML-based method. The last part of this work shows different methods of outlier detection. These methods are used to find objects with catastrophic photometric redshift estimation errors and catalog contamination from the galaxy sample class. Chapter 6 presents a summary of obtained results. Appendix A contains information about the software used in the present work. In Appendix B reader will find additional data describing the performance evaluation.

2

Active Galactic Nuclei

2.1 Unified Model of Active Galactic Nuclei

Active galaxies are defined as galaxies during a period of intensive accretion of matter on their central supermassive black hole. The centers of the majority of galaxies are occupied by a supermassive black hole (SMBH). In active galaxies, material infalling onto SMBH radiates a large amount of energy in the X-ray and optical/UV part of the electromagnetic spectrum, which is then re-emitted at longer wavelengths in the close neighborhood of SMBH. This luminous region in the vicinity of SMBH is called *active galactic nucleus* (AGN). In this work, we will use the terms active galaxy and AGN interchangeably and refer to the galaxy hosting an active nucleus in the center as a *host*.

There were identified a large number of AGN classes, which differ in their relative emission strength in various parts of the electromagnetic spectrum as well as spectroscopic properties or presence of strong relativistic radio jets (for a comprehensive review, see, e.g., Padovani et al. (2017)). A large number of these properties can be explained in terms of the *unified model* of AGN. In its basic form described in Antonucci (1993), AGN classes were explained in terms of three parameters: inclination angle of AGN with respect to the line of sight (LOS), AGN luminosity, and covering factor of AGN. Visualization of the basic unified model is shown in Fig. 2.1

The structure of AGN is axisymmetric and divided into several main regions. The Central black hole is surrounded by a sub-pc accretion disc of fully ionized, dust-free matter. High-density clouds of ionized, dust-free gas form the *broad line region* (BLR) at the distance of $\simeq$1pc up to 10^{3-5} gravitational radii from the central BH. *Torus* placed outside of BLR is a mixture of gas and dust with a partially clumpy structure. The inner radius of the torus is roughly demarcated by the *sublimation radius*. This radius is a distance at which the temperature falls to $\sim$2000 K, which is the temperature above which dust grains start to evaporate (Netzer, 2015; Jones, Lambourne, and Serjeant, 2015). A region perpendicular to the torus plane forms an ionization cone, where low-ionization clouds of gas located at the distance of hundreds of parsecs from the torus plane are forming the *narrow-line region* (NLR). Additionally, a *relativistic jet* might occur along the central axis perpendicular to the torus plain. underlies the classification of aGNs as radio loud or quite Presence or lack of the jet underlies the classification of AGNs as radio loud or quiet. This leads to a complementary radio classification of AGNs which is beyond the scope of this work. The reader is encouraged to refer to other publications such as Urry and Padovani (1995) for the review.

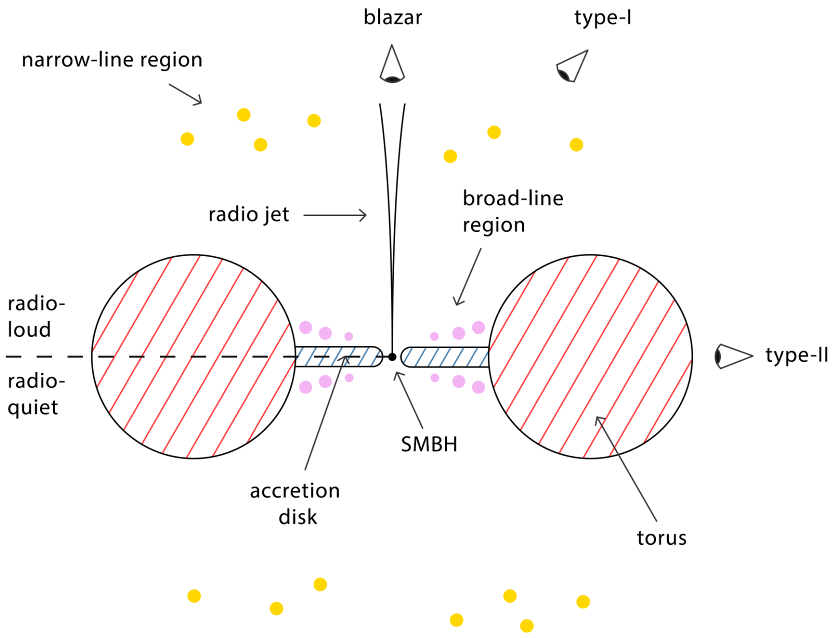


FIGURE 2.1: General picture of AGN Unified Model proposed in Antonucci (1993).

The inclination of the torus with respect to the LOS gives the primary division into two classes of AGN. The *type-I* AGN is characterized by the LOS facing ionization cone not obscured by the dusty torus. In this case, the UV-optical-NIR spectrum shows broad permitted, and semi-forbidden lines with typical gas velocities of 1000-20 000 km s^{-1} originated in BLR together with a bright, non-stellar central component. In addition, type-I AGN spectra (except for some high-luminosity sources) show forbidden narrow emission lines [1] which originated in the NLR with typical gas velocities of 300-1000 km s^{-1} are present. Despite referring to these lines as "narrow," they are still broad compared to emission lines that originated in the galaxies. Based on the source luminosity, such objects can be divided into *Seyfert type-I* galaxies, characterized by lower luminosity and *quasars*. In addition, a separate class of AGN with a relativistic jet pointing toward the observer is referred to as *blazars*. The *type-II* AGN occurs when the dusty torus blocks the LOS between observer and BLR. Such a situation makes emission from the BLR invisible to the observer. In this case, UV-optical-NIR spectra show only low-ionization narrow emission lines, which originate in the NLR outside the torus plane, without the broad line component. Type-II can be further divided into two sub-groups. The first one referred to as *hidden type-I* shows a broad line component obscured by the dusty torus in the polarized light. Hidden type-I AGNs can also be divided into *Seyfert type-II* galaxies and *type-II quasars* based on their luminosity. The second one, *true type-II* AGN, does not have any signatures of broad emission lines. These objects with typically lower luminosity compared to hidden type-I constitute approximately 30% of type-II AGN in the local Universe (see, e.g., Brightman and Nandra, 2011; Merloni et al., 2014a).

[1] Spectral emission lines in astronomy can divided with respect to the gas density at the place of their origin. Contrary to permitted lines, forbidden lines are produced in very low gas density regions. These lines cannot originate in denser environments because they are related to long-living excited states, and in high-density gas, these excited atoms are likely to be de-excited by collisions (Jones, Lambourne, and Serjeant, 2015).

In addition to these main classes there are objects with mixed properties of normal and active galaxies, such as host-dominated AGNs (Kauffmann et al., 2003b), low-ionization nuclear emission-line region or LINERS (Ho, 2008), low-luminosity optically dull AGNs (Trump et al., 2009) or weak line quasars (Meusinger and Balafkan, 2014). These objects are often interpreted as representations of intermediate evolutionary stages between active and normal galaxies or objects with an insufficient accretion flow near the SMBH. A full explanation of physical mechanisms causing the behavior of weak line AGNs is still an open question (see discussion in (Trump et al., 2009) and references therein). Another puzzling class of objects was recently discovered *changing look AGNs* (CL AGN, e.g., LaMassa et al., 2015a; Charlton et al., 2019), which can transition between properties of type-I and type-II AGNs. Discussions on possible explanations of CL AGN properties can be found in LaMassa et al. (2015b), Stern et al. (2018), and Dodd et al. (2021). Further discussion on the properties of these objects is beyond the scope of the present work, and readers are encouraged to refer to the cited sources.

Unified model issues such as difficulties with explaining properties of WLQ and CL AGN (Dodd et al., 2021), poor prediction on SMBH evolution and modeling of merging hosts or indication of much more complex structure of torus as well as the connection between torus interaction with a host galaxy show incompleteness of the basic AGN unification scheme. Discussion of unified model problems as well as possible ways to overcome them is presented in detail in Netzer (2015) and Heckman and Best (2014) and references therein.

Besides the AGN Unified Model, there is another way to interpret AGN properties and classify them based on the mechanism that dominates the energy outflow near the central engine. In order to better understand this distinction, we need to introduce a concept of *Eddington limit* (or *Eddington luminosity*). The Eddington limit defines the maximum luminosity, which an object can achieve in the state of hydrostatic equilibrium, i.e., when the radiation force acting outwards is balanced by the inward gravitational force (Peterson, 1997). Thus, when the radiation pressure exceeds the gravitational force at all radii, the gas surrounding the source will be blown away by the occurring outward winds. For a central black hole with mass M_{BH}, the Eddington limit is given by Shapiro and Teukolsky (1983)

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

Another essential object property tightly connected to the Eddington limit is the *Eddington ratio*, which is the ratio of object bolometric luminosity to its Eddington limit: L/L_{Edd}.

The classification based on the dominating mechanism of the energy outflow near the SMBH distinguishes two main AGN populations or modes. The first class of objects is referred to as the *radiative mode* (often called quasar mode or wind mode). This mode occurs in a luminous AGN, where the SMBH is surrounded by a geometrically thin and optically thick accretion disk. These objects are characterized by the efficient accretion flow and high Eddington ratios ($L/L_{Edd} > 0.01$). The second class is referred to as the *kinetic mode*, also called radio jet mode. These objects are associated with inefficient accretion at rates that do not exceed the 1% of the Eddington limit ($L/L_{Edd} < 0.01$). In the case of kinetic mode AGNs, the geometrically thin accretion disk is truncated in the inner regions. Another property of this mode is that most of the energetic output is translated through the large radio jets in bulk kinetic form (however, radiative-mode AGNs may also show jet structures). A visual comparison of the general structure of both modes is shown in Fig. 2.2. This work is primarily focused on the radiative mode AGNs. However, a further in-depth discussion on the nature of both modes and the implication of their presence on host galaxy properties is present in many reviews such as, e.g., Kormendy and Ho (2013) and Yuan and Narayan (2014). The mode-based classification of AGNs is crucial for understanding the connection between the AGN activity and galaxy evolution and the role AGN plays in observational cosmology. These topics are further discussed in Sec. 2.3.

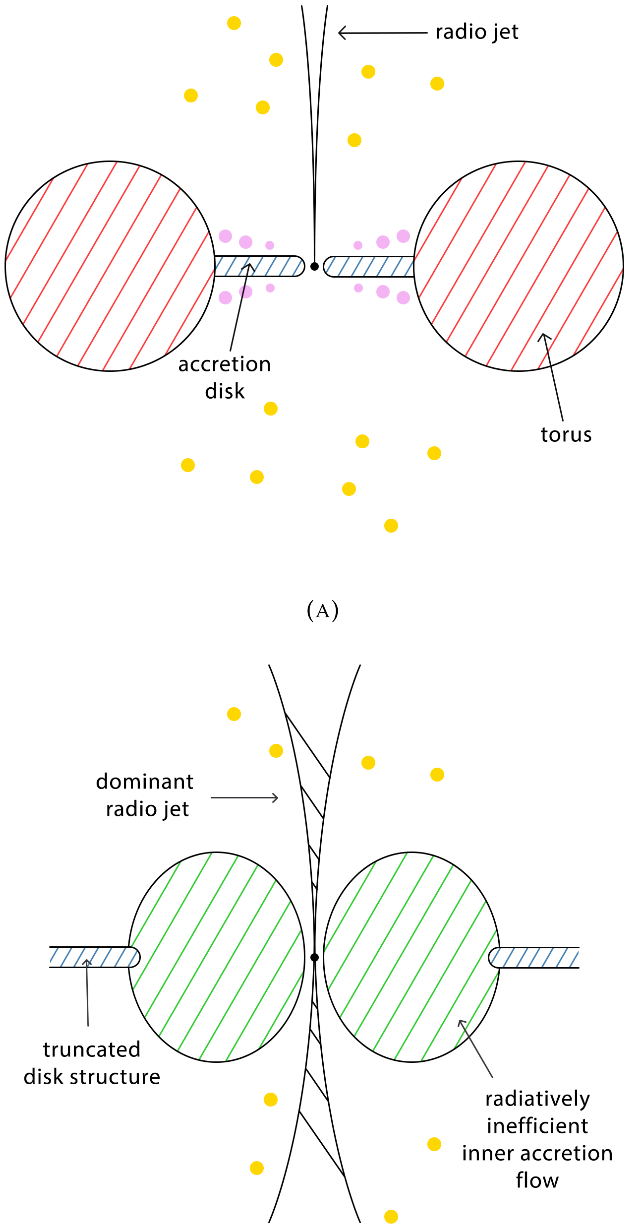
2.2 Multi-wavelength AGN emission and corresponding selection methods

Observations of AGNs in different parts of the electromagnetic spectrum highlight emission from specific parts of the AGN structure and are tightly connected to the broadband AGN selection properties. This section reviews the most important mechanisms producing AGN SED shape in the range from X-rays to IR and describes how they are used to recover AGN samples from sky surveys catalogs. This section does not cover γ-ray and radio properties of AGN since they are broad and complex topics not related to the scientific problem of this work. A large part of the description of AGN selection and bias effects in different parts of the spectrum was inspired by Padovani et al. (2017) and Donley et al. (2012) and references therein, where a reader can find a more profound analysis of these problems.

The **X-ray** band defined as 0.2–200 keV energy range gives one a well-established field for high completeness and purity AGN selection. The main contribution to the X-ray AGN flux comes from the inverse Compton scattering of the accretion-disk photons in the region of X-ray corona in the vicinity of SMBH (Gilfanov and Merloni, 2014). Additional X-ray flux may come from thermal emission inside the accretion-disk (Sobolewska, Siemiginowska, and Zycki, 2004) as well as relativistic jets (Harris and Krawczynski, 2006).

Despite additional mechanisms, AGN X-ray emission is strongly correlated with the accretion-disk emission and shows isotropic properties across the AGN population (Lusso and Risaliti, 2016). Universal properties of AGN X-ray emission combined with moderate resilience to the dust obscuration (especially in hard X-ray band) and low contamination from the host galaxy make X-ray studies well suited for AGN selection. The general approach to the AGN X-ray selection is to impose the X-ray luminosity (L_X) threshold onto the compact, point-like source catalog. In this case one can distinguish quasars with $L_X > 10^{44}$ erg s^{-1}, Seyfert galaxies placed in the range $L_X = 10^{42} - 10^{44}$ erg s^{-1} and low-luminosity AGNs $L_X < 10^{42}$ erg s^{-1} (Padovani et al., 2017). Possible contamination of AGN candidates catalog may come from high-redshift galaxy clusters or compact groups of galaxies when objects are probed at lower-energy X-ray bands, which are sensitive to the X-ray thermal radiation (Bulbul et al., 2021). The additional contaminant may occur in the case of low-luminosity sources, where X-ray emission from the host galaxy X-ray binaries may become a significant contribution to the overall X-ray flux (see, e.g. Fabbiano, 2006).

Several biases are present in X-ray AGN selection in which the main role plays energy-dependent X-ray absorption. Soft X-ray emission has higher absorption efficiency than hard X-ray emission (Wilms, Allen, and McCray, 2000). This phenomenon translates to the redshift bias, where high-redshift sources are probed at higher rest-frame energy and thus show a lower effect of absorbed X-ray flux. Additionally, a Compton-thick AGNs (CT AGN) class is defined as highly obscured sources in which material column density exceeds the cross-section value of the inverse Thomson



(A)

(B)

FIGURE 2.2: Classification based on the dominating mechanism of the energy outflow near the SMBH. *Panel a:* Radiative mode. The SMBH is surrounded by the geometrically thin and optically thick accretion disk. Radiative mode AGNs are characterized by the efficient accretion flow and high Eddington ratios. *Panel b:* Radio jet mode. These objects are characterized by the dominant radio jet and insufficient accretion of the material onto SMBH and low Eddington ratios.

scattering. These AGNs are especially difficult to detect at the X-ray range. They can escape purely X-ray-based selection (see, e.g. Comastri, 2004; Comastri and Fiore, 2004), and recovery of these objects often requires multi-wavelength studies. There is a good agreement between absorption in UV-optical passband and X-rays: the vast majority of type-I AGNs do not suffer from X-ray obscuration, while type-II objects are mainly recognized in X-ray bands as CT AGNs (see, e.g. Merloni et al., 2014b; Padovani et al., 2017).

Emission from the accretion-disk in the rest-frame **optical-UV** band with characteristic power-law continuum together with broad and narrow emission lines originated in BLR and NLR, respectfully, provides us a lot of information about the structure and kinematics of AGN. A major part of our picture of accretion-disk and broad- and narrow-line regions comes from spectroscopic studies (see, e.g. Elvis et al., 1994; Vanden Berk et al., 2001; Netzer, 2015 and references therein) as well as AGN reverberation mapping. In reverberation mapping, the spatial resolution is swapped with time resolution. Analyzing the time lag between optical and UV variability of the spectrum and response variability of broad emission lines allows one to reconstruct processes occurring in the neighborhood of SMBH. A detailed review of the reverberation mapping method can be found in, e.g. Peterson, 1993; Cackett, Bentz, and Kara, 2021. This section will only discuss the properties of broadband optical AGN selection.

Optical photometric identification of AGNs provides large catalog volumes at the expense of significant bias and contamination. Optically-selected AGN catalogs probe mainly type-I sources with high Eddington ratio ($L/L_{Edd} > 0.01$) (Vestergaard et al., 2008; Trakhtenbrot and Netzer, 2012). The AGN appearance in broad optical passbands is sensitive to the power-law continuum and broad emission lines. If observed in the rest-frame, these properties give AGN a specific location in optical color space. However, the color appearance becomes stellar-like at certain redshift ranges, making them hard to separate (especially quasars, which are point objects). This phenomenon introduces large stellar contamination, especially at low galactic latitudes. These problematic redshift ranges are placed around $z \sim 2.6$ and $z \sim 3.5$ (Richards et al., 2002; Richards et al., 2006; Padovani et al., 2017). Another noticeable decrease in the completeness of optical AGN catalogs comes from low sensitivity to type-II objects (see e.g. Zakamska et al., 2003).

Now let us analyze more deeply **infrared** properties of AGN since they are the most important for this work. The infrared part of the spectrum can be divided into three main parts: near-infrared (NIR, 1–5 μm), mid-infrared [2] (MIR, 5–50 μm) and far-infrared (FIR, 50–500 μm). The IR radiation of AGN occurs primarily in the range of NIR and MIR due to the reprocessing of emission from the accretion disk by the silicate dust located in the torus. This dust produces NIR-MIR continuum with distinctive power-law shape $F_\nu \propto \nu^\alpha$ in the range $3 - 8\mu$m, with $\alpha < 0$, being a characteristic of AGN (Klaas et al., 2001; Alonso-Herrero et al., 2001; Alonso-Herrero et al., 2006a). In particular Alonso-Herrero et al. (2006a) found AGN-dominated galaxies exhibit $-2.8 < \alpha < -0.5$ in the range $3.6 - 8\mu$m. In addition, the silicate torus dust also produces two important spectral features located at 9.7μm and 18μm. These features originate from the SiO stretching and bending modes respectfully (Thompson et al., 2009). Other prominent spectral features present in MIR AGN spectra, which may be important for AGN selection, are strong high-excitation lines [NeV] at 14.3μm and [OIV] at 25.9μm. They are connected to the ionization by AGN continuum, most likely originated in NLR (Lutz et al., 2003) and can have a significant contribution to the broadband MIR flux.

Studies of IR AGN emission established fundamentals of the unified model (e.g., research on IR emission together with optical polarization confirmed the nature of

[2] In some publications, one can find an alternative division with NIR covering a range of 1–3 μm and MIR, 3–50 μm respectfully.

hidden type-I sources as AGNs with torus located at the LOS of the observer (Tran, Miller, and Kay, 1992; Smith et al., 2002)), as well as allowed to study the distribution of gas and dust in the torus (see, e.g., Thompson et al., 2009; Sirocky et al., 2008; Marin et al., 2018; Lopez-Rodriguez et al., 2018). There exist three main groups of torus material distribution models, which are trying to fit observed AGN multiwavelength SEDs: continuous distribution models (Pier and Krolik, 1992; Fritz, Franceschini, and Hatziminaoglou, 2006), clumpy torus models (Nenkova, Ivezić, and Elitzur, 2002; Nenkova et al., 2008a) and composite models which are trying to combine continuous and clumpy distributions (Stalevski et al., 2012). Comparison of these models can be found in Feltre et al. (2012) and Lira et al. (2013) (see also detailed discussion in Netzer, 2015).

Several observational results strongly oppose the continuous distribution model. If the material in the torus were distributed smoothly, it would produce a strong temperature gradient along the torus radius and a deep absorption feature since flux emitted by the inner region of the torus would be absorbed by the dust material located at a larger distance from the center. Thus the view of a hot inner region of the torus should produce strong silicon emission at 9.7μm and smaller 18μm features. In contrast, analysis of the outer torus regions should demonstrate absorption in these features. Consequently, for a smooth torus model, a MIR picture of type-I AGN, where the inner region of the torus is visible, should produce silicon features in emission. On the other hand, in type-II AGN, where an observer can see only the outer region - MIR silicon features will be present in absorption. However, observations show much weaker 9.7μm absorption than predicted by continuous distribution model, indicating clumpy structure of central obscurer (Shi et al., 2006; Nenkova et al., 2008b; Martínez-Paredes et al., 2020). In addition, a comparison of the 9.7μm line with the second silicon feature at 18μm allows one to make even better discrimination between two dust distribution models and identify the dust composition of the torus (Hao et al., 2005; Thompson et al., 2009) (however, see discussion on the impact of the different chemical composition of models on silicate features predictions in (Sirocky et al., 2008)). Secondly, strongly isotropic MIR emission of AGNs (see, e.g., Horst et al., 2006) contradicts smooth torus models, which predict much brighter type-I AGNs for a fixed intrinsic luminosity (Thompson et al., 2009). Finally, an extensive range of dust temperatures observed at a specific distance from the central engine (Jaffe et al., 2004; Beckert et al., 2008) is not possible in the case of a smooth torus, where a strong temperature gradient should be present. Therefore, only a clumpy structure, where gaps between clumps allow the central engine to heat more distant material, can produce observed behavior.

As suggested in Netzer (2015), probably a more realistic torus structure is described by the two-phase composite model, where the space between clumps is filled with diluted dusty gas causing additional attenuation of incident radiation. One of the reasons why diluted medium should be present in torus is the inevitable collisions of dusty clouds. However, despite the observational clues, there is still an ongoing debate on the structure of the torus (see, e.g., González-Martín et al., 2019a; González-Martín et al., 2019b; Victoria-Ceballos et al., 2022). In addition to the above discussion on torus geometry, the dust component also seems to be present on the NLR's outskirts. This dusty structure may strongly contribute to the 10–30μm continuum emission (Schweitzer et al., 2008).

The IR selection of AGNs is focused on signatures of dust emission in NIR and MIR passbands. The dust emission gives AGNs characteristic red colors at NIR and shorter MIR passbands, distinguishing them from stars and normal galaxies. [3] The FIR emission is rarely used for AGN selection since it is primarily caused by the star-forming activity of the host (Netzer et al., 2007; Hatziminaoglou et al., 2010; Mullaney et al., 2011). Obscuration of the central region does not limit the selection capabilities of IR broadband methods severely as it does in optical selection. Moreover, it also has an advantage over X-ray selection in terms of much faster survey speed resulting in larger catalog volumes (Padovani et al., 2017; Gorjian et al., 2008). The NIR-MIR color AGN selection was applied in all major IR telescopes covering that part of spectral range, such as Spitzer (Werner et al., 2004; Stern et al., 2005), WISE (Wright et al., 2010; Stern et al., 2012; Assef et al., 2013) or AKARI (Murakami et al., 2007; Lee et al., 2007; Oyabu et al., 2011).

There are, however, several sources of contamination and selection effects present in the IR selection. The primary contamination at lower galactic latitudes comes from brown dwarfs (Stern et al., 2007) and young stellar objects (Koenig et al., 2012), which can mimic AGN IR colors at specific redshifts. The problem of proper identification of these objects becomes less important at a larger angular distance from the Galactic plane, where the main source of the catalog contamination comes from star-forming galaxies (SFG). To better understand the nature of SFG misclassification, we need to know how the stellar component contributes to the IR SED of the galaxy.

At the shorter NIR wavelengths, crucial role is played by 1.6μm spectral feature called *stellar bump* (see, e.g, Padovani et al., 2017; Stern et al., 2005; Alonso-Herrero et al., 2006b; Donley et al., 2012). It is caused by the opacity of the H^- ion, which has a minimum at the corresponding wavelength range (John, 1988). This feature is characteristic of the atmospheres of cool stars and is an important SED element of almost all stellar populations. However, this feature is missing in the very young hot stars ($\sim$1 Myr). In their case, the SED takes the form of the Rayleigh-Jeans power-law shape. It might be also less prominent in low-metallicity stars since the production of H^- ions relies on the presence of free electrons (John, 1988). Another phenomenon that might occur is a shift of the apparent wavelength of stellar bump towards the red end caused by the ISM dust absorption of shorter wavelength component (see, e.g., Sawicki, 2002 for an in-depth discussion of 1.6 μm stellar bump properties). This 1.6μm stellar bump can strongly affect AGN selection. Especially in the case of SFG at higher redshifts ($z \geq 1$), 1.6μm stellar bump enters longer NIR wavelengths characterized by a strong decrease of accretion-disk emission as well as low emission from the dusty torus. As the result, 1.6μm may mimic presence of power-law shape AGN SED and give high-z SFG AGN-like colors (Stern et al., 2005; Donley et al., 2012; Assef et al., 2013).

The SED of non-active SFG shows a characteristic dip between 1.6μm stellar bump and FIR (and longer wavelength MIR) emission from dust heated by star-forming component. In the case of AGN, this dip is filled by a power-law shape thermal continuum produced by emission from dusty torus (Donley et al., 2012). While the host component becomes less prominent in the MIR and longer wavelength NIR range, a significant contribution may come from several dust emission features connected to the ISM presence. The most prominent are spectral features connected to emission from large *polycyclic-aromatic-hydrocarbon* molecules containing $\sim$50 carbon atoms (PAHs) (Leger and Puget, 1984; Allamandola, Tielens, and Barker, 1985; Allamandola, Tielens, and Barker, 1989). The PAH emission comes from reprocessed UV (and to some degree optical) light (Uchida, Sellgren, and Werner, 1998) and, as such, is often

[3] Another strength of NIR-MIR AGN selection is based on the ability to find the most distant, high-redshift quasars, which are undetectable in the optical passbands due to the Lyα absorption (Bañados et al., 2016).

used as a tracer of star formation. There was a long discussion on how accurately PAH features may estimate star-forming components. Analysis of MIR spectroscopic data from the Spitzer telescope as well as earlier studies of Infrared Space Observatory (ISO; Kessler et al., 1996) showed that PAH emission does not trace the massive star formation as marked by the presence of O stars (Haas, Klaas, and Bianchi, 2002), but rather it traces the presence of B stars population, which make the main contribution to the stellar component of the galaxy (Peeters, Spoon, and Tielens, 2004; Smith et al., 2007; Maiolino et al., 2007).

Main PAH features are located in NIR and MIR ranges and are centered at 3.3μm, 6.2μm, 7.7μm, 8.6μm, and 11.3μm. Emission from PAH bands alone can contribute up to 20% of infrared luminosity (L_{IR}) in the case of galaxies with star-forming solid component (Smith et al., 2007). Thus, PAH emission may significantly impact the distribution of SFG in the IR color space. It can also hide the presence of AGN in PAH-dominated spectra from the color-based selection.

The most prominent feature located at 7.7μm also played a major role in establishing observational evidence for AGN unified model. In Clavel et al. (2000) authors compared MIR spectra of type-I and type-II Seyfert galaxies. By analyzing the 7.7μm PAH feature properties, they showed that host properties are unrelated to the AGN type. They also showed similarities in IR SED properties between Seyfert-II galaxies and starburst galaxies, except for the NIR-MIR range dominated by torus emission (Peeters, Spoon, and Tielens, 2004).

The PAH emission, together with IR AGN properties, was also used to study the nature of ultra-luminous infrared galaxies (ULIRGs, $L_{IR} > 10^{12} L_{\odot}$). The ULIRG population becomes an important group of sources when analyzing AGN selection and the impact of star formation in IR data. Morphological studies of these objects showed that most of them are complex systems of ongoing mergers Sanders and Mirabel, 1996. From the combined analysis of PAH emission, where the PAH spectrum of ULIRGs is very similar to obscured starburst galaxy and silicate absorption features, it was found that galaxies with the domination of both AGN and star-forming component can appear as ULIRGs (Spoon et al., 2007). Also, no significant evidence for classical model predictions, where more advanced mergers are AGN-dominated, was found (Rigopoulou et al., 1999).

Regarding biases generated by IR AGN selection, one must consider several important mechanisms. One crucial bias comes from the change of AGN IR colors in the case of strong, broad emission Hα, which may fall into the short-wavelength NIR band and reduce the object's red color. This problem occurs at the redshift range $3.5 \leq z \leq 5$ and can be mitigated by using a combination of IR and optical photometry for AGN selection (Richards et al., 2009). Another source of bias comes from the identification escape of AGNs with a low ratio of AGN dust to host dust emission (Hao et al., 2010). IR color selection methods might not capture these objects due to the lack of the main distinguishing component - AGN dust emission in MIR. Another, more subtle selection bias is coming from the host-AGN interplay. In particular a major role is played by a correlation between the luminosity of the spheroidal component of the host galaxy and AGN luminosity (Marconi and Hunt, 2003). This relation translates into the dependence of specific AGN to host luminosity ratio $L_{\nu,AGN}/L_{\nu,host}$ on the Eddington ratio and, as a consequence, creates a bias against low Eddington ratio AGNs in IR catalogs ratio (Padovani et al., 2017). It makes the most severe limitation of IR selection techniques. According to Hickox et al. (2009) and Mendez et al. (2013) IR selection might probe only few percent upper limit of X-ray AGN catalogs L/L_{Edd} distribution. This effect also translates into bias in M_{BH} distribution, i.e., a large number of spiral galaxies with prominent disk component and lower M_{BH} will be missing from IR catalogs (Padovani et al., 2017; Magorrian et al., 1998).

When analyzing catalog properties and selection effectiveness, one should consider the depth of the survey. While MIR AGN selection is regarded as a very effective approach to AGN selection, it shows different properties for shallow and deep surveys. As it was shown in Padovani et al. (2017) and Assef et al. (2013) in the case of shallow surveys, MIR AGN selection can recover the AGN catalog characterized by both high purity and high completeness. However, in the case of deep surveys, a strong completeness-purity trade-off exists.

2.3 AGN Feedback and Its Connection to Observational Cosmology

Studies of statistical properties of AGN catalogs give us important insights into galaxy evolution and the impact of AGN activity on the processes occurring in the host galaxy, as well as the connection between AGN distribution and Large Scale Structure of the Universe. The phenomenon of AGN affecting the star formation in the host galaxy is known as *AGN feedback*. This process occurs via an interaction between radiation from the matter accretion onto the SMBH and the bulge gas component. The outward radiation pressure sweeps galaxy bulge out of ISM gas, terminating the star formation in the galaxy's central region. Simultaneously, the lack of gas in the bulge limits the material needed for the accretion process, ending the activity of the central engine. In particular, the relation between M_{BH} and stellar velocity dispersion in the host bulge (Gültekin et al., 2009) is often treated as an important case of indirect observations of AGN feedback caused by the radiative-mode AGNs. Another piece of evidence comes from the analysis of the brightest cluster galaxies (BCG) mass. Without an energy input coming from kinetic AGN feedback, they would appear even more massive as giant starburst galaxies. See Fabian, 2012 for in-depth discussion on both of these evidences. Thus, there are separate fundamental roles for the two AGN modes. The AGN in the radiative mode tends to sweep gas out of the bulge. On the other hand, the kinetic-mode AGN keeps the gas hot enough, so it does not cool and does not reinforce star formation processes in the vicinity of SMBH.

Many studies on the AGN feedback suggest that AGN activity may play a major role in the bimodal nature of the galaxy population and quenching of the star formation in blue galaxies in particular (see, e.g., Croton et al., 2006b; Croton et al., 2006a). A bimodality of galaxy population (see, e.g., Kauffmann et al., 2003a; Blanton et al., 2003) defines two main groups of galaxies. One group referred to as a blue population or blue cloud, consists of late Hubble type galaxies with significant ongoing star formation, small stellar masses, and tight relation between star formation rate and stellar mass (Brinchmann et al., 2004). The second group referred to as a red population or a red sequence, comprises early Hubble type galaxies with little ongoing star formation and significantly larger stellar mass. A simplified scenario of galaxy evolution (Lilly et al., 2013; Hickox et al., 2009) may be described as follows . First, the galaxy evolves along the blue star-forming main sequence, accommodating mass through accretion of IGM and, to a smaller degree, through mergers. After reaching a critical mass, both accretion of cold gas and the star formation become quenched. This step determines the galaxy entering the red population. Thus, according to the current paradigm, the AGN feedback plays a major role in this process, keeping galaxies entering the red population from further star formation. However, research on AGN feedback is a relatively new field, and further research should be done to understand this process better. The key to obtaining a complete picture of the interaction between the AGN activity and galaxy evolution lies in a subtle analysis of large astronomical data sets. For this purpose, a study of host galaxy properties of different AGN types is combined with an analysis of AGN properties and their clustering. One of the major studies presented in Hickox et al. (2009) revealed several important observations. Analysis of the clustering of different types of AGNs and their local environments shows the connection between AGN mode and its host galaxy mass and age. It is a well-established fact that red galaxies show larger clustering amplitudes and are found in denser environments than objects from the blue population (Zehavi et al., 2005; Coil et al., 2008). In the case of AGNs, the X-ray-selected AGNs predominantly occupy host galaxies being in the transition between blue and red populations. These AGNs are, in general, characterized by a larger clustering amplitude, inhabiting dense environments (Gilli et al., 2005; Miyaji et al., 2007). Even more robust clustering is shown by radio-selected AGNs, which occupy red sequence hosts, are found in groups and clusters of galaxies, and show clustering properties similar to local elliptical galaxies (Wake et al., 2008; Mandelbaum et al., 2009). Infrared-selected AGNs residing in bluer hosts show, on the contrary, much weaker clustering properties, with some indications of IR-bright AGN activity being triggered not only by the host properties but also by a local, preferably underdense, environment (Hickox et al., 2009).

This type of study opens up an opportunity to bring AGNs into the context of observational cosmology. Clustering analysis on a scale of $\sim$Mpc allows one to estimate the dark matter halo masses (Sheth and Tormen, 1999) and connect them to different types of AGNs, similarly as it is done with normal galaxies (Faber et al., 2007). In particular, radio-selected AGNs at relatively small redshifts ($0.25 < z < 0.8$), with the strongest clustering, occupy most massive dark matter halos ($M_{\rm halo} \sim 3 \times 10^{13} h^{-1} M_{\odot}$), with an increasing bias on a smaller scales ($< 0.5 h^{-1}$Mpc). Halos of these masses characterize large galaxy groups or small clusters. The increasing bias on a smaller scales implies that cross-correlation of radio-selected AGNs with normal galaxies on these scales will tend to select pairs of objects occupying the same dark matter halo (Zehavi et al., 2004; Coil et al., 2008; Brown et al., 2008). The estimated dark halo mass obtained from X-ray selected AGNs clustering gives one smaller value of $M_{\rm halo} \sim 10^{13} h^{-1} M_{\odot}$, typical for smaller galaxy groups. What is interesting, a cross-correlation of X-ray AGNs with normal galaxies at smaller scales ($\sim 1 h^{-1}$ Mpc) shows a smaller slope compared to the X-ray AGN autocorrelation function. A similar result was obtained on the optically selected sample (Li et al., 2006), showing weaker AGN-galaxy clustering on small scales. Authors explained this behavior as the AGNs' tendency to occur in the central galaxies. However, in the case of X-ray AGNs, a small AGN catalog size did not allow for an assured conclusion. The IR-selected AGNs occupy even smaller halos of $M_{\rm halo} \leq 10^{12} h^{-1} M_{\odot}$ masses. In this case, the authors also made a hypothesis of IR-selected AGNs being present in the central galaxies of the small halos.

As it was briefly mentioned in the Sec. 2.2, radio, X-ray and IR AGN selection methods sample different ranges of Eddington ratio. In the Hickox et al. (2009), authors showed that X-ray selected AGNs sample almost the whole Eddington ratio range of $10^{-3} \leq L/L_{Edd} \leq 1$, while radio-selected AGNs sample lower limit of the distribution ($L/L_{Edd} \leq 10^{-3}$) and IR-selected AGNs represent the upper limit ($L/L_{Edd} \geq 10^{-2}$).

These observations were interpreted by the authors of the discussed work as an AGN complement of the previously discussed galaxy evolution model. In this scenario, a radiative phase of the AGN and formation of the stellar bulge occurs when the dark matter halo reaches a mass of 10^{12} - $10^{13} M_{\odot}$. After this episode, star formation in the host becomes suppressed, and AGN mode changes from radiative (represented by optically and IR-bright AGNS) to kinetic mode (i.e., radio-bright sources). Let us briefly summarize the picture of AGN and galaxy co-evolution described in this section from the perspective of the present work. Different methods of AGN selection allow us to trace different episodes of the AGN accretion modes, different host galaxies, and different environments and dark matter halo sizes. These properties make the statistical analysis of AGN catalogs a very powerful astrophysical and cosmological tool. Thus, increasing the multi-wavelength catalogs volume and preserving the high purity and completeness of these catalogs is crucial for modern cosmological studies.

3

Data

3.1 Sky surveys of the North Ecliptic Pole Field

The region of the North Ecliptic Pole (NEP; $\alpha(\mathrm{J2000}) = 18^h00^m00^s$, $\delta(\mathrm{J2000}) = 66°33'88''$, in Galactic coordinates $l = 96°.4$, $b = 29°.8$) is one of several deep sky fields observed by unprecedentedly large number of instruments giving us a broad, panchromatic view on the nature of astronomical objects. Space telescope missions often choose the NEP region due to the convenience of observation from geocentric orbit and its properties with respect to the extragalactic studies: NEP is a non-distinctive region of the sky at relatively high galactic latitude. This combination allows one to perform a deep survey and reduce the cosmic variance.

The NEP region, observed in various wavelength ranges, gives us a broad, panchromatic picture of the deep sky, from X-ray to radio data. The high-energy picture of NEP was first investigated by ROSAT X-ray space mission (Truemper, 1982). The ROSAT NEP survey (Henry et al., 2001; Voges et al., 2001; Henry et al., 2006) observed a wide area of 80.6 deg^2 centered around NEP (see also recent release of the updated ROSAT NEP catalog in Hasinger et al., 2021). This catalog, together with optical follow-up (Gioia et al., 2003), was used for many studies on galaxy clusters (Gioia et al., 2001; Mullis et al., 2001; Gioia et al., 2004; Pratt and Bregman, 2020), AGN studies (Wolter et al., 2005)) or study of AGN-cluster interplay (Branchesi et al., 2006; Cappelluti et al., 2007). Other X-ray telescopes performed observations of the NEP region as follow-ups to the existing (or forthcoming surveys). The Chandra telescope (Weisskopf et al., 2000) performed follow-up in the AKARI NEP-Deep field (Krumpe et al., 2015) and NuSTAR (Harrison et al., 2013) performed follow-up in the now planned James Webb Space Telescope North Ecliptic Pole time-domain field (Zhao et al., 2021), which is a small field of 0.16 deg^2 area near the NEP center. In the near future the NEP region will be also observed with the eROSITA telescope (Predehl et al., 2021) if the eROSITA comes back to operational mode.

Optical and ultraviolet studies of the NEP region were done almost exclusively as follow-ups for particular fields observed by other, mainly IR instruments. Several major IR missions observed the NEP field. The first deep IR observations were done by the IRAS satellite (Neugebauer et al., 1984) in the MIR-FIR range (Hacking and Houck, 1987) with succeeding radio follow-up with the VLA (Hacking et al., 1989). This catalog provided observations for pioneering works on IR properties of galaxies and galaxy evolution models (see, e.g., Hacking, Condon, and Houck, 1987; Hacking and Soifer, 1991; Ashby, Houck, and Hacking, 1992). Next, IR observations in the NIR-MIR range were performed by the AKARI (Murakami et al., 2007; Matsuhara et al., 2006), which performed probably the two most important and deepest IR observations of the NEP region: AKARI NEP-Wide survey of 5.4 deg^2 region (Lee et al., 2009) and AKARI NEP-Deep survey, which covered 0.6 deg^2 area (Wada et al., 2008). Both of them were observed by numerous optical telescopes providing photometric and spectroscopic data (e.g., (Jeon et al., 2010; Goto et al., 2017; Hwang et al., 2007; Huang et al., 2020; Oi et al., 2021). Additional far-infrared follow-up of the AKARI NEP region was performed by the Herschel satellite (Pilbratt et al., 2010). It observed two AKARI NEP fields with two different instruments: AKARI NEP-Wide field was covered by longer-wavelength SPIRE (Griffin et al., 2010) instrument (Pearson et al., 2017), while AKARI NEP-Deep was observed by PACS (Poglitsch et al., 2010; Pearson et al., 2019). Moreover, the AKARI NEP field was also covered by GALEX satellite (Martin et al., 2005) providing observations in near- and far-UV (Burgarella et al., 2019). Additional long-wavelength observations of the AKARI NEP-Deep were performed by the JCMT/SCUBA2 (Holland et al., 1999) in the sub-mm range (Geach et al., 2017; Shim et al., 2020) and by The Westerbork Radio Synthesis Telescope in the radio wavelengths (White et al., 2010)). After accomplished observations of AKARI, the deep field NIR-MIR data of the NEP region was collected by two other major IR space observatories. The first one was the WISE telescope (Wright et al., 2010). With several hundreds of passages through the NEP, it was able to create a deep field catalog of 1,5 deg^2 area (Jarrett et al., 2011). Second one was the Spitzer telescope (Werner et al., 2004), which covered larger, 7.04 deg^2 area (Nayyeri et al., 2018). The NEP field was also observed by the specialized CIBER (Bock et al., 2013) instrument used for studies on the nature of the near-infrared cosmic background (Zemcov et al., 2014).

This multi-wavelength data collected in the NEP region was used in a large number of extragalactic studies. These were numerous studies of AGN properties (see, e.g. Wang et al., 2020; Yang et al., 2020; Chiang et al., 2019; Santos et al., 2021; Barrufet de Soto et al., 2017, galaxy evolution (see, e.g., Kim et al., 2015; Oi et al., 2017; Kim et al., 2019; Goto et al., 2019; Barrufet et al., 2020; Kim et al., 2021a) or galaxy cluster studies (see, e.g., Solarz et al., 2015; Seo et al., 2019; Huang et al., 2021) to name a few. Besides strictly astrophysical research, the NEP data was used for applied machine learning studies. These were methods for automatic AGN selection in IR and multi-wavelength data (Poliszczuk et al., 2019; Poliszczuk et al., 2021; Chen et al., 2021), star-galaxy separation (Solarz et al., 2012) or deep learning models for galaxy mergers selection in the optical data (Pearson et al., 2022).

3.2 Multi-wavelength catalog of AKARI NEP-Wide field

As discussed in previous sections, understanding the IR properties of different galaxy populations is crucial to understanding star-formation history, galaxy evolution, and AGN feedback processes. For these purposes, the most complete multi-wavelength catalog in the NEP region is provided by the recently published multi-wavelength merged catalog of the AKARI NEP-Wide field (Kim et al., 2021b). This catalog is based on the AKARI NEP-Wide source catalog (Lee et al., 2009; Kim et al., 2012), cross-matched with optical counterparts provided by the broadband optical follow-up of SUBARU/HSC (Oi et al., 2021). A final, cross-matched catalog of 91,861 AKARI sources is supplemented with existing measurements from other instruments making this data set a broad, panchromatic picture of the NEP field.

The AKARI space telescope (Murakami et al., 2007) was launched in 2006 and performed three main surveys: an all-sky infrared survey in the MIR-FIR range (Ishihara et al., 2010; Doi et al., 2015), as well as two deep-sky surveys in the NEP region in the NIR-MIR range (Matsuhara et al., 2006). The larger, circular area of $\sim$5.4 deg^2 centered around NEP was covered by the AKARI NEP-Wide survey (Lee et al., 2009; Kim et al., 2012). In contrast, the smaller $\sim$0.6 deg^2 area of the AKARI NEP-Deep survey (Wada et al., 2008) is located on the side of the NEP center in order to match pre-existing optical follow-up (Takagi et al., 2012). The AKARI satellite carried two separate focal-plane instruments. The first one was the Far-Infrared Surveyor (FIS; Kawada et al., 2007) created for the FIR survey in four bands spanning the 50–180 μm range. The second was InfraRed Camera (IRC; Onaka et al., 2004), covering the 2–26 μm range with three different channels: NIR (2–5 μm), MIR-S (5–12 μm), and MIR-L (12–26 μm).

The NEP field was also observed with the Spitzer telescope. However, the AKARI satellite has several unique properties, making it better suited to portrait NEP in the NIR-MIR range. First, a very important property is the unique continuous coverage of the 2–26 μm range. It allows filling the 8–24 μm gap in the Spitzer photometry, which is crucial for studying SFG and AGN MIR properties. Moreover, IRC FOV is almost four times larger than the FOV of the Spitzer/IRAC instrument, which allowed to significantly increase the mapping speed of the survey (about 100 times faster than IRAC) and reach a similar survey depth, despite a shorter lifetime of the AKARI instrument in its cryogenic phase (Matsuhara et al., 2006).

Initial broadband optical follow-up of the AKARI NEP-Wide field was performed with two telescopes: Canada-France-Hawaii Telescope (CFHT; Hwang et al., 2007) and Maidanak (Jeon et al., 2010). These catalogs had two main limitations: first, the depth of both catalogs was insufficient to match optical counterparts to the IR source of the AKARI catalog. Moreover, these two follow-ups were characterized by different depths and bandpass transmission curves, making it impossible to homogeneously analyze the multi-wavelength AKARI NEP-Wide catalog (Goto et al., 2017; Kim et al., 2021b). In order to solve this issue, a deep uniform optical survey with SUBARU/HSC instrument (Miyazaki et al., 2012) was proposed (Goto et al., 2017). Observations were taken in two campaigns. First observations were performed with r band in August 2015 and suffered from bad seeing as well as the dome shutter opening error, giving shallower observations compared to the rest of the passbands (g, i, z, and Y), which were used in the second observation campaign in 2018 (Oi et al., 2021). The data reduction was performed with the official pipeline, hscPipe 6.5.3 (Bosch et al., 2018). Obtained data were presented in terms of Cmodel AB magnitude system photometry, which is relatively robust against fluctuations caused by seeing conditions(Huang et al., 2018). Comparing the final SUBARU/HSC catalog with previous CFHT and Maidanak data shows that new HSC observations are 1.7 – 2.5 deeper in g, r, i, and z bands. The new merged catalog is richer in $\sim$20,000 AKARI sources matched to optical counterparts. Most of these objects are presumably faint galaxies (Kim et al., 2021b).

The total number of 111,535 IR AKARI sources were matched with a 3-σ radius (corresponding to 1''.78) with HSC optical counterparts. Out of these objects, catalogs of final, clean matches (91,861) and spurious/flagged matches (19,674) were selected. In addition to pipeline flags, magnitude limits were applied to the NIR AKARI passbands. The bright end limit is occupied by spurious sources as well as flagged stars (Kim et al., 2012), while the faint end does not go beyond the detection limit. This new merged catalog is also supplemented with improved photometric redshifts (Ho et al., 2021), smaller spectroscopic follow-up data (Shim et al., 2013) as well as multi-wavelength data from X-ray to the sub-mm spectral range described in the previous section.

3.3 Training and generalization samples

This section describes important methods of training sample creation and determination of a generalization sample limit. Both of these results were described and published in Poliszczuk et al. (2021). The specific way of training sample creation (see Sec. 3.3.1) allows one to indirectly impose the MIR information into the model structure during the learning phase. In the first publication Poliszczuk et al. (2019), research on the extrapolation risk during IR photometric data classification was performed. It showed difficulties in controlling model performance in the regions of the generalization sample, which were not represented by the training data. In order to avoid extrapolation and a consequent reduction of model reliability, we use the generalization sample limit created by the MCD algorithm (see Sec. 3.3.2). This method was never used in astronomy before. It allows one to precisely control the performance of the model and avoid the risk of extrapolation.

3.3.1 Training sample

In the present work, AGN selection is made using optical and NIR broadband photometry. However, for supervised machine learning algorithms (see Sec. 3), one also needs a sample with known class labels called *training sample* in order to train a classification model and later predict labels on the sample without labels via *generalization* procedure. The latter unlabeled sample is referred to as the generalization sample. Most of class labels were taken from optical spectroscopic follow-up performed in the AKARI NEP-Wide field (Shim et al., 2013) by MMT/HECTOSPEC (Fabricant et al., 2005) and WYIN/HYDRA (Barden et al., 1993) instruments, together with small number of additional spectra taken by Keck/DEIMOS (Faber et al., 2003), GTC/OSIRIS (Cepa et al., 2000) and SUBARU/FMOS (Kimura et al., 2010) spectrographs, by the members of AKARI NEP Team.

The main spectroscopic survey described in Shim et al. (2013) observed two categories of objects. The first group was primary spectroscopic targets, which were selected as objects bright in AKARI MIR passbands ($S11 < 18.5$ mag and $L15 < 17.9$ mag). An additional optical limit in the Maidanak R band was imposed ($16 < R < 21 - 22.5$ mag depending on a spectrograph) to select objects bright enough in optical light for spectroscopic observations. The second group was secondary targets of specific classes. In particular AGN candidates in Shim et al. (2013) were selected according to color-cut method developed for AKARI NEP-Deep data (Lee et al., 2007). This technique defines AGN-dominated region as:

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

These limits allowed to select objects with a high probability of being characterized by the prominent torus dust emission and characteristic power-law NIR-MIR SED shape. In the further part of this work, these objects, with spectroscopic class confirmation as well as at least one emission line with Full Width at Half Maximum (FWHM) larger than 1000 km/s (Shim et al., 2013) will be referred to as *AGN1*. It is crucial to note two important aspects of the training sample construction. First is that stars present in the NEP field (mostly in the bright end of the NIR distribution) were removed from the AKARI/HSC merged catalog through procedures described in Kim et al. (2021b), thus in this work, we are performing two-class galaxy-AGN separation. The second is

band	catalog size	labeled data
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)

TABLE 3.1: Number of objects detected in particular SUBARU/HSC and AKARI/IRC passbands. Numbers in brackets show sources with measurements existing in all previous passbands corresponding to shorter wavelengths. Catalog size refers to all objects present in clean sources catalog (Kim et al., 2021b). Labeled data refers to the subsample of objects with existing spectroscopic class (Shim et al., 2013) or detection of strong X-ray emission detected by Chandra (Krumpe et al., 2015).

that an optically confirmed AGN training sample was selected using targets obtained from the MIR color technique. Thus, the MIR properties of these AGNs indirectly affect the properties and distribution of the sample in optical and NIR bands.

An additional set of class labels was taken from X-ray observations performed by the Chandra telescope in the AKARI NEP-Deep field (Krumpe et al., 2015). In this case, objects with high X-ray luminosity defined as $\log L_X > 41.5$ erg/s in the 0.5-7 keV range were defined as X-ray AGNs. This luminosity limit allows one to include both Seyfert and quasar objects in this group. This class will be referred to as *XAGN*.

Final training (and generalization) data were expected to be detected in all SUBARU/HSC (g, r, i, z, Y) bands as well as all AKARI/IRC NIR bands ($N2$, $N3$, $N4$). It results in a training sample consisting of 1547 objects: 1348 galaxies and 199 AGNs (163 AGN1 and 36 XAGN). Table 3.1 shows how the number of objects detected in particular bands changes while moving from the optical to MIR range. While this change does not seem to be so crucial in the case of the labeled sample (target pre-selection for the spectroscopic observations was limited to the bright, prominent objects), the whole catalog suffers a severe limitation while moving towards MIR passbands. The number of detections in the $L24$ band comprises only $\sim$3% of the catalog with detection in the g optical band. Another visible tendency is the significant decrease in the number of objects moving from NIR to MIR samples. This table perfectly illustrates the necessity for a more efficient AGN selection mechanism.

To analyze the main properties of the training sample, let us look at Fig. 5.3, where class distributions are analyzed with respect to the $N2$–$N4$ color. The choice of this particular color property is not accidental. The difference between the object's brightness in $N2$ and $N4$ passbands can show us two important properties: steepness of the power-law AGN SED shape in the 3–8μm range as well as the presence of

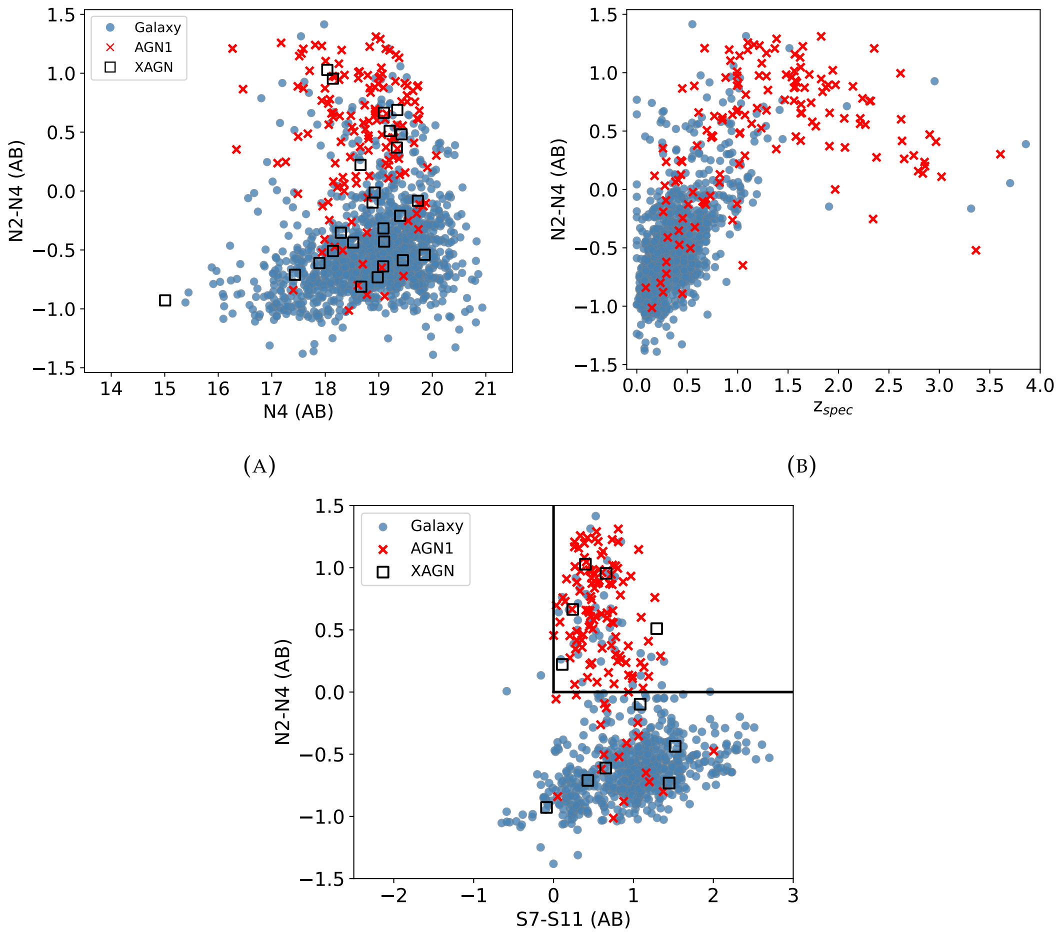
1.6μm stellar bump in the case of high redshift SFG. Due to the above reasons and good AGN-galaxy separation, the color-magnitude plot of $N2$–$N4$ color combined with $N4$ band is often used for AKARI data visualization (see, e.g., Lee et al., 2007; Lee et al., 2009. As we see in Fig. 3.1a, $N2$–$N4$ gives a good distinction between the main AGN1 locus and the center of the galaxy distribution. Due to the lack of stars in the catalog, the stellar locus is empty in this plot. The stellar class is located in the lower-left corner, i.e., the place for objects characterized by the large flux in the NIR band and blue $N2$–$N4$ color. Comparing Fig. 3.1a to Fig. 3.1b, we see that the main contamination of the AGN locus in the $N2$–$N4$ distribution comes from the high-redshift SFG, as it was discussed in the previous Chapter. Thermal dust emission from the torus is further taken use of in the traditional MIR AGN selection technique applied for the AKARI data (Eq. 3.1). Results of this classification is presented in Fig. 3.1c. AGN selection conditions were slightly relaxed in this figure, i.e., the $S11 < 18.5$ requirement was removed, so the additional small sample of spectroscopically confirmed AGNs and XAGNs were included in the classification scheme. The lack of this limit does not strongly affect classification efficiency, and one can still see a clear separation between the galaxy and AGN classes.

The XAGN sample distribution causes a separate problem. As one can see, it does not have a well-defined locus, and XAGN sources are scattered all over the space assigned to both AGN and galaxy classes. The main reason for this behavior is the difference between X-ray and IR AGN selection. As was discussed in the previous Chapter, NIR-MIR AGN selection can only identify AGN with a high Eddington ratio (i.e., $L/L_{Edd} \geq 0.01$). Thus AGNs which are fueled by radiatively inefficient accretion flows (which are also often lacking signs of broad-line region (Trump et al., 2009)) may escape NIR-MIR (as well as optical) selection (Donley et al., 2012).

The final topic that will be addressed in this Section is the accuracy of photometric redshift estimation. Photometric redshifts used in this work were included in SUBARU/HSC - AKARI/IRC merged catalog (Ho et al., 2021). They were obtained via the χ^2 template-fitting method Le Phare (Arnouts et al., 1999; Ilbert et al., 2006) and were based on multiple photometric bands from the UV-IR range. A comparison between spectroscopic redshift and photometric redshift estimation based on methods described in Ilbert et al. (2006) is presented in Fig. 3.2. We see a good agreement between spectroscopic and photometric redshift at lower redshift ($z < 1.5$) for both galaxy and AGN classes. The larger redshift range, occupied mainly by the AGN class, shows a significant bias towards lower redshifts, smearing high-redshift AGNs all over the redshift range. This discrepancy is caused by the lack of an AGN template used for photometric redshift estimation (Ho et al., 2021), making analysis of redshift properties of AGN candidates catalog very challenging.

3.3.2 Generalization sample and MCD limit

In the very nature of supervised learning, a classification model can only recognize observations that relate to the training sample's properties. This characteristic leads to two important consequences. First, we can manipulate the properties of the training sample in order to emphasize some specific behavior of the classifier. Instance weights of specific properties or MIR-based target selection of spectroscopic observations used in the training sample belong to this class of manipulations. The second consequence is the inability of a classifier to properly extrapolate its predictions beyond the region in the feature space occupied by the training sample. This second consequence imposes an important limitation on the generalization sample. Its properties in the feature space should not be substantially different from the properties of the



(A)

(B)

(C)

FIGURE 3.1: Training sample properties. *Panel A: N*2–*N*4 vs *N*4 color-magnitude plot of the training data. *Panel B:* Plot of the *N*2–*N*4 color against spectroscopic redshift. Here XAGN are missing since they do not have measured spectra. *Panel C: N*2–*N*4 vs *S*7–*S*11 color-color plot used in Lee et al. (2007) for AKARI MIR AGN selection.

training sample. Such a situation is problematic because the classifier does not have information about the regions outside the training sample and because we cannot test its performance in those regions. Without training data observations, we cannot calculate any evaluation metric. Thus both a model and a scientist who trains it remains blind in those outside regions.

Limits imposed on the astronomical data usually take the form of color and magnitude cuts. However, while this approach allows one to preserve a simple form of the selection function, it does not satisfy the general demands of supervised learning. The main reason is that simple color and magnitude cuts cannot properly define the limits of the high-dimensional manifold created in the feature space by the training sample and may leave regions where a classifier would have to extrapolate in order to make a prediction.

This work was an attempt to fulfill both requirements: to construct an effective limit for a generalization sample in the high-dimensional space and preserve a relatively simple, where reconstruction of the selection function remains possible. To achieve both of these goals, a well-established method of minimum covariance

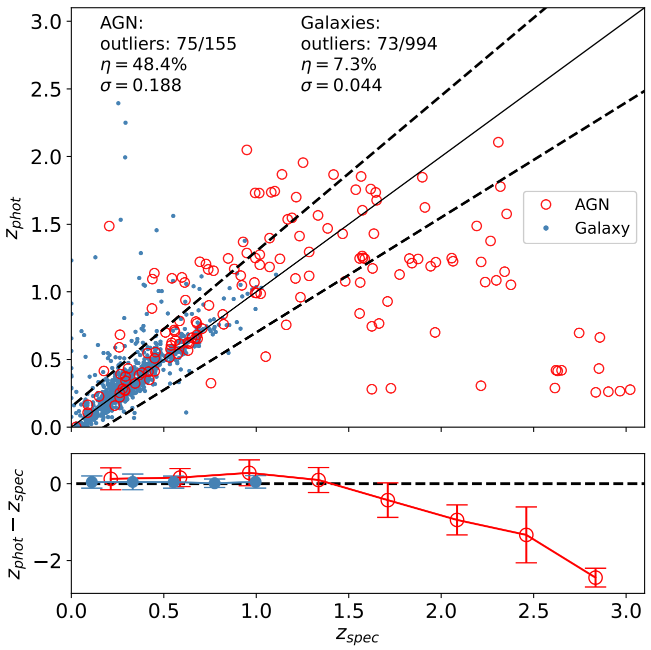


FIGURE 3.2: Comparison between the spectroscopic redshift and photometric redshift estimation from Ho et al. (2021) for labeled galaxy (blue dots) and AGN (red circles) data. The cone created by dotted lines refer to $z_{phot} = z_{spec} \pm 0.15 \times (1 + z_{spec})$. The η describes the fraction of outliers (or catastrophic errors) defined as objects outside the cone. Sigma is the normalised median absolute deviation defined as $\sigma = 1.48 \times \text{median}(|\Delta z|/(1+z))$. Lower plot shows the mean residuals with standard deviations. Only $z < 3$ objects are shown.

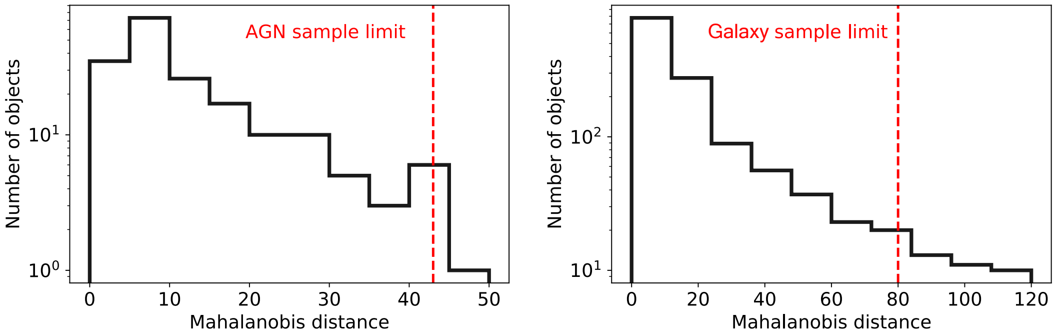
determinant estimator algorithm (MCD, Rousseeuw and Driessen, 1999) was used. The MCD method allows one to fit a high-dimensional ellipsoid to the training data set and limit the shape of the generalization sample to the ranges of this ellipsoid. The MCD method has a single free parameter α called *contamination rate*. This parameter controls the amount of data (in our case, training sample observations) allowed outside an ellipsoid.

In order to create a shape reflecting the distribution of the training sample, two separate ellipsoids were fitted to galaxy and AGN classes separately. It was done to avoid two problems. First, galaxy and AGN classes may have very different distributions in the high-dimensional space, and an effective fit of a single ellipsoid might not be possible. Second, a strong imbalance of two classes may lead to a further diminishing of AGN observations during the α parameter tuning process. Search for an adequate α parameter value was done via inspection of Mahalanobis distance distribution. The Mahalanobis distance d_M is defined as:

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

where $\vec{x}$ points towards observation location, $\vec{\mu}$ is the vector of mean feature values, and Σ is the covariance matrix. Fig. 3.3 presents Mahalanobis distance histograms for AGN and galaxy classes. Choice of α value is subjective and depends on the specific application of the MCD algorithm. There was no need for a large α value, which corresponds to very conservative limits. The main purpose of MCD limit was the extrapolation avoidance during the generalization phase. Thus, the α value was selected to correspond to the d_M range, where histogram discontinuities or where a large decrease in the number of objects occurs. It was chosen to be $\alpha = 0.065$ ($d_M \simeq 43$) and $\alpha = 0.05$ ($d_M \simeq 80$) for AGN and galaxy classes respectfully. Two different ellipsoids with corresponding α values were fitted to AGN and galaxy classes, and the generalization sample shape was limited to these ellipsoids. Thus final generalization sample consists of objects which belong to at least one ellipsoid. These limits were created in the optical-NIR magnitude space, not the final feature space where predictions were made, to preserve the simple shape of the imposed cuts.

Application of the MCD limit (together with the detection requirement in optical HSC and NIR IRC passbands, which limits catalog size to 45 841 sources) gives us a final generalization sample made of 33 119 objects. Statistical comparison between training and generalization samples is presented in Tab. 3.2. Here we see that median values of the generalization sample are shifted towards the faint end of the distribution. It is caused by the properties of the training sample, where additional optical brightness conditions were imposed to perform spectroscopic measurements. At the same time, ranges of distributions in particular bands of generalization sample remain within training sample boundaries (except in a few cases, such as the faint end of the g band). Another analysis of generalization sample properties is presented in Fig. 3.4, 3.5. Fig. 3.4 shows brightness distribution in optical and NIR-MIR IRC passbands. We see that MCD-limit cuts the faint end of the distribution, especially in the optical part of the electromagnetic spectrum. Fig 3.5 shows distribution of $N2$–$N4$ color. Here we see a very strong reduction of objects characterized by $N2$–$N4 = 0$, which is a part of the color distribution occupied mainly by low-redshift non-active galaxies.



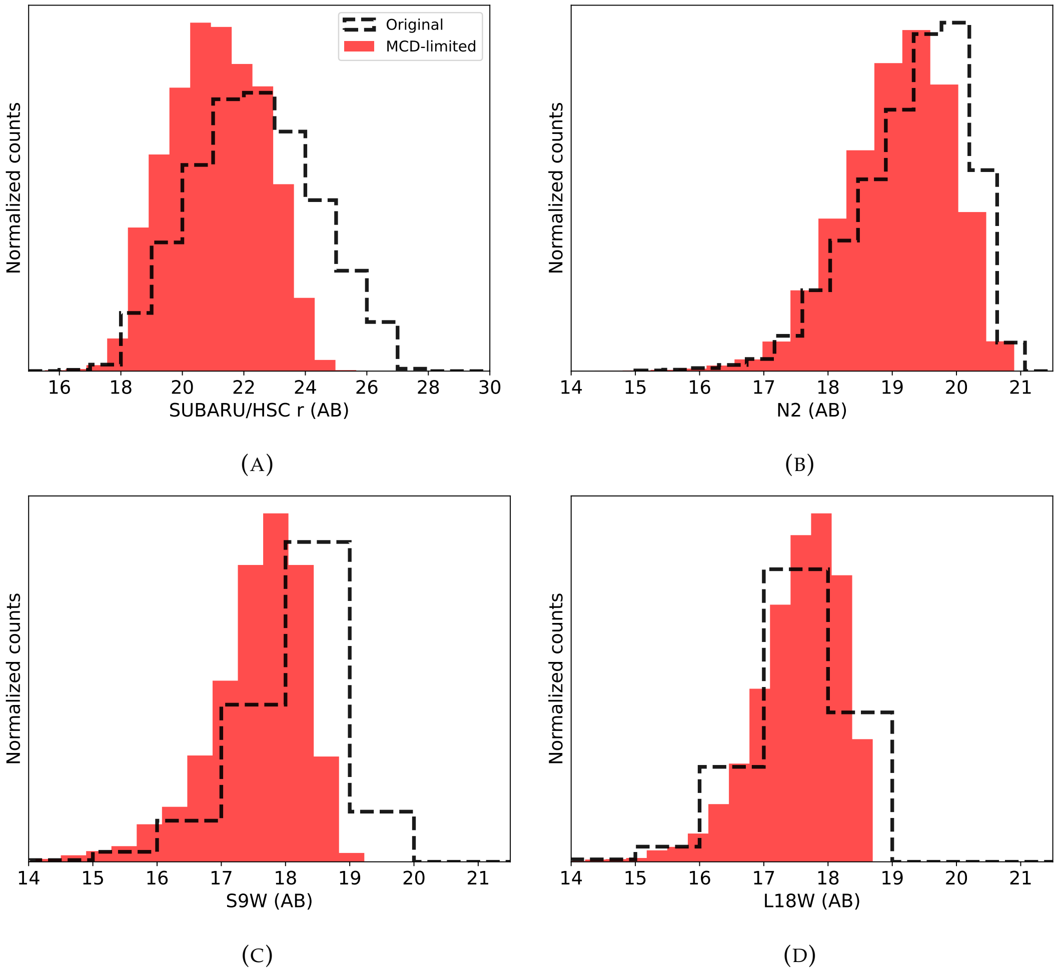
(A)

(B)

FIGURE 3.3: Mahalanobis distance histograms for AGN and galaxy training samples. *Panel a:* AGN sample. *Panel b:* galaxy sample. Dashed red lines correspond to a particular contamination parameter value of the MCD algorithm used to limit generalization data set to the training sample range.

	median	**MAD**	**min.**	**max.**
		Training sample		
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
		Generalization sample		
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

TABLE 3.2: Statistical properties of the training and generalization samples. Median, median absolute deviation (MAD), minimal and maximal values for redshift and photometry in broad-band fileters used in training and generalization.



(A)

(B)

(C)

(D)

FIGURE 3.4: Histograms of the brightness distribution in optical and NIR passbands in MCD-limited generalization sample (red, filled histogram) and original SUBARU/HSC-AKARI/IRC merged catalog (Kim et al., 2021b) (black, dashed line). *Panel A:* SUBARU/HSC *r* band. *Panel B:* AKARI/IRC *N*2 band. *Panel C:* AKARI/IRC *S*9*W* band. *Panel D:* AKARI/IRC *L*18*W* band.

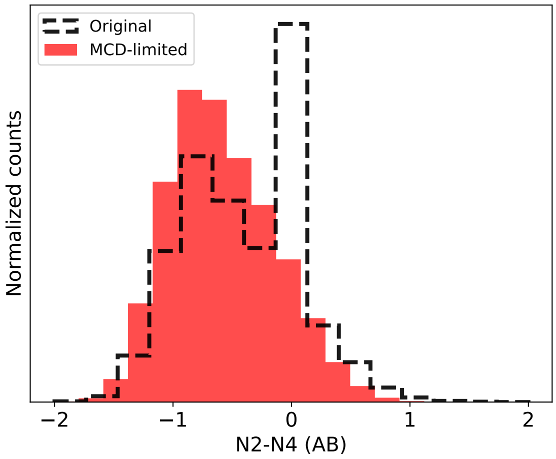


FIGURE 3.5: Histogram of $N2 - N4$ color distribution for MCD-limited generalization sample (red) and original SUBARU/HSC-AKARI/IRC merged catalog (Kim et al., 2021b) (black, dashed line)

4

Machine learning techniques

4.1 Supervised classification

This section will discuss the concept of supervised learning and present several supervised learning algorithms used in this. This section was inspired by the explanation presented in the following two books Hastie, Tibshirani, and Friedman (2001) and Bishop (2006). The *supervised learning* is a branch of machine learning, where a model is trained on the data with known labels. These labels often referred to as a *target*, might be discrete class labels in the case of the classification task or continuous values in the case of a regression problem. In the supervised setting, labeled data is represented by two sets of parameters. One set relates to the data labels. This label might be a one-dimensional label or a vector of labels in the case of more complex learning tasks. Label (or target) will be denoted by y. Another set of parameters is called a *feature set*, and it is common for both labeled and unlabeled data and will be denoted by x. Thus each object in the labeled data, which will also be referred to as *observation* is defined by the pair (x_i, y_i).

The supervised model tries to learn the connection between features and labels on the labeled data. Once the model learns this mapping, it can predict labels on the unlabeled data using a representation of the unlabeled data in the feature space. This process of prediction on the unlabeled data is known as *generalization*. We will refer to the labeled data as *training data* and to unlabeled data as *generalization data*. In the classification problem, the supervised model tries to learn how to assign class labels to the data based on their properties in the feature space. The basic form of classification, which is used in this work, is the discrimination between two classes. This type of classification is known as *binary classification* and classes are often referred to as *positive* and *negative* with assigned targets $y = 1$ and $y = -1$ respectively.

In order to find the most suitable model for a specific learning task, one should perform a *model selection* process. First, a machine learning algorithm should be tested with various *hyperparamters* combinations of the model. Hyperperparameters define specific tunable properties of the algorithm. It is also a good practice to test different types of machine learning algorithms in order to find the best one. The need for hyperparameter tuning and testing of various learning algorithm types can be explained from the perspective of *bias-variance trade-off*. The model's bias describes the strength of the assumptions model makes about the prediction. For example, linear models tend to have a larger bias than more flexible non-linear models. Thus, too high bias lead to the model rigidity and inability to adapt to the training data. This phenomenon is known as *underfitting*. However, reducing the bias of the model cannot be treated as a remedy for poor classifier performance. A model which will be too flexible may learn the distribution of the training data too precisely. In this case, the model learns not only the real properties of class populations but also the statistical noise present in the training sample. This phenomenon is known as *overfitting*. Overfitted model is characterized by low bias and high variance. The high variance model is very sensitive to every small change in the training data, which makes it prone to learning the noise component of the data distribution. Such a model, while having good results on the training data, will not generalize well. The model training and later model selection are important steps allowing to find a trade-off between bias and variance of the final classifier.

4.1.1 Linear models and logistic regression

If the decision boundaries are linear, we refer to the model as a *linear classification method*. In the simplest approach a linear decision boundary can be created by fitting a linear model for the target variable:

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

where i is the class label, x is the observation and $\{\theta_{i0}, \theta_{i1}\}$ are the model parameters. The decision hyperplane between the classes 1 and 2 is created by the point for which $f_1(x) = f_2(x)$ that is:

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

In this work I used the *logistic regression* algorithm as a representation of the linear classification models family. A linear regression algorithm models the posterior probability as a linear function of observations. In a *binary classification case* the posterior probability is modeled as:

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

where $\theta = (\theta_0, \theta_1)$. In order to preserve linear decision boundaries one can apply a monotone *logit* transformation:

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

Thus a separating hyperplane is created by the set of points for which logit transformation is equal to zero.

A logistic regression model is fit by a maximum likelihood method, using a conditional log-likelihood of class k, given observation x. In terms of machine learning, we are working on the minimization of the negative log-likelihood, which gives the *cross-entropy* error function. Working on a binary classification example, where $k_i = 1$ gives a response target value $y_i = 1$ and $k_i = 2$ gives $y_i = 0$, a log-likelihood for N observations is defined as:

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

Here we assume the vector n-dimensional input vector x_i becomes (n+1)-dimensional and includes 1 to accommodate an intercept parameter θ_0. Thus the error function is defined as

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

To minimize the error function (and to maximize log-likelihood) we set its derivatives to zero obtaining n+1 nonlinear equations:

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

The error function is minimized by the Netwon-Raphson iterative optimization technique (Fletcher, 1987)

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

where **H** is a Hessian matrix made of second derivatives of error function:

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

After the optimization of the model parameters, it is used to make final predictions on the data. A model can be further modified by adding a penalty term, which is often based on L_1 or L_2 norms. Adding the normalization imposes restrictions on the model during the learning phase and improves its performance. In this case, an error function gains an additional regularization term $R(\theta)$, which form depends on the regularization strategy:

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

The C parameter controls the strength of regularization and can be tuned. Moreover, the logistic regression model can be further tuned with the class or instance weights by introducing constant weights under the sum present in the log-likelihood function.

4.1.2 Support vector machine

Another group of methods for class separation is based on searching for a hyperplane, which maximizes the margin between classes in training data. A popular algorithm that uses this approach is the *support vector classifier* (SVC). For a non-separable case, an SVC optimization problem can be defined as

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

where M refers to the margin between class and separating hyperplane (thus margin between two classes is equal $2M$). The $\xi = \{\xi_i\}_{i=\overline{1,N}}$ parameter controls the misclassification rate of the model. A particular point x_i is misclassified when $\xi_i > 1$. The ξ_i is also proportional to the distance of prediction displacement on the wrong side of the margin. By introducing bounding condition $\sum_{i=1}^{N} \xi_i < \text{const}$, we can control the total amount of misclassifications on the training data.

Equations 4.11 describes a standard formulation of the support vector classifier and forms a convex optimization problem (a quadratic programming problem with linear inequality constraints). It can be solved via the Lagrange multipliers method. We remove the norm constrain on the θ_1 parameter by defining

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

which transforms 4.11 problem into minimization of $\|\theta_1\|$. By additional replacing a constant, which controls the number of misclassifications by the tunable *cost parameter* C, we obtain an equivalent form of 4.13 optimization problem, which is more useful from the computational point of view:

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

Now we can construct a Lagrangian primal function

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

After minimizing L_p with respect to θ_1, θ and ξ_i and setting respective derivatives to zero, we obtain three conditions:

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

By applying these conditions together with positivity constraints on Lagrange multipliers and ξ_i, we obtain a Lagrangian dual objective function:

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

which gives a lover bound on the objective function presented in Eq.'s 4.13. After maximizing L_d subject to $0 \leq \lambda_i \leq C$ and second condition from 4.15 we obtain three additional conditions:

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

Both sets of constraints, 4.15 and 4.17 form so-called Karush-Kuhn-Tucker (KKT) conditions. These, together with 4.18 uniquely characterize the solution of the optimization problem. Moreover, the KKT conditions define an important subsample of observations characterized by non-zero λ_i parameter value called *support vectors*. Some of these points with $\xi_i = 0$ will lie on the margin and would be used to create a separation hyperplane, while support vectors with $\xi_i > 0$ will lie beyond the margin.

The *support vector machine* (SVM, Cortes and Vapnik, 1995) is an extension of the support vector classifier (SVC). The support vector classifier is a linear method that operates in the original feature space. It can become more flexible by adding artificial dimensions made of combinations of original features. This way, a starts acting in a non-linear way. In the SVM, these artificial features are introduced via *kernel function* $k(x, x')$ containing original features only in the form of inner products. This modification is known as *kernel trick* and does not require knowledge about the exact mapping between the initial feature space and the new high-dimensional feature space (see Cortes and Vapnik, 1995 for an in-depth discussion on the kernel trick properties). The introduction of the kernel function transforms the Lagrangian dual objective function into

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

In this work, we use radial basis kernel function (RBF), defined as

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

where γ is a tuning parameter of the model. The SVM model was used in this work in the traditional version described above as well as weighted version of SVM called *fuzzy SVM* (Lin and Wang, 2002). In this case an additional fuzzy membership s_i is introduced into the optimization problem:

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

This way, different observations present in the training data can have different impacts on the learning process. Such weights may be applied in an instance weights manner, where each object has a different weight. They can also be applied in the form of class weights, where smaller class objects have larger weights in order to increase their importance and shift the decision boundary outwards from the smaller class.

Posterior probability estimation for the SVM is obtained in a similar way to the logistic regression method. The output of the SVM classifier provides one a distance of the object from the separating hyperplane. In order to calibrate them and transform them into probability estimation, this distance is then fitted to the sigmoid function similarly to the Eq. 4.3. A detailed description of this procedure can be found in Platt (1999).

4.1.3 Ensemble methods and decision trees

Another popular family of methods applied to classification problems is based on the construction of *decision trees*. Most decision tree algorithms (including ones that are used in this work) divide the feature space into rectangle fields with assigned labels by applying recursive binary splitting. Thus each division obtained by the external node m (without child nodes) in the tree can be identified with the obtained region R_m. This type of decision tree is referred to as *CART* (classification and regression tree) and was described in Breiman et al. (1984).

The response in a particular region R_m is modeled as a majority class label present in this region in the training set. Let us define N_m as a number of observations in R_m and p_m as a probability of observing the positive class member in R_m, which can be estimated via proportion of positive class observations in node m:

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

In order to define the structure of the tree, we need to set up two mechanisms: one is the search for an optimal binary partition performed in the node, and the second is the way to establish an optimal tree depth. A decision tree is built by creating a splitting node that uses a selected feature with a specific threshold. The threshold applied to the feature divides the region of the feature space into two separate sub-regions, above and beyond the threshold value. In a most basic form, feature and separating threshold are selected in a greedy manner by minimizing impurity measure in the node. Usually, a Gini index is used for this purpose. For a two-class problem it can be simplified into a form:

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

where $\hat{p}_{mk}$ is the proportion of the class k in node m.

The creation of a fully grown tree is not always optimal. Tree size is treated as a hyperparameter of the model, which controls the complexity (and thus the variance) of the model and can be tuned. A simple approach might be the termination of the growth of a tree when the accuracy of prediction does not improve with additional splits. Such a method may not give optimal results because sometimes improvement of the performance occurs not with every split. Instead, it may take place after several splits. To avoid this problem and better control the tree size parameter, a method called *cost-complexity pruning* is used. Let us denote a fully grown tree as T_0 and T being any pruned version of T_0. Also let $|T|$ denote number of terminal nodes in T, where *terminal node* refers to a node with no child nodes. Let us define *cost-complexity criterion* as

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

where m refers to a region R_m, N_m is the number of observations in that region, Q_m is the node impurity measure, and α is a tuning parameter. The cost-complexity pruning method searches for a subtree that can most effectively minimize C_alpha.

Tuning parameter α allows one to control the bias-variance trade-off. Control on the tree size and its general properties can be further developed via the application of class-based and instance-based weights. This can be done during the procedure of tree growth. The probability $\hat{p}_m$ is obtained by calculating objects from a specific class in the m region. Thus one can re-weight these objects before they are used for $\hat{p}_m$ estimation. Such modification will affect the impurity measure during tree growth and, as a consequence, will modify the properties of the classifier. The same approach to weighting is used for tree ensembles discussed in the following paragraphs.

While the decision trees perform very well on the training data, they are generally characterized by a high variance and tendency to overfitting (Breiman et al., 1984). To overcome this problem, the *bagging* averaging method is often used to reduce the variance of the model. The bagging method is based on training an ensemble of models on the bootstrap samples (training sample subsets drawn with replacement) and averaging model predictions to obtain the final one. Thus if we define a set of B bootstrap samples and obtain a class vote for observation from each tree, then the final class prediction would be obtained via majority vote.

The task of the variance reduction of a tree-based model can be further modified as it was done in *random forest* algorithm (Breiman, 2001). The variance of an average result obtained from B identically distributed trees is defined as

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

where σ^2 is the variance of a single tree, and ρ is the pairwise correlation. The first term cannot be reduced by the increase of B. As a result, the correlation between bagged trees limits their ability to reduce variance. The main idea of the random forest is to combine the bagging method with the de-correlation of component trees in order to further reduce the variance. Such de-correlation is done via random selection of feature subset m before each split in the tree (i.e., this subset of features is used to obtain the best feature and split point in each node). For the classification tasks, the default size of the feature subset is $\lfloor\sqrt{p}\rfloor$, where p is the total number of input features. Randomization in the tree construction can be pushed even further as it is done in the *extremely randomized trees* algorithm (Geurts, Ernst, and L., 2006), where the final threshold in the node split is obtained by setting random thresholds for features in subset m and taking the best threshold as the final one. It is worth noting here that both ensemble tree structure of both random forest and extremely randomized trees algorithm allows one to obtain a straightforward way of probabilistic classification. In most implementations, probability estimation is given by a majority vote among the single classification trees in the ensemble.

In the traditional formulation of bagging, the expectation of an average of any of the bootstrap samples is the same because of the tree models identical distribution. Thus the bias of the bagged models is the same as of the individual component trees and cannot be lowered. It is also true for random forests and extremely randomized trees. An improvement in the prediction is made only by variance reduction.

In order to reduce the bias, one would have to build an ensemble out of trees that are not identically distributed. Such an approach is used in the *boosting* method. Boosting algorithms use *weak classifiers* - models, which are only slightly better than assigning class labels by random guessing. Such weak learners are then stacked into an M sequence, where the subsequent models are trying to correct the prediction of the previous ones. The final prediction of the boosting algorithm can be expressed as a weighted combination of component models:

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

For the convenience of this equation, we changed labeling from $\{0,1\}$ to $\{\pm 1\}$. Weight α_m describes the contribution of f_m to the final prediction. During each step, m training observations are re-weighted, which forces a subsequent classifier f_m to focus on problematic cases. Here β_m is a set of parameters describing weak learner f_m. In the case of a decision tree, these would be split variables and split points in particular nodes. A version of the boosting model that uses decision trees as base learners is referred to as *boosted trees*. The forward stagewise boosting procedure is a very greedy strategy. Thus, usually, boosting algorithms incorporate numerical optimization via gradient boosting (see Hastie, Tibshirani, and Friedman, 2001 for the review on boosted trees algorithm). In this work, we used a very popular and effective XGBoost (Chen and Guestrin, 2016) implementation of gradient boosted trees algorithm.

In addition, this work incorporates two ensemble methods that use strong classifiers, i.e., fully developed and fine-tuned models. These input models might be selected classifiers of any type that showed the best performance. These additional ensemble methods will be referred to as *voting schemes* or *voters*. The first type of voter will be referred to as *hard voter*. It takes a majority vote from the input models and uses it as a final prediction. The second type of voter will be referred to as *stacked classifier*. It is a simple logistic regression model which uses probability estimations from the input models as a feature set and makes a prediction based on the estimated probability distribution properties. The basic idea of these methods is to further reduce the variance of the final model and stabilize prediction in the case of the small training data sample.

4.2 Performance evaluation

The effectiveness of the classification performed by a machine learning model on the unlabeled data, destined for generalization, is usually estimated via *evaluation metrics* calculated on the labeled data. Metrics calculated on the labeled data used to train the model might show too high values due to the possibility of overfitting. Thus, in order to obtain reliable metric scores, they should be calculated on the labeled data that is taken out from the model training process. Such a procedure is referred to as the training-test split. In such a split, a classifier is trained on one subset of the labeled data, called the training set, and evaluated on another one, called the test set. In order to further increase the accuracy of the evaluation and to reduce the effect of the random division of the labeled data, this procedure is repeated several times. This multiple-step evaluation is known as the *n-fold cross-validation*. Here the labeled set is divided into n subsets, trained on $(n-1)$ and validated on the last one. This process is repeated n times with different combinations, and the mean value of all validation results is taken as a final estimation of the metric score. Cross-validation is especially important in the case of a small amount of training data, where the risk of overfitting is higher. It can be mitigated even further one can perform cross-validation many times, shuffling the data before each iteration of the n-fold split.

Another problem relative to this work is the learning of imbalanced data. Different sizes of classes in the training sample affect two aspects of the model learning process. First, it has a major impact on the way in which the model separates classes. Usually, if one class is much smaller than another, the model tends to shift the separation plane towards a smaller class. It may have benefits due to the reduced contamination of the smaller class at the expense of reduction of its completeness. Secondly, it affects metric scores making them often overestimate the efficiency of model performance, even when the performance of the model in a small class is very poor. Thus, in the case of imbalanced data, one should choose evaluation metrics that will not be prone to such effects (see, e.g., Fernández et al., 2018 for the in-depth discussion). In this work, the main goal was to obtain a reliable catalog of AGN candidates. As was discussed in Chapter 3, the AGN training sample is much smaller than the galaxy sample. Thus, the emphasis was made on the proper evaluation of the classifier performance on the smaller AGN class.

From now on, in this section, we will refer to the AGN class as *positive* and the galaxy class as *negative*. From the astrophysical perspective, the most common measures which describe catalog properties are its purity and completeness. The purity of the positive (AGN) class is known in machine learning as *precision* and is defined as

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

It shows what fraction of all objects classified as positive entities, i.e., true positives (T_p) and false positives (F_p) are the real observations from the positive class (T_p). Completeness of the AGN catalog is referred to as *recall* and is defined as

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

Recall defines the ratio of properly classified observations from a positive class (T_p) to the whole sample of positive class, true positives, and false negatives. In order to optimize both precision and recall during model training and to have an efficient way to compare the performance of various models, one may use more compound metrics, which are related to precision and recall. To be able to effectively control the model performance on the positive class, a popular and well established choice is the F1 score, which is a harmonic mean of precision and reacall:

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

Another metric tightly related to precision and recall is the *precision-recall area under curve* (PR AUC). The precision-recall curve is created by calculating precision and recall scores for different decision thresholds. Here comes an important limitation of this score. In order to define various class separating probability-based decision thresholds, one needs to be able to obtain probability estimations from a model. Thus it is only suitable for algorithms that allow for probability estimation. Then area under the precision-recall curve refers to the general effectiveness of the model without focusing on any specific threshold. The larger area under the PR-curve, the better the model is. The last metric used in this work is known as *balanced accuracy* (bACC) and is less connected to both precision and recall. It was used to highlight a general performance of the model, not focused specifically on a positive class. Balanced accuracy is defined as

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

The TNR in the equation refers to the *true negative rate* and is a precision counterpart for a negative class:

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

4.3 Outlier detection and high dimensional visualization

4.3.1 Outlier detection with Isolation Forest algorithm

In addition to supervised classification, this work also addresses the problem of unsupervised outlier detection. Making a simplification, we can define an *outlier* (or a *novelty*) as an observation that does not fit the general trends of the data distribution. Searching for this type of object may be challenging, specifically in the case of high-dimensional data representation or large data volumes. This problem is addressed by the branch of machine learning techniques created for outlier detection.

Most outlier detection algorithms follow the same general principle. First, they fit the data by constructing a profile of the typical observations and then try to find observations that do not fit this profile. In this work, we use a fundamentally different approach applied in the *Isolation Forest* algorithm (Liu, Ting, and Zhou, 2008). Instead of building a typical observation profile, it explicitly locates outliers in the data. This feature makes the Isolation Forest algorithm a very fast and scalable tool often used in modern machine learning pipelines.

The fundamental property of the Isolation Forest is based on a simple assumption that outliers are, in general, easier to separate from the rest of the data compared to typical objects. This property is capitalized in the following way. A forest of fully grown binary decision trees is created. During the growth of the tree, a node split is created by a random selection of splitting features and the corresponding threshold. To obtain an effective outlier detection, the degree of anomaly is related to the tree depth and relative position of the observation. In particular, a *path length* $h(x)$ is defined as the number of splits an object passes before reaching an external node of the tree. Observations are sorted according to corresponding path lengths, and objects with the shortest paths have a higher probability of being an outlier. A final anomaly score derived from $h(x)$ was described in the original paper as follows. Given the set of N observations, an anomaly score for observation x is defined as

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

where $\overline{h}(x)$ is the average path length obtained from the set of Isolation Trees and $c(n)$ is the average path length of unsuccessful search defined in Liu, Ting, and Zhou (2008). Such anomaly score is monotonic to path length. These two measures can occupy ranges of $0 < s(x,N) \leq 1$ and $0 < h(x) < N - 1$ respectively. Within these conditions authors define detection rules as follows. If an observation x returns $s(x,N) \simeq 1$, then x is identified as outlier. Observations with $s(x,N) \leq 0.5$ are identified as typical object. Finally if the whole sample is characterized by $s \simeq 0.5$, then there are no obvious outliers in the data.

4.3.2 tSNE algorithm

The Isolation Forest outlier detection mechanism, despite its high effectiveness, gives us limited information about the environmental context of a specific outlier. We do not know how this object is relative to other objects located possibly in different samples. In this work, we apply Isolation Forest in order to identify different sources of AGN catalog contamination, and in some cases, this type of context is crucial. Thus additional step of high-dimensional visualization in form of *t-distributed stochastic neighbor embedding* or *tSNE* (Hinton and Roweis, 2002; Maaten and Hinton, 2008).

The tSNE algorithm is a non-linear dimensionality reduction technique that is usually used for high-dimensional data visualization. There are two main steps in the tSNE algorithm. First, the joint probability distribution is constructed out of similarities between objects in the high-dimensional feature space. The similarity of objects in the basic version of the tSNE algorithm is the Euclidean distance between objects in the feature space. Next, the tSNE algorithm tries to learn low-dimensional embedding that can preserve similarities present in the high-dimensional representation. The tSNE algorithm does it via gradient descent minimization of Kullback-Leibler divergence between the high-dimensional representation of the data in the original feature space and low-dimensional embedding created by tSNE. Such minimal value corresponds to the objects position in the low-dimensional embedding.

5

Construction of Machine Learning Pipeline and Resultant catalogs

5.1 Construction of Machine Learning Pipeline

This section describes the general ML-based pipeline for AGN selection presented in this work and is further discussed in the following sections. The chart of main pipeline steps is shown in Fig. 5.1 and run as follows.

First, the data preparation procedures were performed. These included feature selection with KS-statistic method described in Sec. 5.2 and limiting the generalization sample using MCD algorithm described in Sec. 3.3.2. Thus training of the model was performed on the labeled data represented with the selected set of features. Generalization performed on the MCD-limited unlabeled data was performed using the same set of features.

During the training procedure, supervised classification models described in Chapter 4, logistic regression, SVM, extremely randomized trees, random forest, and boosted trees in the form of XGBoost implementation were created in several settings. Each model was tested in two main versions: the non-modified model and the model with class-balanced weights. Both of these versions were further modified based on the fuzzy logic (instance weights) strategies described in Sec. 5.3. Besides the non-modified version of the model, these were error-based and distance-based fuzzy logic models.

First, during training, hyperparameter tuning was performed for most models. Logistic regression, SVM, and XGBoost models were tuned using randomized grid searches with 1000 different hyperparameter combinations. This type of greedy search of hyperparameters combination allows one to find the variant of the model most suitable for a specific learning task. Tree ensemble methods were not tuned and left in the default versions present in the Scikit-learn library (Pedregosa et al., 2011) due to the very small sensitivity of these methods to hyperparameter tuning (Probst, Boulesteix, and Bischl, 2019). The optimal hyperparameter combination was searched via maximization of the F1 metric. Each time a specific hyperparameter combination was chosen, the F1 metric value was estimated via 100 shuffled 5-fold cross-validation train-test splits to minimize the risk of overfitting on the small data set. The same split technique was used on the final, best hyperparameter combination to estimate the rest of the metric scores and their uncertainties. Values or ranges of hyperparameters which were used to create a parameter grid, are presented in Tab. 5.1. Description of some of these parameters can be found in Sec. 4.1. The exception are XGBoost

Parameter	**Range**
	logistic regression
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]

TABLE 5.1: Grid of hyperparameter values used in grid search during model training. Some logistic regression and SVM parameters were sampled from a log-uniform distribution with fixed range.

parameters, some of which are connected to the specific numerical optimization. Full description of XGBoost hyperparameters can be found at XGBoost documentation webpage [1]. In the case of all tree-based algorithms, we used 500 trees to build an ensemble. In this case if the number of trees turned out to be too large, it will not cause an overfitting. Instead additional trees will simply not improve the model performance. Python codes used for the training are available on the GitHub page [2].

Next, a trained model with the best hyperparameter combination was used to make predictions on the labeled and generalization data sets. Due to the small amount of labeled data, a separate test set that would be held out during training and used only for predictions by the final model was not created. Instead, an approximate version of testing model predictions was performed. For this purpose, a model with fixed chosen hyperparameter combinations was run through additional 5-fold cross-validation. Predictions combined from five folds were used to estimate the model prediction of the labeled data. One needs to keep in mind when analyzing visualization of model predictions on the labeled data that this type of procedure slightly underestimates the model's performance. After prediction on the labeled data was finished, a model with fixed hyperparameters was fitted to the whole training data and used for prediction on the generalization sample to form an AGN candidates catalog. In addition to these methods, two voting classifiers were created out of the best models (see Sec. 5.4). The hard voting classifier, which used a majority vote out of predictions from the set of models, did not need to be trained. The stacked classifier, which used probability estimations of the best models, was trained in the same way as other models using probabilities as features.

Once the best model was selected, the obtained AGN catalog was analyzed with outlier detection methods described in Sec. 5.5. One method using the Isolation Forest

[1] https://xgboost.readthedocs.io/en/stable/parameter.html

[2] https://github.com/ArtemPoliszczuk/NEPWide_AGN

model was applied to select improperly estimated photometric redshifts. The second method, consisting of the Isolation Forest model combined with tSNE visualization, was used to find contaminants in the AGN candidates catalog.

5.2 Feature selection

Original representation of the data often is not optimal for ML applications. There are two main reasons. First, a very large number of features may result in a *curse of dimensionality*, (i.e., large sparsity of the data occurring in the high-dimensional space), reducing the effectiveness of the model performance. Second, a non-representative feature set may introduce significant noise to the data, leading to overfitting. To overcome this problem, different methods of *feature engineering* are used. A common approach consists of two steps: *feature creation*, where artificial features are created from the original ones via, e.g., feature ratios, multiplication, or difference, and *feature selection*, where the best subset from all of these features is chosen.

Feature selection methods can be divided into two main approaches. The first one is a general, problem-agnostic approach focused on feature space dimensionality reduction. This approach is exemplified by many popular approaches, such as PCA, which tends to reduce noise in the data representation without any reference to a defined ML problem. The second approach used in this work tends to find features most suitable for a specific prediction problem. In order to find a feature set, which will allow for effectively selecting AGNs in the data, a *Kolomogorov-Smirnov* statistic (KS-statistic) was used. The KS-statistic, in this case, is defined as the biggest distance between the empirical cumulative distribution function between AGN and galaxy training samples. The KS-statistic can be treated as a measure of the difference between the two distributions. A thus larger value of the KS-statistic calculated for a particular feature corresponds to the better usability of this feature for the AGN-galaxy separation. The KS-statistic was calculated for HSC optical bands (g, r, i, z, Y), AKARI NIR bands ($N2$, $N3$, $N4$) and all possible colors made of these eight filters. Fig. 5.2 shows KS-statistic values for the subset of best features. In order to avoid an increase in the data dimensionality, the number of final features was limited to the number of initial optical and NIR bands. This resulted in a feature set made of eight colors, using all available passband information. One can also observe evident high information contribution of the NIR colors to the AGN selection.

The KS-statistic has several pros and cons as a feature selection method. On the one hand, its simplicity is its main asset. It does not require a large amount of data, and it is prone to overfitting, as opposed to more complex, model-dependent wrapper methods (see, e.g., Jović, Brkić, and Bogunović, 2015). It can also capture differences in both position and shape of two distributions. On the other hand, this feature selection method focuses on single features and cannot capture interrelationships between them.

5.3 Fuzzy logic instance weights

In this work we used two types of instance weights (or fuzzy membership): distance- and error-based. In both cases instance weight s_i of i-th observation was normalized to the range [0, 1] (where larger value relates to a higher importance of an observation) and calculated via equation:

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

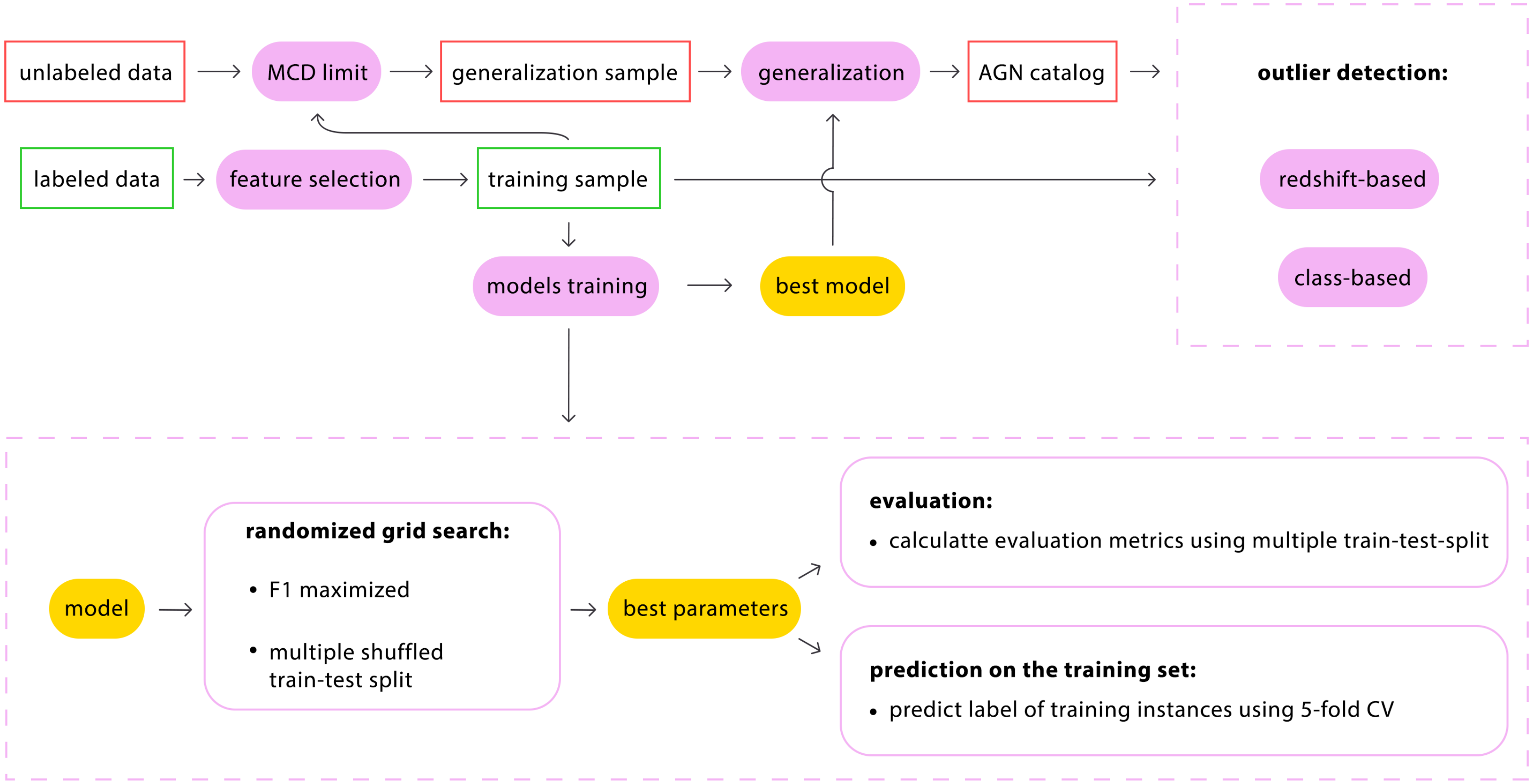


FIGURE 5.1: Machine learning pipeline scheme described in the present work. The upper part of the scheme shows the general outline, the lower part of the scheme, shown in the violet rectangle, refers specifically to the training process.

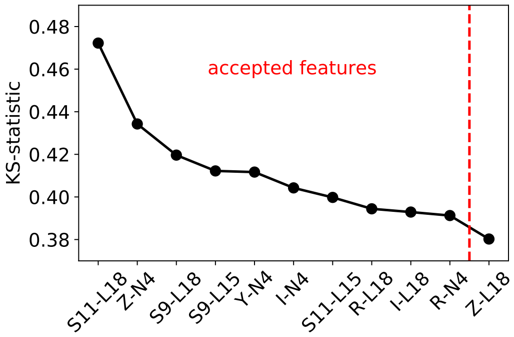
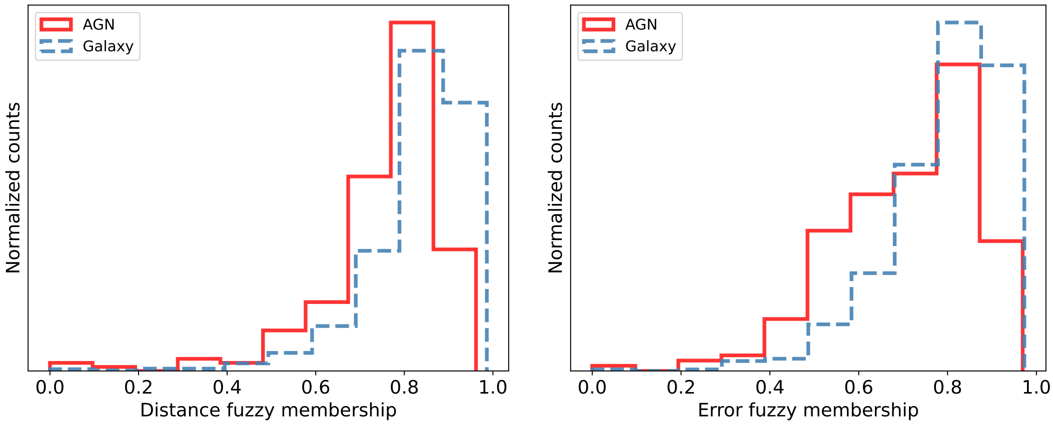


FIGURE 5.2: Results of the main feature selection via Kolomogorov-Smirnov statistic method used in this work. Only a subset of features with the highest KS score is shown.

where u_i is a quantity characteristic for a specific type of instance weight calculated for i-th observation in the training set, $u_{\max}$ is a maximal u_i value in the sample and δ is a a small value used to avoid division by zero. Because the importance of particular observations in the training sample is affected only by the relative difference between fuzzy memberships, the presence of δ has no impact on the training process and serves only numerical safety purposes. In this work δ parameter was set to 10^{-4}.

The idea of fuzzy membership based on distance from the class center as well as the definition of s_i as in Eq. 5.1 was in-depth described in Lin and Wang (2002), where it was applied to the SVM algorithm. However, instance weights might be also applied to other types of learning algorithms (see Sec. 4.1). In the case of distance weights, u_i is a Euclidean distance from the center of a class, defined in the feature space, and is used as a form of instance weight. Here the purpose of such a weight is to minimize the impact of outliers on the training process. In order to properly construct such weights, they were calculated separately for AGN and galaxy classes. This approach is dictated by the different distribution of classes in the feature space and the class imbalance, where the latter causes a systematic underestimation of the smaller class importance. Histograms of a distance-based fuzzy membership distribution for both classes are shown in Fig. 5.3a. Here we see that a large number of outliers characterizes the AGN class. Thus distance-based fuzzy memberships should provide us with a more conservative classification characterized by a higher purity and lower completeness trade-off.

The error-based fuzzy membership approach follows the same principles but focuses on the impact of measurement uncertainty. This approach was first discussed in our preliminary work Poliszczuk et al. (2019) and further developed together with distance-based weights in Poliszczuk et al. (2021). In this approach, u_i is defined as a sum of absolute values of measurement uncertainties calculated for optical and NIR passbands. Here a specific systematic bias was allowed. Mainly, optical bands are characterized by a smaller measurement uncertainty due to the higher accuracy of the measurement and specificity of SUBARU/HSC pipeline, which tends to cause magnitude error underestimation. Thus the accuracy of measurement in AKARI passbands dominates the final outcome of s_i. We avoided rescaling and unifying measurement errors for a practical reason: NIR passbands have the main informational contribution to the AGN selection. Thus, they should be treated as more important. Histograms of an error-based fuzzy membership distribution for both classes are shown in Fig. 5.3a. Again, we see a larger impact of instance weights on the AGN class.



(A)

(B)

FIGURE 5.3: Histograms of different types of fuzzy memberships. *Panel a:* Distance-based fuzzy membership. *Panel b:* Error-based fuzzy membership.

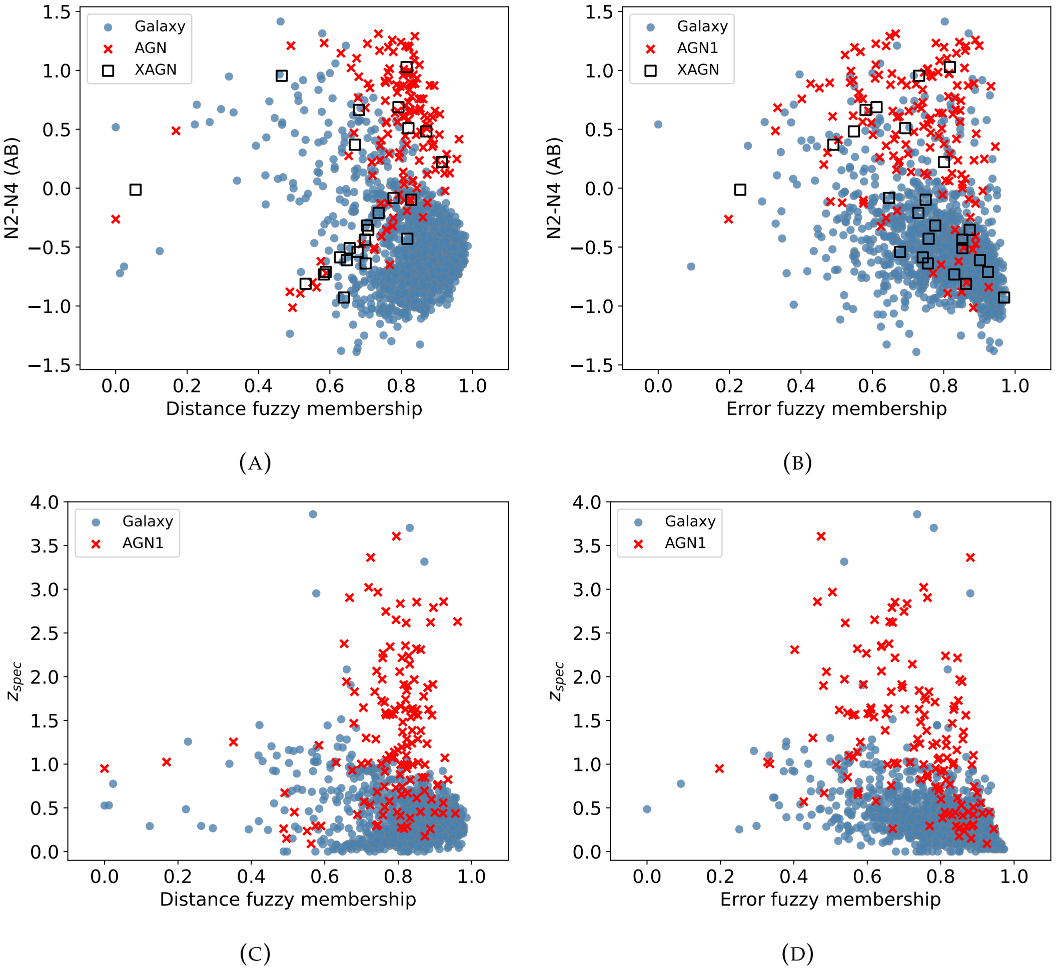
Further comparison between two fuzzy-weight strategies is presented in Fig. 5.4. In particular, Fig. 5.4a shows significant tendencies created by distance-based fuzzy memberships. Specifically, one can observe an importance decrease of high-redshift SFG, characterized by red $N2$–$N4$ color, as well as the lower impact of XAGN as well as other AGNs characterized by blue colors. On the other hand, error-based fuzzy membership shows a general, strongly scattered tendency to decrease the importance of red $N2$–$N4$ objects. The relation between distance-based fuzzy membership and redshift distribution of the training sample is shown in Fig. 5.4c, and no significant dependency can be observed. Analogous relation for error-based fuzzy membership is shown in Fig.5.4d. Here, one can see a weak correlation, which is caused by the fact that more distant objects are dimmer in general.

5.4 Performance evaluation

5.4.1 Impact of class-balanced weights on model performance

The result of different models training described in Sec. 3.3.2 is evaluated via comparison of several metrics: F1 metric, precision, recall, area under the precision-recall curve (PR AUC), and balanced accuracy (bACC). All of these metrics were described in detail in Sec. 4.2.

General analysis of model performance is presented in Fig. 5.5. Here we see metric values for various non-balanced and class-balanced models. Since the class balance is considered a more fundamental change in the model than a more subtle instance weighting (fuzzy logic), in this first step of model evaluation analysis, instance weighted models were not used. There are three main goals at this stage of the analysis: find a subset of the best models to create voting classifiers, analyze the impact of the class balance on the model performance and find, if possible, the best model. Fig. 5.5a shows set of all classifiers together with so-called *dummy classifier*. The dummy classifier assigns class labels at random, where the sizes of predicted class samples correspond to the fraction of classes in the training data. The performance of this basic model serves as the lowest boundary in our analysis, i.e., if the model was able to learn the classification task in the training process, it should show higher



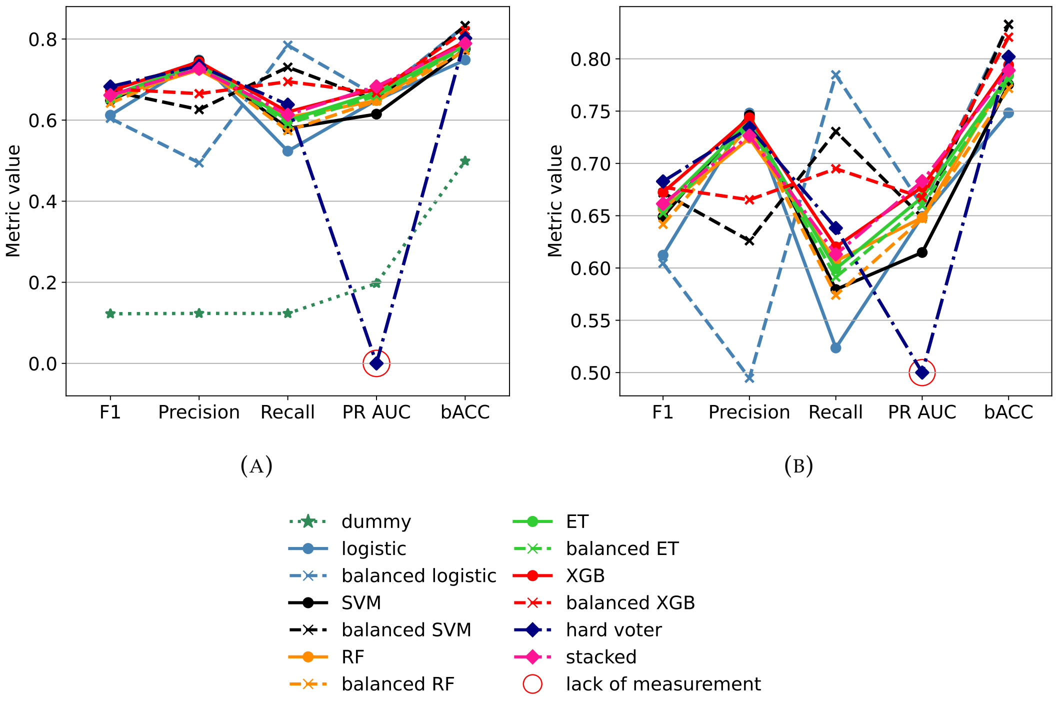
(A)

(B)

(C)

(D)

FIGURE 5.4: Impact of different fuzzy memberships (instance weights) on properties of the training sample. *Panel A:* Impact of distance-based fuzzy membership on the $N2$–$N4$ color distribution. *Panel B:* Impact of error-based fuzzy membership on the $N2$–$N4$ color distribution. *Panel C:* Impact of distance-based fuzzy membership on the redshift distribution. *Panel D:* Impact of error-based fuzzy membership on the redshift distribution.



(A)

(B)

(C)

FIGURE 5.5: Performance evaluation for different classification models. Only models with no fuzzy logic application and stacked classifiers are presented. *Panel A:* Evaluation metrics for different models compared with dummy classifier. *Panel B:* Evaluation metrics for different models. Dummy classifier is not included *Panel C:* Legend.

metric scores than the dummy classifier. Based on the comparison with the dummy model, Fig. 5.5a shows that all of the presented models can learn the classification task effectively. The lack of PR AUC score for the hard voting classifier comes from the fact that construction of the precision-recall curve requires the usage of classification probabilities, which are not present in the case of the majority vote created by the hard voting classifier.

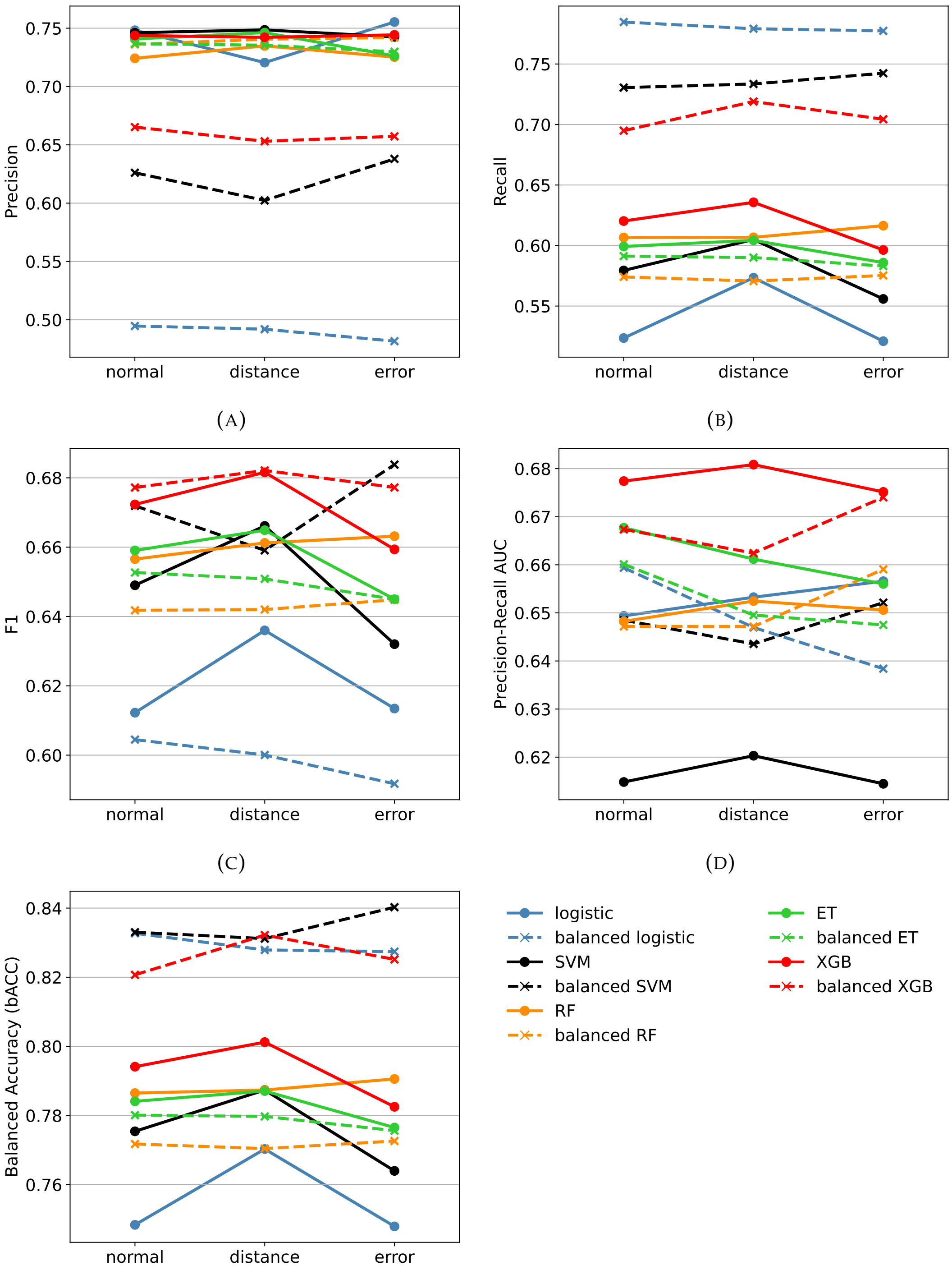
Fig. 5.5b shows same results but zoomed in and without dummy classifier. Here we see several interesting tendencies. First, there is present a split between two types of impact that the class balance can have on the classifier performance. Moreover, this split, prominent in precision to recall ratio, is not dependent on the general class of the algorithm. One class of the behavior represented by logistic regression, SVM, and XGBoost (i.e., linear, non-linear, and boosted tree methods) shows a significant increase in the recall with a simultaneous drop in the precision when the class balance is applied. This impact of class-balance weights can be interpreted intuitively. In a non-balanced case, the separation between class regions is pushed towards a smaller class (i.e., AGNs or Positive class). Thus, a smaller class label is assigned to objects in a very close neighborhood of the smaller class training sample. This results in prediction, which promotes classification as a positive class object only for very typical observations, which lie close to the class center, producing a high precision and low recall catalog. Application of the class weight moves classification boundaries in the feature space away from the smaller class region, resulting in the recall increase followed by the higher contamination of the positive class catalog (i.e., lower precision). The rest of the models, i.e., extremely randomized trees and random forest, does not show any significant impact of class weight on the model performance. This result cannot be explained by the properties of the tree ensemble models and might be problem-specific since class weighting is known to boost ensemble tree model performance (see, e.g., Chen, 2004).

Both voting classifiers, i.e., a hard voting classifier and a stacked classifier, were created from the subset of best classifiers. Groups of models for this subset were selected by the analysis of the non-balanced and class-balanced models without instance weighting. In other words, if, for example, some specific class-balanced model was chosen as an input for a voting classifier, then all instance weight realizations: no instance weight, distance-based weights, and error-based weights were used. It was done because class weights are considered a more fundamental weighting method, while instance weights are more subtle and play the role of the final tuning. Models selection for voting classifiers was done based on metric values comparison presented in Fig. 5.5. In order to reject some of the classifiers, models with significantly lower values of any of the metrics were chosen. We removed non-balanced logistic regression due to the very low precision, class-balanced regression due to the very low recall, and non-balanced SVM due to the low PR AUC value. Class-balanced SVM and XGBoost were left in the final set to obtain diversity in the performance of particular models despite low precision values. Hard voting classifier and stacked classifier show tendencies in the metric values similar to non-balanced. This departure from the class-balanced behavior comes from the fact that, in some cases of instance weight strategies bring class-balanced models closer to the non-balanced cases (this phenomenon will be discussed later in this Chapter).

The hard voting classifier was chosen as the best final model. This decision has several reasons. At this stage, the comparison of the model performance must expand beyond the analysis of the F1 score, which was the main metric during the training process. The small amount of the training data makes it difficult to estimate metric values precisely, and conclusions from obtained scores should be drawn with caution. This uncertainty, as well known contamination from the high-redshift SFG, led to the shift towards models with higher precision and lower completeness during the analysis of the precision-recall trade-off. The best results were shown by the non-balanced fuzzy-distance XGBoost and the hard voting classifier, where the differences between metric values were situated within the uncertainty limit (exact values of metrics and corresponding uncertainties for all models can be found in the Appendix **??**, Tab. **??**). Again, the hard voting classifier was chosen out of these two models due to the limited control of the classifier performance caused by the small amount of the labeled data. The hard voting classifier, due to its ensemble nature, allows us to reduce the variance of the model further and avoid the risk of overfitting. However, the simplicity and effectiveness of this model were obtained at the expense of the lack of classification probability estimation. As a result, the final model was characterized by the 0.73 precision and 0.64 recall values, which correspond to the estimation of purity and completeness of the AGN catalog respectfully. The generalization performed on the unlabeled data gave us the catalog of 465 AGN candidates, which constitutes 1.4% of the whole generalization sample. More information on values of evaluation metrics for different models can be found in Appendix B in Tab. B.1 and B.2.

5.4.2 Impact of the fuzzy logic on model performance

Before we dive into the analysis of the properties of the final classifier and obtained AGN catalog, let us investigate the impact of different fuzzy logic strategies. Fig. 5.6



(A)

(B)

(C)

(D)

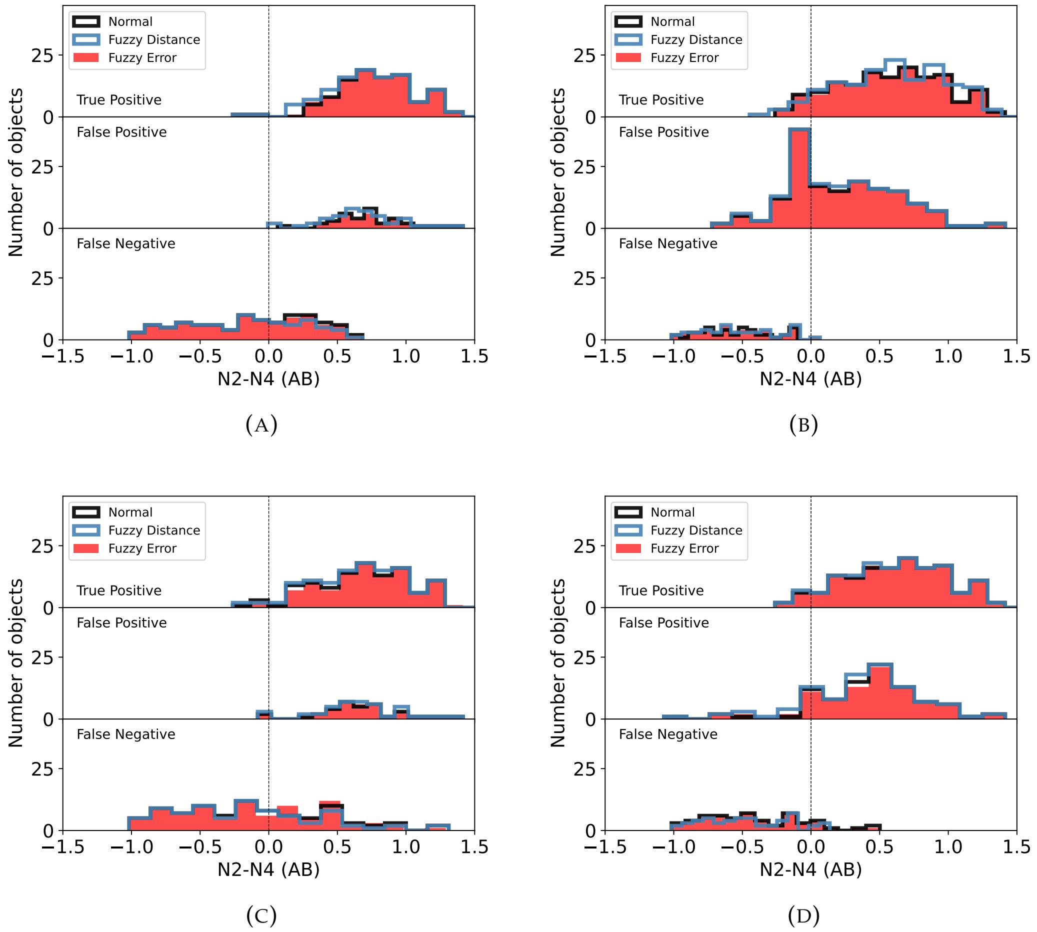
(E)

(F)

FIGURE 5.6: Visualization of the evaluation metric values for different fuzzy membership strategies. *Panel A:* Precision. *Panel B:* Recall. *Panel C:* F1 score *Panel D:* Area under Precision-Recall curve. *Panel E:* Balanced Accuracy.

shows metric values for different fuzzy logic strategies. Fig. 5.6a and Fig. 5.6b show values of precision and recall respectfully. Since both of these metrics are the basis for more complex scores of F1, PR AUC, and partially bACC we will start our analysis from them. Application of the fuzzy logic, in general, does not affect the precision of the model, and similar values are preserved for all cases of instance weighting. Two exceptions are the class-balanced logistic regression and class-balanced SVM, which show significant differences between fuzzy strategies and the opposite impact of fuzzy membership on the non-balanced and class-balanced cases. To a much smaller extent, this tendency is visible also in the case of the XGBoost. Thus we can observe the separation between models into two different categories of behavior, which overlapped with general differences between models present in Fig. 5.5 and discussed above. This separation is present not only in the comparison of non-balanced and class-balanced cases but it is also preserved in different fuzzy logic strategies. Comparison of the recall metric values shown in Fig. 5.6b gives us a very different picture. Here we see that fuzzy logic impacts the non-balanced algorithms, leaving class-balanced models relatively intact. Moreover, the distance-based fuzzy logic application gives better results than the application of error-based fuzzy logic or lack of instance weights. Tendencies of the precision and recall translate into behavior observed in the F1 (Fig. 5.6c) and bACC (Fig. 5.6e) scores. The PR AUC score, shown in Fig. 5.6d, often exhibits different tendencies, implying a significant difference one would obtain if PR AUC score was used instead of F1 during the hyperparameter tuning procedure.

A general tendency one can observe during the analysis of metric values is the model performance improvement with distance-based fuzzy logic and a relatively subtle difference between error-based fuzzy models and models with no instance weights. A significant increase in the recall value for non-balanced models with distance-based fuzzy logic with no significant change in precision and lack of such tendency in the class-balanced models has an intuitive explanation. The distance-based fuzzy logic reduces the importance of outliers in the training process with a simultaneous increase of the typical objects' importance. Thus the importance of the training AGNs lying in the galaxy locus becomes lower. This change does not significantly affect the completeness of the catalog because galaxies predominantly occupy these regions of the feature space. At the same time purity of the catalog increase because the probability of selection of AGN candidate far away from the center of the AGN class become lower. An increase in the typical objects' importance increases the class separation boundaries outwards class center. Thus AGN catalog completeness will increase in the region occupied mainly by AGNs in the training set. Additionally, one may observe a decrease in the contamination of the AGN catalog in the region occupied mainly by AGNs due to the decrease in the high-redshift SFG importance caused by their low distance-based weights. To summarize, we can see a clear tendency in the AGN catalog completeness (recall) growth and opposite mechanisms that may increase or reduce AGN catalog purity (precision). As a result, we have a distance-based fuzzy logic model with a higher recall and similar precision compared to the initial model with no instance weights. We see that fuzzy logic has the most impact on the non-balanced models, where the class sizes are not even, and decision boundaries are pushed towards smaller classes. Here distance-based weight can gain their accumulative impact on the classification. However, in the case of class-balanced models, the decision boundary is already pushed away from the smaller class by the class weights. The main impact distance weights can have on the class-balanced models is the reduction of the importance for outliers lying in the other class locus.



(A)

(B)

(C)

(D)

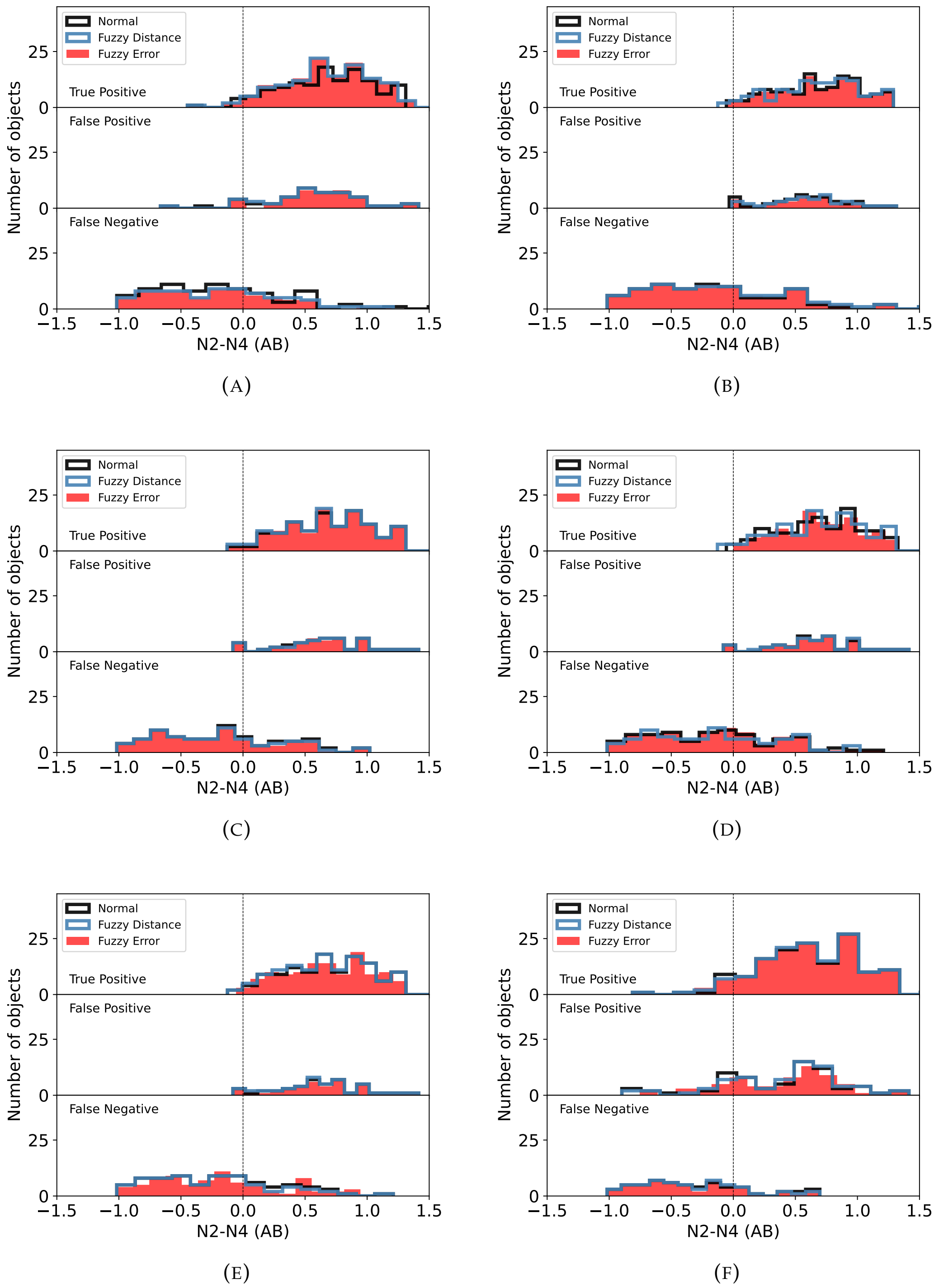
FIGURE 5.7: The *N*2–*N*4 color distribution presents the impact of different fuzzy memberships (instance weights) on the prediction performed on labeled data. The positive class refers to the AGN class, and the negative class refers to the galaxy class. True Positive objects are properly classified AGNs. False Positive objects are misclassified galaxies, i.e., contamination of the AGN catalog. False Negative objects are the AGNs misclassified as galaxies. *Panel A:* Non-balanced logistic regression. *Panel B:* Class-balanced logistic regression. *Panel C:* Non-balanced SVM *Panel D:* Class-balanced SVM

The error-based fuzzy logic does not have, in general, a significant impact on the classification performance in both class-balanced and non-balanced models. Despite physical motivation lying behind error-based fuzzy logic, it can cause only a minor change in the classifier's performance. This phenomenon is most probably, caused by the low correlation between a specific problem and measurement uncertainty, i.e., observations that should be treated as the most important due to their position in the feature space may not be characterized by the highest accuracy of the measurement. Thus this type of fuzzy logic may cause two contradictory behaviors. On the one hand, objects that were measured poorly and, as a consequence, were shifted into a different part of the feature space would not be a major issue because error-based weights reduce their impact on the classification. On the other hand, poorly measured objects that are crucial for the classification and are located within a proper feature space volume would not be able to give a model enough information.

Now let us analyze how does the the fuzzy logic impact the $N2$–$N4$ color prediction distribution presented in Figures 5.7 and 5.8. In all the presented cases, we can see the hierarchical nature of the class and instance weights, i.e., the class weights are more fundamental and have a major impact on the color distribution, while instance weights should be treated as a fine-tuning technique. In the case of logistic regression, SVM, and XGBoost, the application of class-weights has several common implications, all of which are connected to the shift of the decision boundary towards the blue $N2$–$N4$ color. First, we see that class weights are causing a shift of the true positive (properly selected AGNs) towards the blue $N2$–$N4$ color. As was shown previously, the blue $N2$–$N4$ color range is mainly occupied by galaxies. Thus increasing weights of the smaller class observation shifts the separation between classes outwards from the smaller (AGN) class. Moreover, an increase of the AGN importance results in the treatment of AGN class locus (i.e., red $N2$–$N4$ color range) as occupied almost exclusively by AGNs with the decrease of the high-redshift SFG impact on classification. Consequently, in class-balanced models, one can observe a significant increase of the AGN catalog contamination in the red $N2$–$N4$ color range, together with a decrease in the number of False Negatives (i.e., AGNs classified as galaxies). The only models where we cannot observe a significant impact of class weights are the random forest and extremely randomized trees, as it was previously discussed. The fuzzy logic impact analysis does not show any significant changes in the $N2$–$N4$ color distribution. However, when making conclusions about the minor impact of fuzzy weights, one must keep in mind that $N2$–$N4$ color, despite its usefulness, is not a fully representative feature, and more significant changes might occur in the high dimensional feature space.

5.4.3 Final model and AGN candidates catalog

Now let us focus on the final model performance, i.e., the hard voting classifier. A familiar $N4$ vs $N2$–$N4$ magnitude-color plot for prediction on the labeled data as well as obtained AGN candidates catalog is shown in Fig. 5.9. We can observe tendencies present here in the constituent models described previously. One of such prominent tendencies is that AGNs are predominantly selected in the red part of the $N2$–$N4$ color. Two factors cause this phenomenon. Firstly we used a higher precision score as an important property of the classifier when comparing models via precision to score ratio. Thus the final model has a lowered tendency to select AGN candidates in the regions of the feature space dominated by galaxies (e.g. blue part of the $N2$–$N4$ color distribution). The training sample AGN observations located in the blue $N2$–$N4$ color are mainly X-ray selected AGNs. As discussed in the previous parts of this work, X-ray-selected AGNs are often hard to recover using IR or optical selection. The main reason for this difficulty comes from the fact that IR (and optical) AGN selection probes only the high value end of the L/L_{Edd} ratio distribution, while the X-ray selection covers most of the distribution range. The main contamination of the catalog comes from the SFG located at relatively high redshifts (see Fig. 3.1b) characterized by a predominantly red $N2$–$N4$ color. A closer look at the redshift distribution of different prediction subsamples presented in Fig. 5.10 shows us three interesting characteristics of the classification model. First, we see a significant difference between the True Positive (properly selected AGNs) and False Negative (AGNs wrongly classified as Galaxies) redshift distributions. It shows us that the main compound of the AGN distribution that escapes our selection comes from the low redshift universe. Another important observation pertains to the redshift properties of the False Positive observations (i.e., galaxies classified as AGNs). We



(A)

(B)

(C)

(D)

(E)

(F)

FIGURE 5.8: The *N*2–*N*4 color distribution presents the impact of different fuzzy memberships (instance weights) on the prediction performed on labeled data. The positive class refers to the AGN class, and the negative class refers to the galaxy class. True Positive objects are properly classified AGNs. False Positive objects are misclassified galaxies, i.e., contamination of the AGN catalog. False Negative objects are the AGNs misclassified as galaxies. *Panel A:* Non-balanced random forest. *Panel B:* Class-balanced random forest. *Panel C:* Non-balanced extremely randomized trees. *Panel D:* Class-balanced extremely randomized trees. *Panel E:* Non-balanced XGBoost. *Panel F:* Class-balanced XGBoost.

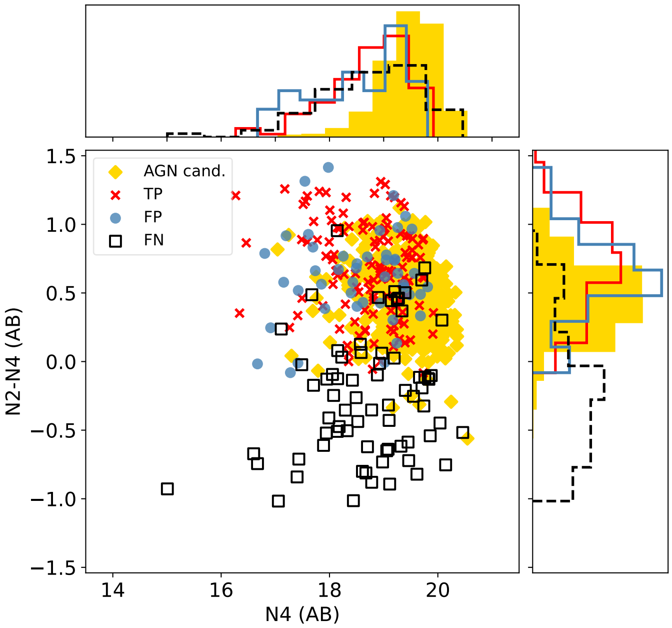


FIGURE 5.9: Color-magnitude plot of $N2$–$N4$ vs. $N4$, together with density histograms of corresponding color and magnitude showing predictions of the final model on labeled and generalization sets. True Positive (TP, red crosses) refers to properly classified AGNs in the labeled dataset. False Positive (FP, blue dots) refers to galaxies incorrectly classified as AGNs. False Negative (FN, black squares) refers to AGNs incorrectly classified as galaxies. AGN candidates, denoted by yellow rhombs, refer to observations from the generalization sample classified as AGNs. Colors on normalized histograms correspond to colors on the color-magnitude plot.

see that contamination comes not only from SFG at higher redshifts but also from low-redshift galaxies characterized most probably by the significant dust component. Finally, a comparison between the spectroscopic redshift of True Positives and the photometric redshift of AGN candidates' distributions shows us similar properties of these two samples. Small shift of the AGN candidates redshift distribution towards smaller redshift values is most likely caused by the systematic underestimation of the AGN photometric redshifts above ∼1.5. Thus, the real redshift distribution of the AGN candidates sample probably covers a much wider range. Another sign of the presence of high-redshift AGNs in the AGN candidates catalog are also present in Fig. 5.9 if we compare it to Fig. 3.1b. Here we can observe a drop of the $N2$–$N4$ color value at the highest redshifts. In addition, high-redshift AGNs, in general, appear to be dimmer. Combining these two facts, we will see a subset of objects characterized by high $N4$ values and $N2$–$N4 \in (0, 0.5)$. The training data sparsely represent this region occupied by AGN candidates. Their absence in the training sample comes mainly from the conditions the training object had to fulfill to be observed by a spectrograph. Objects characterized by such properties are possible candidates for high-redshift AGNs. Statistical properties of the final AGN candidates catalog are presented in Tab. 5.2.

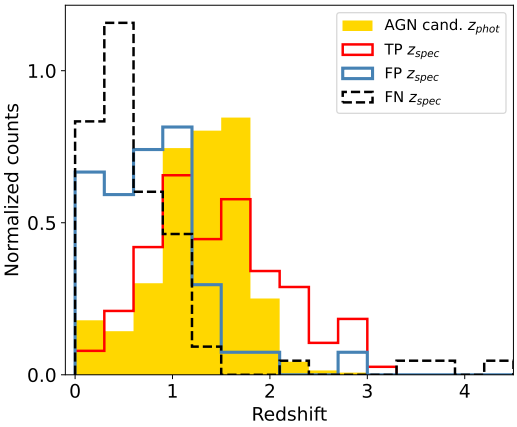


FIGURE 5.10: Histogram of redshift distribution with respect to the prediction results of the best, final model on the labeled and generalization sets. True Postive (TP, red color) refers to properly classified AGNs in the labeled dataset. False Positive (FP, blue color) refers to galaxies incorrectly classified as AGNs. False Negative (FN, black color) refers to AGNs incorrectly classified as galaxies. AGN candidates, denoted by the yellow color refer to observations from the generalization sample classified as AGNs.

band	median	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

TABLE 5.2: Statistical properties of the final catalog of AGN candidates. Median, median absolute deviation (MAD), and minimal and maximal limits are shown.

5.4.4 Extrapolation experiment

One of the prominent features of the classification performed by the hard voting model was the difficulty of AGN selection in the region of the feature space characterized by the blue $N2$–$N4$ color. Nature of these objects and difficulties of selection using optical, NIR and MIR passbands was already discussed. In order to test if modern ML techniques, which operate in complex, high-dimensional spaces, can overcome this problem, a supplementary experiment was performed. We will refer to this experiment as the *extrapolation experiment* and previous classification will be referred to as the *main classification*. In order to increase amount of information accessible to the model, we added measurements from MIR passbands: $S7$, $S9W$, $S11$, $L15$ and $L18W$ (we did not use $L24$ due to very small number of detections in this passband).

A training sample was modified in order to create a model focused on the problematic cases from the main classification. The observations classified by the hard

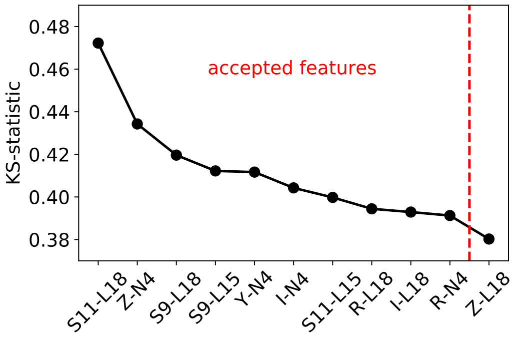
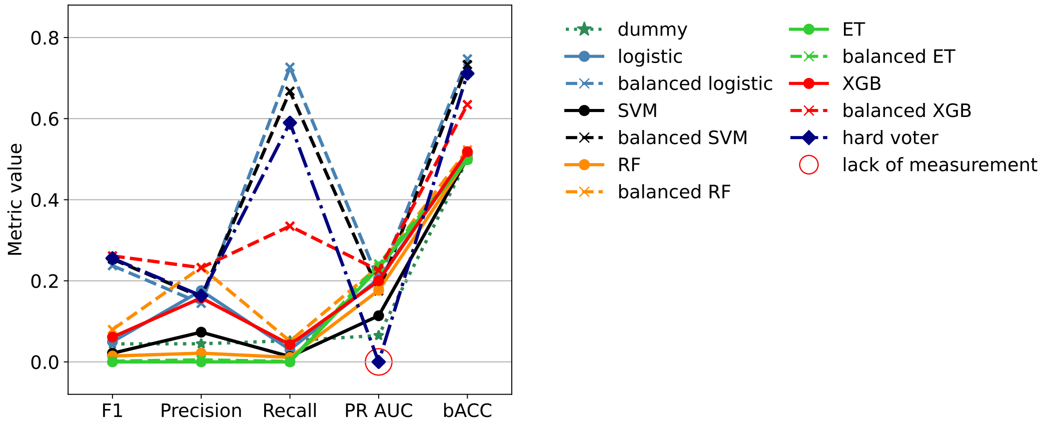


FIGURE 5.11: Results of the feature selection via Kolomogorov-Smirnov statistic method used in the extrapolation experiment. Only a subset of features with the highest KS score is shown.

voter during main classification as false negatives (i.e. wrongly classified AGNs) were used as the new AGN training sample. The galaxy training sample remained unchanged. This modification together with limits imposed by the requirement of MIR detection gave us a final training sample made of 705 galaxies and 39 AGNs. The galaxy training sample was not modified into subsample of the false positives observations in order to do not further reduce training sample size. The new generalization sample was made out of the main classification generalization sample, by adding a simple requirement of the MIR detection. We did not apply a new MCD limit on the unlabeled data in order to be able to compare generalization results from the main classification and the extrapolation experiment. This way we obtained a generalization sample made of 2207 observations.

Feature selection was based on the same method as previously. The KS-statistic was used to choose best features for this specific classification task. Values of the KS-statistic are shown in Fig. 5.11. A subjective decision was made to select ten features with the highers KS values as the final feature set. Because some of filters were present only in features with low KS-statistic value, previously applied requirement of using all filters in the final feature set was omitted. It is worth noting that many of the selected features in the extrapolation experiment are made of MIR filters.

Performance evaluation of the extrapolation experiment is shown in the visual form in Fig. 5.12 with detailed values presented in Appendix B in Tab. B.3 and B.4. Here one can see that only some of the models were able to learn the classification task. Both, non-balanced random forest and SVM showed metric results with uncertainties within the range of dummy classifier performance, or even more catastrophic results in the case of non-balanced and class-balanced extremely randomized trees. Two of the remaining models, class-balanced random forest and non-balanced logistic regression, were able to obtain a significant increase in the precision value, but could not overcome problems with low F1 and recall values. Due to this issue, a small subset of the best models: class-balanced SVM, logistic regression and XGBoost, were used to build an additional hard voting classifier. Due to the higher risk of overfitting caused by a small amount of data, the stacked classifier was not used in the extrapolation experiment. It is worth to note that in the extrapolation experiment all of the best models are using class weights. It is an opposite tendency to what we were observing in the main classification. Comparison between training sample AGN to galaxy class ratios in the main classification and extrapolation experiment show us a change from $\sim$ 15% to $\sim$ 5%. Thus it is probable that class weights in this classification task become an important part of the model in the case of a very strong



(A) (B)

FIGURE 5.12: Performance evaluation for different classification models for the extrapolation experiment. Only models with no fuzzy logic application and stacked classifiers are presented. *Panel A:* Evaluation metrics for different models compared with dummy classifier. *Panel B:* Legend.

class imbalance (i.e. when one the classes constitutes only few percent of the other).

Due to a very strong contamination of AGN catalogs produced by all of the remaining models, a high precision score became a main criterion of the best model selection. The highest precision value was obtained by the class-balanced XGBoost and the class-balanced random forest. Out of these two models, the XGBoost was characterized by a much better results in the F1 score, recall and balanced accuracy. Thus, the class-balanced XGBoost, characterized by the 0.25±0.11 precision and 0.37±0.16 recall, was chosen as the final classifier.

Initial generalization performed by the final model gave us catalog of 354 AGN candidates. The main idea of the extrapolation experiment was to overcome the problem of AGN selection in the blue $N2$–$N4$ color range. We did not impose the $N2$–$N4 < 0$ condition on the generalization set right at the beginning of the extrapolation experiment construction, due to the possibility of a very precision in the red $N2$–$N4$ color range once MIR information was added to the optical and NIR data. Fig.5.13 shows us, however, a presence of a strong contamination of the AGN catalog in the red color range. Thus, the only useful part of the sample produced during generalization procedure, that would be supplementary to the main classification catalog, were objects characterized by the $N2$–$N4 < 0$ condition. This way we obtained catalog of 198 objects (∼9% of the extrapolation experiment generalization sample). Beside a high contamination of the AGN catalog in the red $N2$–$N4$ color range, we still can observe the same problem with loss of AGNs in the blue part of this color. These are, again, mainly X-ray selected AGNs located at low redshifts (see Fig. 5.14). Comparision between Fig. 5.13 and Fig. 3.1a gives us some idea about the capabilities of this classification model - it can recover some of AGNs located in the galaxy locus, i.e. probably objects with significant host component at the expense of a very significant AGN catalog contamination. However, even with a very low AGN catalog purity, classifier created in the extrapolation experiment cannot recover X-ray selected AGNs. It proves once again a fundamental difference between X-ray and optical-IR selection of AGNs. Due to a very high level of contamination of AGN catalog (i.e. low precision), its low completeness (i.e. low recall) and mentioned above inability to recover XAGNs, the extrapolation experiment is treated in this work only

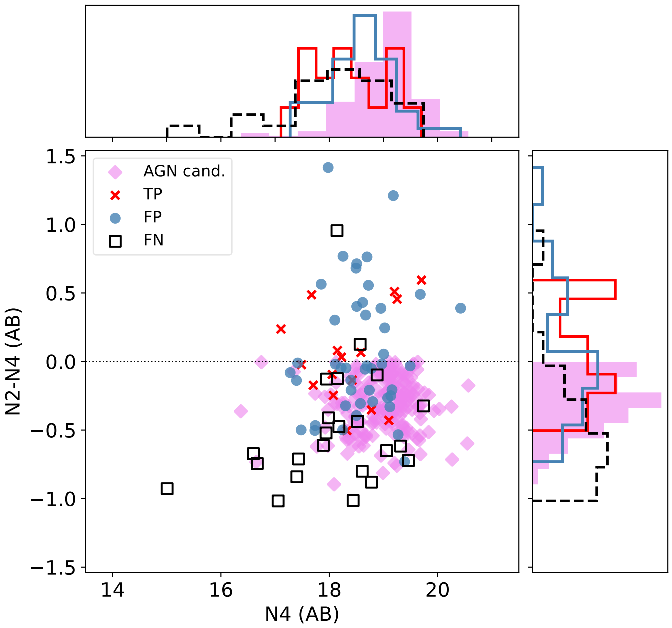


FIGURE 5.13: Color-magnitude plot of $N2$–$N4$ vs $N4$, togheter with density histograms of corresponding color and magnitude showing predictions of the extrapolation experiment model on labeled and generalization sets. True Postive (TP, red crosses) refers to properly classified AGNs in the labeled dataset. False Positive (FP, blue dots) refers to galaxies incorrectly classified as AGNs. False Negative (FN, black squares) refers to AGNs incorrectly classified as galaxies. AGN candidates, denoted by purple rhombs refer to observations from the generalization sample classified as AGNs. Colors on histograms correspond to colors on the color-magnitude plot.

as a way to determine limits of the ML-based AGN selection. Thus AGN candidates selected during this experiment were not included into the final AGN catalog.

5.4.5 Comparison with MIR-based AGN selection

Let us compare new ML-based method for AGN selection with the original MIR-based technique. We assume some common properties between two selection methods due to the training sample construction, used in this work. The AGN target preselection for the main spectroscopic follow-up (Shim et al., 2013) which were included into the training sample were based on the AKARI MIR selection technique used in Lee et al. (2007). These AGN objects constitute a majority of AGNs present in the training data (additional AGNs come from X-ray selected sample and few spectroscopic observations coming from other follow-ups).

In the original MIR method, the color cut AGN selection was performed with an additional $S11 < 18.5$mag limit imposed on the sample. This limit was created in order to fit a method to characteristics of the AKARI NEP-Deep catalog. In the case of data used in the present work, only a very small number of objects overpass this limit. These are two X-ray AGNs and four type-I AGNs in the training sample and nineteen AGN candidates in the final catalog. Due the small number of these objects,

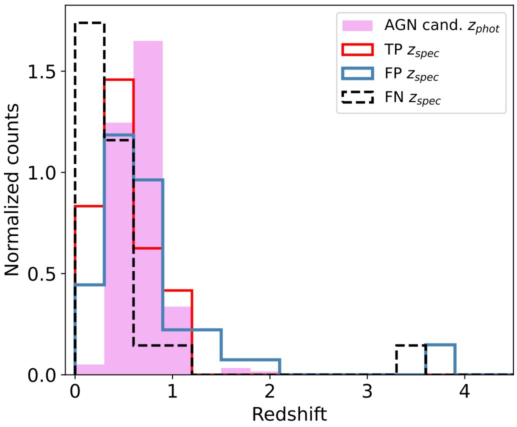
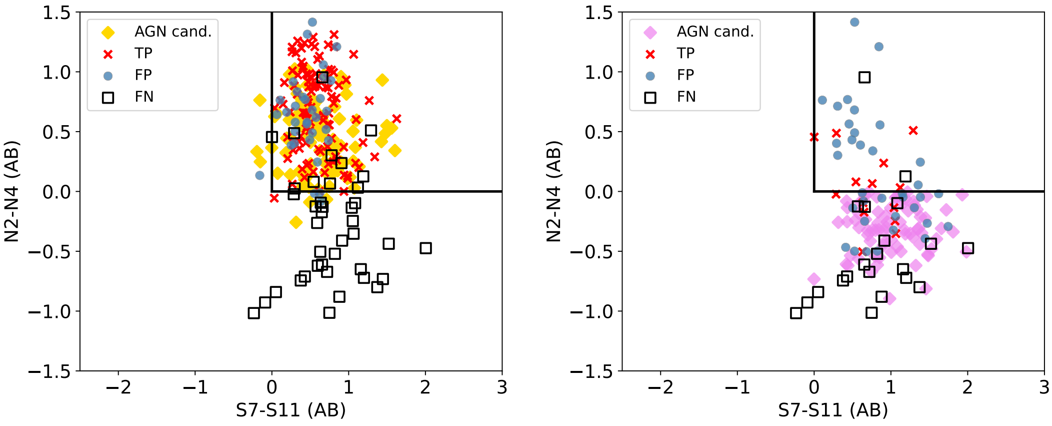


FIGURE 5.14: Histogram of redshift distribution with respect to the prediction results of the extrapolation experiment model on the labeled and generalization sets. True Postive (TP, red color) refers to properly classified AGNs in the labeled dataset. False Positive (FP, blue color) refers to galaxies incorrectly classified as AGNs. False Negative (FN, black color) refers to AGNs incorrectly classified as galaxies. AGN candidates, denoted by the purple color refer to observations from the generalization sample classified as AGNs.

fact that this work is performed on the AKARI NEP-Wide catalog (and not on the AKARI NEP-Deep) and in order to have a more straightforward comparison between two selection methods, we decided to omit the $S11$ limit.

To make a comparison between an ML-based method and MIR-based color selection, training sample and obtained catalog of AGN candidates were limited to objects detected in $S7$ and $S11$ passbands. This additional condition prune training and AGN candidates catalog samples to 815 (out of 1547 initial training observations) and 113 (out of 465 initial AGN candidates) objects respectively. In the case of extrapolation experiment, we did not further modify the data, since $S7$ and $S11$ detection conditions were already imposed during creation of the training and generalization samples.

Figure 5.15 shows visual comparison between MIR-based AGN selection method and ML-based methods from the main classification (Fig. 5.15a) and the extrapolation experiment (Fig. 5.15b). The MIR color-cut method is based on the power-law shape of the AGN SED in the NIR and MIR range. Such SED produces red NIR and MIR colors placing AGN locus in the $N2$–$N4 > 0$ and $S7$–$S11 > 0$ square on the color-color plot. Figure 5.15a compares MIR-based color method with the prediction on the training data and generalization performed by the final model from the main classification. We see a significant similarity between the two methods. The vast majority of AGN candidates selected in the main classification occupy upper right square (red $N2$–$N4$ and $S7$–$S11$ colors), which is used by the MIR-based method to select AGN candidates. Analysis of the classification performed on the labeled data shows us confirmation of tendencies already discussed in Sec. 5.4.3. First, we see an inevitable contamination of the IR-selected AGN catalog by the SFG component presented here as a False Positive sample. This contamination, present in our ML-based method, is also clearly visible in the MIR-based AGN selection. Second, visualization of the MIR-based AGN selection confirms a fundamental discrepancy between X-ray and IR AGN selection. Blue $N2$–$N4$ color which characterize most of the XAGN sample strongly separate False Negative population from the AGN locus. Figure 5.15b shows us results of the extrapolation experiment on the same color-color plot. We see here that classification



(A)

(B)

FIGURE 5.15: Near-IR and mid-IR color-color plot used for the MIR-based AGN selection (Lee et al., 2007). Selection criteria of this method are demarcated by the upper right square, marked by the black lines. Points present on the plots refer to predictions of the ML-based model performed on the labaled and generalizatio data. True Postive (TP, red crosses) refers to properly classified AGNs in the labeled dataset. False Positive (FP, blue dots) refers to galaxies incorrectly classified as AGNs. False Negative (FN, black squares) refers to AGNs incorrectly classified as galaxies. AGN candidates, denoted by yellow (main classification) and purple (extrapolation experiment) rhombs refer to observations from the generalization sample classified as AGNs. *Panel A:* Main classification *Panel B:* Extrapolation experiment

model is able to perform AGN selection outside of the NIR-MIR square used in the traditional color-cut method. Even when reaching outside of the red color region, model is unable to recover XAGN sample. Thus one can conclude that ML-based approach cannot overcome fundamental problems of the IR AGN selection and bring X-ray and IR selection methods together using only optical and IR data.

Table 5.3 shows comparison of evaluation metrics between the MIR color-cut method and ML-based method created during the main classification. Both methods have very similar F1 score and balanced accuracy values. The ML-based model shows higher precision and lower recall compared to color-cut MIR method. This difference may be partially caused by the nature of final model selection were we used higher precision as a more important property of the model. In general metric values show good consistency between the two methods with ML-based model being more conservative approach with output AGN catalog characterized by higher purity and lower completeness. At the same time ML-based method shows to be more universal due to the lack of MIR detection condition: only 24% of AGN candidates present in our catalog were in $S7$ and $S11$ MIR passbands. Let us now analyze of the precision-recall trade-off shown by the final ML-based model in the main classification and in the case where additional MIR detection condition was imposed. Additional MIR detection condition does not have a strong impact on the purity of AGN catalog - precision increase from 0.73 to 0.77. At the same time we see a significant increase in the completeness of the AGN catalog when consider data with MIR detection - recall increase from 0.64 to 0.74. Thus incorporation of the MIR information does not help to significantly reduce the contamination from the SFG sample. It helps however to increase completeness of the AGN catalog. Such behavior can be explained as follows.

Method	F1	Precision	Recall	bACC
$\left\{\begin{array}{l} N2 - N4 > 0 \\ S7 - S11 > 0 \end{array}\right.$	0.76	0.73	0.80	0.84
hard voter	0.75	0.77	0.74	0.86

TABLE 5.3: Performance comparison of final model of the main classification with the MIR-based color-cut method described in Lee et al. (2007). Only objects detected in $S7$ and $S11$ passbands were used.

The AGNs which were detected in optical and NIR passbands but were lacking measurements in MIR passbands may be AGNs with low AGN activity and/or small dust component. In both cases these objects are very hard to recover using only optical and NIR information due to their similarities to galaxies in this spectral range.

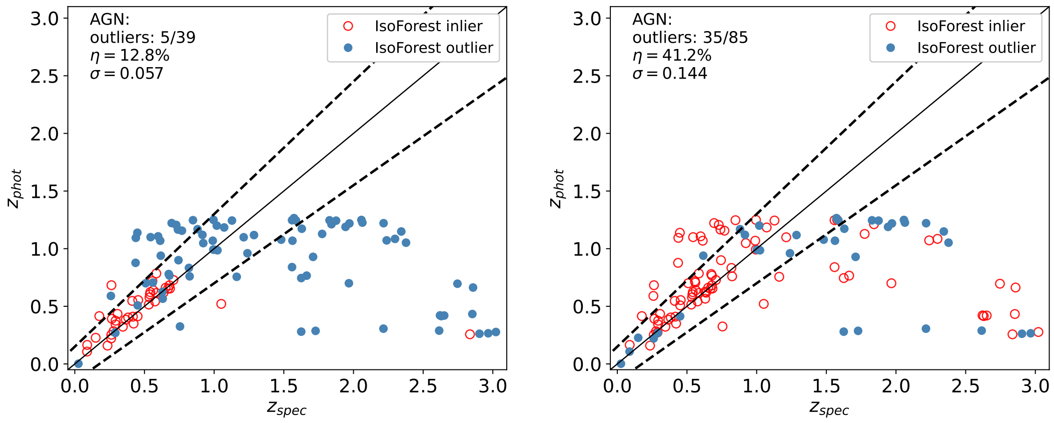
5.5 Outlier detection

The AGN catalog obtained in this work may be used in astrophysics for various purposes, such as the base for creating the target sample for spectroscopic observations or AGN studies in general. However, in many cases, statistical analysis of such a catalog may be very sensitive to biases and contamination present in the data. Reduction of these issues is particularly important in galaxy evolution and observational cosmology studies, such as the ones discussed in Sec. 2.3.

This section addresses some catalog issues that might be effectively removed using outlier detection machine learning techniques. In particular, we study two applications of such methods. One of them is the possibility of removing catastrophic photometric redshift estimation errors from the AGN candidates catalog. Many currently existing photometric catalogs suffer from large photometric redshift errors, which make them not applicable for many cosmological studies, where any type of clustering properties are used. Thus removing objects with non-congruent photometric redshifts is crucial for this catalog to be useful. The second method application of outlier detection focuses on removing contaminants from the AGN candidates catalog. Here, a combination of outlier detection and high-dimensional visualization is used to identify suspicious objects and join them with specific contamination sources such as high-redshift SFG, low-redshift dusty galaxies, or low-activity AGNs. The presence of these objects in the AGN catalog makes it unreliable for any AGN population studies. Thus this type of catalog cleaning is also highly important.

5.5.1 Redshift-based outlier detection

The first type of outlier detection performed in this work was redshift-based outlier detection with Isolation Forest. This method's main purpose was to detect AGN candidates present in our catalog with improperly estimated photometric redshifts. The idea of procedure was created as follows. First, a proper redshift limit was chosen to reduce the probability of a catastrophic redshift estimation error. Next, an Isolation Forest model was fitted to the training data in the feature space created out of features used in the main classification together with spectroscopic redshift. After the model was fitted to the data, it was used to predict both labeled data and the AGN candidates catalog. During prediction feature set was modified: now, it contains photometric redshift from Ho et al. (2021) instead of spectroscopic redshift. During the fitting procedure, a specific value of the spectroscopic redshift value is matched to



(A) (B)

FIGURE 5.16: Comparison between spectroscopic redshift and photometric redshift estimation from Ho et al. (2021), showing results of Isolation Forest outlier detection. Cones determined by dotted lines as well as η and σ parameters were calculated in the same way as described in Fig. 3.2. Red circles and blue dots refer to objects identified by the Isolation Forest model as inlier and outlier, respectively. *Panel A:* Prediction of Isolation Forest model trained on the combined galaxy and AGN data. *Panel B:* Prediction of Isolation Forest model trained on the AGN data only.

the position of the observation in the high-dimensional color space. At the next step, a model fitted to the real spectroscopic redshift values will see a photometric redshift value, which strongly differs from the spectroscopic one as a significant shift of the observation position in the feature space and will classify such entity as an outlier. This way, it is possible to detect doubtful cases of a photometric redshift estimation and obtain a clean AGN catalog for specific applications in observational cosmology and environmental studies.

The first important step in constructing this outlier detection mechanism is the selection of a uniform redshift upper limit for both model fitting and prediction. Since only a photometric redshift is available for both AGN candidates and labeled data, it was used to create a redshift limit. To find the most optimal value, we used two criteria: on the one hand, we want to preserve the largest number of AGN candidates in the limited catalog. On the other hand, we want to avoid redshift ranges with significant bias present in photometric redshift estimations. To find a redshift limit, which does not limit AGN candidates catalog size in a way that makes a limited catalog unusable for a reasonable statistical analysis, quartiles of the photometric redshift distribution were calculated. The first quartile, which contains 25% of the whole AGN candidates catalog, is set at $z_{phot} = 1.033$. The second and third quartiles (50% and 75% of the catalog) are set at $z_{phot} = 1.264$ and $z_{phot} = 1.644$ respectively. Analysis of these values together with redshift properties of the labeled data shown in Fig. 3.2 gives us a clue why we should choose a second quartile as a reasonable upper redshift limit. When looking at residual plot (lower part of Fig. 3.2) we can observe a strong change at $z_{spec} \simeq 1.5$. At $z_{spec} \leq 1.5$, we can observe small residual values with a slight overestimation bias. Once we move towards a higher redshift range $z_{spec} > 1.5$, we see a significant bias toward photometric redshift underestimation.

The main reason for this behavior comes not only from the difficulty of accurate photometric redshift estimation at larger distances and estimation of AGN redshift

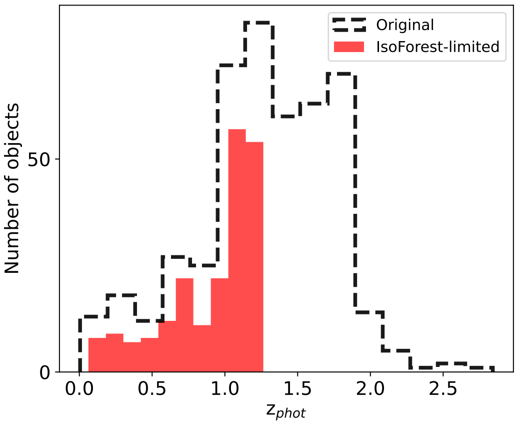


FIGURE 5.17: Histogram comparing photometric redshift distribution of the whole AGN candidates catalog obtained in the main classification and sub-sample of this catalog limited by the redshift-based Isolation Forest model.

in particular but also from the specific purposes of the particular redshift estimation described in Ho et al. (2021). This work aimed to develop an efficient way of the photometric redshift estimation for galaxies present in the AKARI NEP-Wide field. Since the galaxy sample is located at lower redshifts than the AGN sample, the approach was optimized towards a lower redshift range. Moreover, the authors of this work showed that the introduction of AGN templates reduces the effectiveness of galaxy sample redshift estimation. Thus, as a result, AGN templates were not used during the final photometric redshift estimation. From the perspective of this work, we can see that photometric redshifts for high-redshift AGNs may be invalid due to two reasons: the inability of the redshift estimator to work with high-redshift objects and to work with AGNs in particular. Thus we should not reach far beyond the range of the galaxy redshift distribution. AGN redshift estimation seems to be accurate enough within the galaxy redshift range with a small number of catastrophic estimation errors. The additional argument comes from purely statistical properties. Using data from both galaxies and AGNs to fit the Isolation Forest model will be more likely to classify high-redshift objects as outliers due to their small number in the sample. In addition, the large span of the redshift range will make it more difficult to properly localize the positions of outliers in the high-dimensional feature space.

In order to test the properties of the outlier detection, we fit the Isolation Forest model to data sets limited to the first three quartiles and made predictions on these datasets. To analyze the properties of each dataset, we see how training AGNs which were selected by the model as inliers fit into the redshift cones described in Fig. 3.2. We see a strong drop in the redshift estimation accuracy with the increase of the redshift limit. Such drop occurs in both cases: when we use both galaxies and AGNs as training data to fit the model and when we use training AGNs only. In the case of combined training data we obtain $\eta_{Q1} = 13.2\%$, $\eta_{Q2} = 12.8\%$ and $\eta_{Q3} = 21.3\%$ for the first, second and third quartile respectively. In the case of the model fitted to AGN data only, we obtain $\eta_{Q1} = 29.6\%$, $\eta_{Q2} = 41.2\%$ and $\eta_{Q3} = 47.0\%$ for the first three quartiles. Here we can see a significantly better performance of the outlier detection model trained on the combined galaxy and AGN data at all redshift ranges. The model's poor performance trained on the solely AGN data probably comes from the high sparsity of the data located in the high-dimensional feature space. In other words, a small amount of the training data with large distances between

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

TABLE 5.4: Statistical properties of the AGN candidates catalog limited to objects selected as inliers by the redshift-based Isolation Forest model. Median, median absolute deviation (MAD), and minimal and maximal limits are shown.

points in the high-dimensional space makes it difficult for the model to find the positions of real outliers. When analyzing the performance of the model trained on the combined galaxy and AGN data, we see that it can stabilize the contamination level of the photometric redshift catalog to a certain redshift range. The percentage of the catastrophic errors for models operating in the first and second quartile ranges is very similar. Due to these observations, the second quartile ($z_{phot} = 1.264$) was chosen as an upper redshift limit for the final outlier detection model, and the combination of the galaxy and AGN samples was used as a training data.

Figure 5.16 shows results of the outlier detection performed on the second quartile limited data using combined AGN and galaxy data (Fig. 5.16a) and AGN-only data (Fig. 5.16b). We can see here tendencies that were described when analyzing the effect of the redshift limit on the model performance. The model trained on the combined AGN and galaxy data detects two kinds of outliers. The first kind is the objects located at higher redshifts in general. These observations were selected as outliers because most galaxies are located at much lower redshifts. The AGN sample, shifted towards higher redshifts, is too small to change the model's behavior. The second type of outliers is the most problematic high-redshift AGNs, which have wrong photometric redshift estimations. The model trained on the combined AGN and galaxy data is very good at detecting such objects. Almost all of these problematic observations with catastrophic redshift errors were found and removed from the catalog. The model trained on the solely AGN data shows different tendencies. Here we do not see a systematic rejection of high-redshift objects. Instead, the model can operate on the whole redshift range defined by the second quartile. This flexibility, however, comes together with much higher contamination of the final catalog. We see that model is unable to detect most of the outliers showing both over- and under-estimation of the photometric redshift.

Using the final model of the second quartile limited Isolation Forest trained on combined AGN and galaxy data, an outlier detection was performed on the AGN candidates catalog. As a result, a catalog of 210 inlier objects was obtained. A comparison of photometric redshift distribution for the original AGN candidates catalog and selected inliers is shown in Fig. 5.17. One can see a similarity of the two distributions. The Isolation Forest catalog cleaning introduced no major change in the redshift distribution within a second quartile range. Properties of the AGN candidates catalog with removed catastrophic photometric redshift errors are presented in Tab. 5.4.

5.5.2 Class-based outlier detection

The second type of outlier detection tested in this work addresses the problem of AGN candidates catalog contamination and creates a ground for search for unusual objects. This class-based problem is more subtle than previously described redshift-based outlier detection. Multiple types of outliers that we want to detect, together with less straightforward differences between them, make it necessary to modify their detection method. We will refer to this outlier detection problem as *class-based outlier detection*, while the method described in the previous section will be referred to as *redshift-based outlier detection*.

The class-based outlier detection was created as follows. First, two separate Isolation Forest models were created. One was fitted to the training galaxy sample (we will refer to it as *Galaxy Isolation Forest*), and the second one was fitted to the AGN training sample (we will refer to it as *AGN Isolation Forest*). We used the same feature set as in the main classification in both cases. In the case of this outlier detection, we are interested in two types of objects. One of them are outliers detected by the AGN Isolation Forest. These might be possible AGNs that are not typical for the training sample AGN population, e.g., high-redshift AGNs or type-II AGNs. Outliers detected by the AGN Isolation Forest might also be contaminants of the AGN catalog, e.g., SFG, which may not fit the AGN population in the color feature space. The second type of objects are the Galaxy Isolation Forest inliers. Such inliers found in the AGN catalog may be different contaminants such as SFG, low activity AGNs or dusty galaxies.

To analyze the work of the Isolation Forest models, we will use previously mentioned sub-classes present in the training data, i.e., AGNs made of AGN1 (163 objects) and XAGN (34 objects) subsamples and galaxies together with a high-redshift SFG subsample (HzSFG, 17 objects). The HzSFG group was made of galaxies characterized by $N2$–$N4 > 0$ color and $z_{spec} \geq 1$. Here we used only observations where spectroscopic redshift was measured using at least two emission lines.

Now let us analyze Isolation Forest models predictions on the labeled data. Figure 5.18b and 5.18a shows us $N4$ vs $N2$–$N4$ magnitude-color plot for predictions of Isolation Forest models on the AGN and Galaxy labeled data respectively. Let us first analyze how Isolation Forest models interpret the distribution of the AGN labeled data. Observations that the AGN Isolation Forest treats as outliers are mainly objects located in the galaxy locus. The same applies to observations selected as inliers by the Galaxy Isolation Forest model. It is worth noticing that samples of outliers and inliers have only a small intersection of a few objects. Thus these two Isolation Forest models are sensitive to different types of AGN objects. In the case of labeled galaxies samples and HzSFG observations, we can see that the vast majority of HzSFG can be classified as outliers by the AGN Isolation Forest. Figure 5.18c shows us the same color-magnitude plot with the location of AGN candidates selected as outliers by the AGN Isolation Forest or inliers by the Galaxy Isolation Forest. The location of these two subsamples is very different. Outliers detected by the AGN Isolation Forest, which constitute the sample of 21 objects, occupy the same region as HzSFG from the labeled sample. Inliers detected by the Galaxy Isolation Forest are located at the region that contains many training galaxies and is underrepresented by the training AGNs. As discussed in previous sections, these objects may be low activity AGN or dusty galaxies, but they might also be high-redshift AGNs. Thus by removing these inliers from the AGN candidates catalog, we reduce contamination of the final catalog at the expense of possible exclusion of high-redshift AGNs.

Analysis of the color-magnitude plot may not be enough to properly study the nature of outliers and inliers detected by different Isolation Forest models. In order to deepen this study, we add the next step into the outlier detection method in the form of tSNE algorithm visualization. We use the tSNE algorithm to obtain a meaningful two-dimensional representation of the high-dimensional feature space and to more accurately visualize the relative distances and connections of different groups of objects. The main tuning parameter present in the tSNE algorithm is the *perplexity* parameter, which defines the most important scale to be kept in the visualization. In other words, high perplexity value gives us more information about subtle, small-scale connections between observations, while low perplexity emphasizes the general, large-scale properties of the data. In order to find the most appropriate perplexity value, we tested how this parameter affects the distribution of different classes of objects in the training data. We tested a set of perplexity parameter values: 30, 50, 80, and 150. It turned out that, while the general shape of a galaxy and AGN distribution remains unchanged in the two-dimensional tSNE visualization, the HzSFG sample is particularly sensitive to the perplexity parameter value. Since this class of objects is a significant source of AGN catalog contamination, it was important to have a well-localized cluster of HzSFG in the tSNE visualization. Such strong clustering was obtained with a perplexity parameter value equal to 80, and this value was used in further analysis.

Figure 5.19a shows us two-dimensional visualization of the training data distribution. We will refer to artificial tSNE dimensions as *tSNE1* and *tSNE2*. Similar to the color-magnitude plot analyzed previously, here we see a very clear separation between the two classes. One large region is strongly dominated by galaxies with a small contribution from type-I AGNs and X-ray AGNs. AGNs predominantly occupy the second large region. In the case of AGN-dominated regions, we see to main contamination regions. One of them is the region located in tSNE1$\in [0, 20]$ and tSNE2$\in [-10, 0]$ range. These are mainly low-redshift galaxies similar to AGN's appearance. The second contamination region is located in tSNE1$\in [20, 40]$ and tSNE2$\in [0, 10]$ range and is predominantly occupied by HzSFG. Figure 5.19b shows the distribution of the AGN candidates catalog with respect to the training data. Here we can see a very good agreement between AGN candidates catalog location and the location of the AGN training data. Except for two observations, all of the AGN candidates are located precisely in the AGN-dominated region. The placement of AGN candidates selected as outliers or inliers by the AGN and Galaxy Isolation Forest models are shown in Fig. 5.19c. Here we can see that both groups are located in the zones of AGN region contamination. The AGN candidates, identified as inliers by the Galaxy Isolation forest, are located in the first contamination region occupied by the low redshift galaxies, which may also contain low-activity AGNs. On the other hand, the AGN candidates which are identified as outliers by the AGN Isolation Forest are located in the HzSFG region. Thus we have coherent results from two separate methods, Isolation Forest outlier detection, and tSNE visualization. Combining both of them to better understand the nature of AGN candidates catalog outliers, we see that such a combined method allows us to detect and reject two possible sources of contamination. One of them is low-redshift galaxies and possible AGNs with strong host component. The second one is the most problematic group of HzSFG. Properties of the limited AGN candidates catalog cleaned from these contaminants (390 objects) as well as the purest version of this catalog where both galaxy contamination and catastrophic photometric redshift errors were removed (157 objects) are presented in Tab. 5.5.

	median	**MAD**	**min.**	**max.**
Class-based clean catalog				
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
Class- and redshift-based clean catalog				
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

TABLE 5.5: Statistical properties of AGN candidates catalog cleaned of class contamination and class and redshift contamination. Median, median absolute deviation (MAD), minimal and maximal values for redshift and photometry in broad-band filters used in training and generalization.

5.6 Results

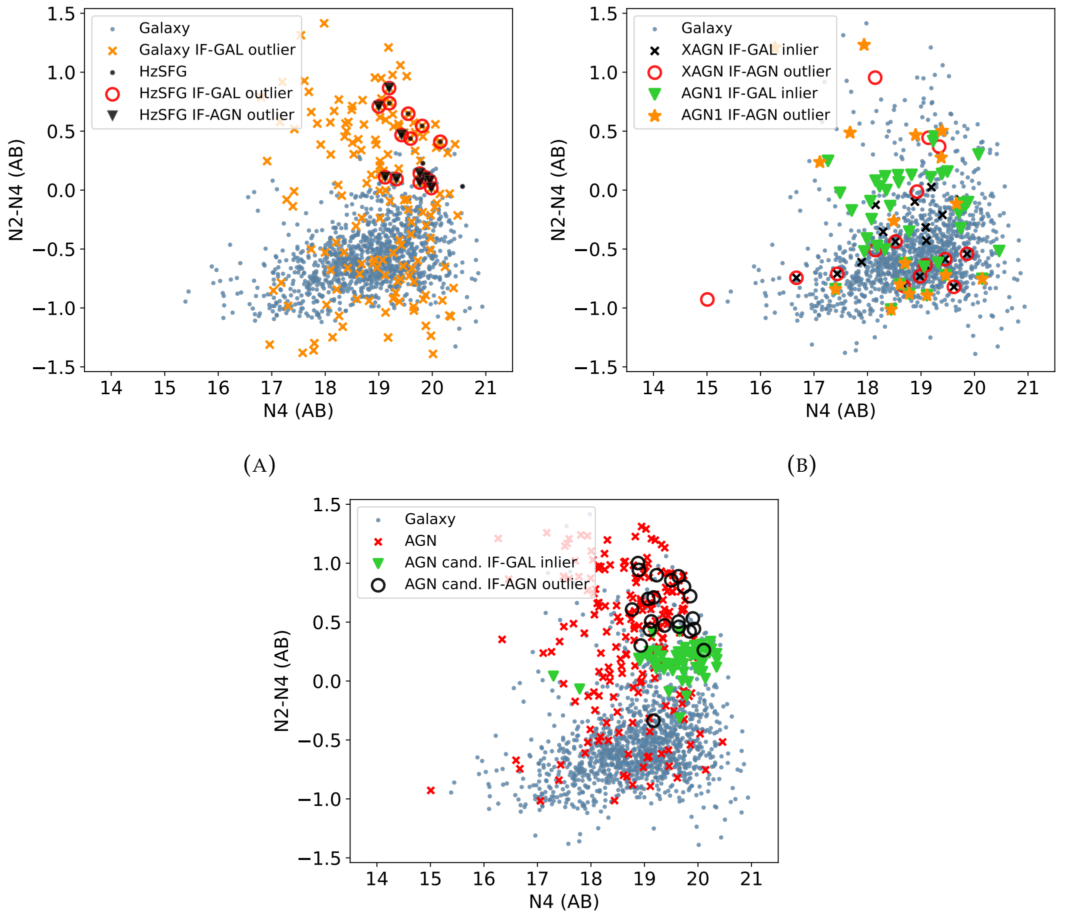
The main results of the work presented in this thesis can be divided into two parts. The methodological result of this research is the construction of the compound machine learning pipeline for the photometric AGN selection from the multiwavelength catalogs. This pipeline has three components, all of which showed to be very effective for a specific task. The first component of the pipeline is related to data preparation. Here, besides traditional machine learning elements such as feature selection, two additional methods crucial for the procedure were applied. One of them was the construction of the AGN training sample, which was based on the mid-IR target preselection and allowed to indirectly provide information about the mid-IR selection into the ML model structure during training. The second method was the MCD limit of the generalization sample. It allowed the selection of AGN candidates exactly in the feature space region defined by the training sample.

These two elements anchored the effectiveness of the second component, i.e., the supervised classification model. By testing various types of classification models with various weighting strategies, including those based on fuzzy logic, it was possible to select a sample of the best classifiers and create a final majority voting classifier. A comparison of the properties of this final best classifier with a simple color-color mid-IR method on the sub-sample of training data with existing mid-IR measurement shows prominent similarities between the two methods. It indicates that the model was able to learn mid-IR selection properties via the training sample and adapt them to the optical and near-IR data. Thus, it makes it possible to recreate the mid-IR-based AGN selection for a sample with no mid-IR data, which was not possible before.

The third component of the ML pipeline was based on the outlier detection methods. These methods were applied to the AGN candidates catalogs created in the previous steps in order to increase their purity. In this case, both redshift-based and class-based methods showed very good results. Class-based methods were able to detect two major sources of contamination present in the AGN catalog. One class of outliers were objects located in-between galaxy and AGN classes, which might be low-activity AGN or low-redshift dusty galaxies. The second type of objects were high-redshift SFG candidates, which are buried inside the AGN locus and comprise the most problematic source of contamination. A combination of Isolation Forest outlier detection with the tSNE visualization allowed us to identify specific properties of high-redshift SFGs, which made it possible to detect a set of objects with similar properties in the AGN candidates catalog. The Isolation Forest outlier detection method is also shown to be very effective in the identification of errors in the input catalogs, in particular incorrect photometric redshift error estimations. The combination of these methods makes the whole pipeline a very effective tool for AGN selection. It can be easily adapted for other types of catalogs and data. The fact that all the elements of this pipeline can be modified or retrained and adapted to suit other astrophysical needs opens a lot of possibilities for its application.

The second, physical, result of this work is the AKARI-NEP AGN catalog itself. The main catalog consists of 465 AGN candidates; its properties are summarized in Tab. 5.2. It is characterized by the 73% purity and 64% completeness. Examination of the MIR properties of the subsample of AGN candidates with the MIR detection shows that objects from this catalog exhibit very typical properties of MIR-bright AGN. The part of the AGN training sample that contributed to the creation of the AGN candidates catalog consisted mostly of type-I AGN. Thus we can make a reasonable assumption that a majority of objects present in the result catalog are type-I AGNs. As shown by Poliszczuk et al. (2021), spectral energy distributions (SEDs) of these AGNs obtained making use of the CIGALE fitter (Boquien et al., 2019) mostly confirm that they have high AGN fraction.

In addition to the main AGN candidates catalog, three clean subsample catalogs were created. One of them is a catalog with removed catastrophic redshift estimation errors. It consists of 210 objects; its properties are summarised in Tab. 5.4. Due to the specific method that was used for photometric redshift estimation in the multiwavelength AKARI NEP-Wide catalog, the high-redshift end of AGNs distribution has significant contamination. This sub-catalog without inaccurate redshift estimations is created for clustering and environmental studies in the low-redshift universe. The second sub-catalog of AGN candidates was cleaned out of class-contaminants. It consists of 390 objects, and its properties are shown in Tab. 5.5. This catalog is well suited for target selection for follow-up spectroscopic observations. Such observations are indeed now ongoing, and further are being planned. Moreover, objects identified as contaminants might also comprise an interesting sample of objects per se. One interesting subclass of contaminants are high-redshift SFG candidates, and the second one is possible low-redshift galaxies hosting low-activity AGNs. The final sub-catalog of AGNs was obtained by a combination of catalog cleaning procedures presented above. It consists of 157 objects. Its properties are shown in Tab. 5.5. This catalog is the purest obtained sample with reduced class- and redshift-contamination. As was mentioned in Sec. 2, IR AGN selection probes a high end of the Eddington ratio distribution. This cleanest sample can be used for further studies of this AGN population with multiwavelength data available for the North Ecliptic Pole field.

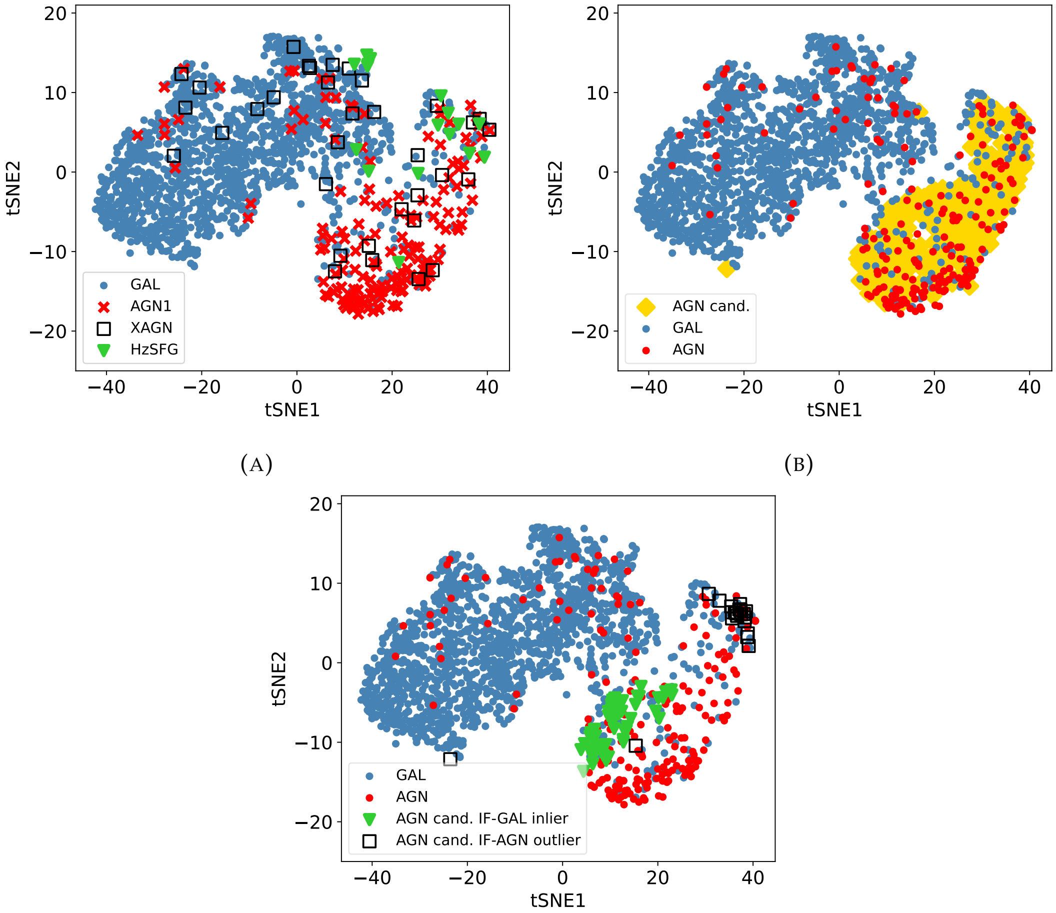


(A)

(B)

(C)

FIGURE 5.18: Magnitude-color plot of $N2$–$N4$ vs $N4$ showing results of class-based Isolation Forest outlier detection. *Panel A:* Results of outlier detection performed by Isolation Forest models on the labeled galaxy observations. Grey dots show the general distribution of labeled galaxies data, and black dots show positions of high-redshift SFG (HzSFG). Orange crosses refer to objects identified as outliers by the Galaxy Isolation Forest. Red circles show HzSFG identified as outliers by the Galaxy Isolation Forest Black triangles show HzSFG identified as outliers by the AGN Isolation Forest. *Panel B:* Results of outlier detection performed by Isolation Forest models on the labeled galaxy data. Grey dots show the general distribution of labeled galaxy observations. Black crosses, and red circles show XAGNs identified as inliers by Galaxy Isolation Forest and outliers by AGN Isolation Forest, respectively. Green triangles and orange stars show type-I AGNs identified as inliers by Galaxy Isolation Forest and outliers by AGN Isolation Forest, respectively. *Panel C:* Results of outlier detection performed by the Isolation Forest models on the AGN candidates catalog. Grey dots and red crosses show positions of galaxy and AGN training data. Green triangles refer to objects identified as inliers by Galaxy Isolation Forest. Black circles refer to objects identified as outliers by AGN Isolation Forest.



(A)

(B)

(C)

FIGURE 5.19: Two-dimensional tSNE visualization of the training data, AGN candidates catalog, and results of Isolation Forest outlier detection. *Panel A:* The tSNE visualization of the training data consisting of galaxies (blue dots), high-redshift SFG (green triangles), type-I AGNs (red crosses) and XAGNs (black squares). *Panel B:* The tSNE visualization of training galaxy (blue dots) and AGN (red dots) data compared to the placement of AGN candidates (yellow rhombs). *Panel C:* The tSNE visualization of training galaxy (blue dots) and AGN (red dots) data compared to the placement of AGN candidates selected as inliers (green triangles) and outliers (black squares) by Galaxy and AGN Isolation Forest models respectively.

6

Summary

In the present work, several significant results were achieved. Let us summarize obtained conclusions and mark out possible topics for future studies.

First, the MCD limit imposed on the generalization sample showed to be a very effective technique to avoid extrapolation. Analysis of prediction results on the training and generalization samples based on the placement of AGN candidates and AGN training data in the NIR and MIR color distributions gives no indicators of significant extrapolation during generalization. The MCD limit imposed on the generalization set seems to efficiently bring generalization close to the model performance on the training set. Analysis of the impact of the MCD limit on the properties of the unlabeled data sample shows it affects strongest the optical passbands, cutting out the faint end of the distribution. This effect is caused by the requirements of target selection for spectroscopic measurements, which were imposed on the training sample. However, we also see a significant reduction of objects characterized by the $N2$–$N4 = 0$ color. Removal of these objects by the MCD algorithm suggests they were not well represented in the training set and were able to cause difficulties during generalization. It is especially important because the $N2$–$N4 = 0$ region was recognized during the analysis of outlier detection methods as a problematic area with possible significant contamination from low-redshift galaxies or low-activity AGN. The model's good performance presented in this work shows the usefulness of the MCD-based generalization sample limit technique in astronomy. It can be effectively applied to various cases where the training sample might not be representative for the unlabeled data. These are in particular situations when a supervised model is trained on data with spectroscopic classes or redshifts, as well as situations where a model is trained on simulated data. Moreover, due to the relative simplicity of the model, it should be possible to recover the selection function created by the MCD limit. This problem should be studied to increase this method's usability further.

Next, studies of class weights and fuzzy logic on different types of machine learning algorithms showed several important results. First, one can observe that the impact a specific weight has on a model is model-dependent, with some models being susceptible to weight application and others showing no significant change. Second, class weight showed a more significant change in model performance than fuzzy weights. In general application of class-based weights shifts the separation boundary outwards from the smaller class (in this case, AGN class), increasing recall of the model and decreasing its precision. This tendency was shown by most models except tree ensemble algorithms, which seem to be a case-specific behavior. Fuzzy logic (or instance weights) had a less direct impact on the classification caused by the superposition of different effects. In general, the biggest change in the model performance comes from the distance-based fuzzy logic. This type of fuzzy logic works the most effectively in models without class weights. In such a case, the decision boundary lies close to a smaller class, and distance-based fuzzy weight can produce a cumulative effect causing an increase in AGN catalog completeness. Class-weighted models have decision boundaries pushed away from the AGN class, and distance-based weights act predominantly to reduce the impact of outliers lying in the galaxy locus. Interestingly, this increase in AGN catalog completeness for models without class balance does not decrease catalog purity, as observed in the application of class-based weights. This result occurs due to the opposite tendencies in distance-based weighting application, causing the increase of purity in one region of feature space and its decrease in another. The error-based fuzzy logic, despite its physical motivation, does not have a significant impact on the classification performance in both class-balanced and non-balanced models. Thus, fuzzy logic should be applied as a fine-tuning of the model after the result of class weights is established.

A comparison between the MIR-based color AGN selection and ML-based selection confirms the genuineness of the main aim of this work. It is possible to create an ML-based AGN selection that mimics the properties of the MIR-based method using only optical and NIR broadband photometry. Its effectiveness is caused by two crucial tools applied at the beginning of the model training. First, the AGN training data set was based on the spectroscopic sample, which used MIR-selected AGN as targets. Thus information about MIR-based AGN selection was imprinted in the structure of the training sample and passed into the model construction. The model effectively used this information and recovered the AGN sample from the data due to the specific power-law shape of the AGN SED present not only in the MIR part of the spectrum but also at longer NIR wavelengths. This information was translated into the AGN properties in AKARI NIR color space. It also shows how effective is the combination of data coming from the combined deep optical ground observations and AKARI NIR measurements. The second tool, crucial for this aim, was the use of the MCD algorithm to limit the distribution of unlabeled data to the training sample shape. This property gave an opportunity to effectively train the model and apply it safely to the generalization data.

The extrapolation experiment, which was trying to push classifier effectiveness even further and overcome the limitations of the MIR selection, showed unsatisfying results. It could not recover objects located within the galaxy class in the NIR and color space. These were predominantly X-ray-selected AGNs. The inability to recover X-ray AGN sample with both ML-based and MIR-based techniques confirms fundamental differences between X-ray-based and IR-based selection methods. However, the lack of success of the ML-based method may be partially caused by the small size of the AGN training data used in the extrapolation experiment. This problem may be solved in the future by training the model on the additional template-based AGN sample. Combining information from such data with an additional MCD-limit, which takes into account the properties of the artificial data, may work similarly to the MIR case and translate some of the X-ray information into the model.

Finally, studies on outlier detection methods showed very promising results. It turned out that it is possible to detect most of the catastrophic photometric redshift errors with the Isolation Forest algorithm as long as they are located near the center of the spectroscopic redshift distribution of the training data. Such an approach allows one to create a catalog suitable for further clustering studies and its application to observational cosmology tasks. The class-based outlier detection was performed by combining the Isolation Forest method with tSNE algorithm visualization. Such a combination showed that the Isolation Forest algorithm trained on AGN data detects outliers with properties similar to high-redshift star-forming galaxies (SFG). This result is particularly important since high-redshift SFGs are the primary contaminant of the MIR-selected AGN catalogs. On the other hand, the Isolation Forest trained on the galaxy data tends to find inliers within AGN observations that are located in a problematic $N2$–$N4 \simeq 0$ region. These might be low-redshift dusty galaxies as well as low-activity AGNs. On the other hand, the sample place in the color distribution might be occupied by a specific class of high-redshift AGNs which are poorly represented in the training data. Removal of these outliers allows one to obtain an AGN candidates catalog with high purity suitable for various evolutionary and environmental studies.

To summarize, in this work, an effective complex machine learning procedure for AGN selection was developed. It allows one to overcome instrumental limitations of MIR telescopes and significantly increase the size of the AGN catalog compared to the obtained with traditional MIR-based AGN selection techniques. In addition, once the AGN catalog is obtained, one can use developed outlier detection methods to find an adequate trade-off between catalog purity and completeness. It can be done by controlling contamination of various catalog properties, such as the accuracy of redshift estimation or the presence of wrongly classified objects. As with most machine learning methods, this compound technique can be modified and applied to various tasks. Some of these modifications might be retraining the classifier to operate on different classes of objects or mimicking different properties of objects (than MIR-based AGN selection) from outside of the available spectral range. Another type of modification might be the creation of fuzzy logic weights suitable for a specific task. Finally, the flexibility of the created outlier detection method allows one to construct more complex and subtle ways to control the properties of the obtained catalog. Techniques presented in this work will be very useful for the application in modern panchromatic big data astronomy as well as data-simulation combined observational cosmology.

Bibliography

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

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

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

Alonso-Herrero, Almudena et al. (Mar. 2001). The Nonstellar Infrared Continuum of Seyfert Galaxies. 121.3, pp. 1369–1384. DOI: 10.1086/319410. arXiv: astro-ph/0012096 [astro-ph].

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

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

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

Ashby, Matthew, J. R. Houck, and Perry B. Hacking (Sept. 1992). Deep Infrared Galaxies. 104, p. 980. DOI: 10.1086/116291.

Assef, R. J. et al. (July 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, p. 26. DOI: 10.1088/0004-637X/772/1/26. arXiv: 1209.6055 [astro-ph.CO].

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

Barden, S. C. et al. (1993). Hydra – KittPeak multi-object spectroscopic system. *ASPCS* 37, pp. 185–202.

Barrufet, L. et al. (Sept. 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 et al. (Mar. 2017). The AGN Population in the Akari NEP Deep Field. *Publication of Korean Astronomical Society* 32.1, pp. 271–273. DOI: 10.5303/PKAS.2017.32.1.271.

Beckert, T. et al. (Aug. 2008). Probing the dusty environment of the Seyfert 1 nucleus in NGC 3783 with MIDI/VLTI interferometry. 486.3, pp. 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. et al. (Sept. 2003). The Broadband Optical Properties of Galaxies with Redshifts 0.02<z<0.22. 594.1, pp. 186–207. DOI: 10.1086/375528. arXiv: astro-ph/0209479 [astro-ph].

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

Boquien, M. et al. (Feb. 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 et al. (Jan. 2018). The Hyper Suprime-Cam software pipeline. 70, S5, S5. DOI: 10.1093/pasj/psx080. arXiv: 1705.06766 [astro-ph.IM].

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

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

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

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

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

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

Bulbul, Esra et al. (Oct. 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 et al. (Jan. 2019). AKARI NEP field: Point source catalogs from GALEX and Herschel observations and selection of candidate lensed sub-millimeter galaxies. 71.1, 12, p. 12. DOI: 10.1093/pasj/psy134.

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

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

Cepa, Jordi et al. (Aug. 2000). OSIRIS tunable imager and spectrograph. *Optical and IR Telescope Instrumentation and Detectors*. Ed. by Masanori Iye and Alan F. Moorwood. Vol. 4008. SPIE Conference Series, pp. 623–631. DOI: 10.1117/12.395520.

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

Chen, Bo Han et al. (Mar. 2021). An active galactic nucleus recognition model based on deep neural network. 501.3, pp. 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 and 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 et al. (Apr. 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, p. 31. DOI: 10.1093/pasj/psz012. arXiv: 1902.02800 [astro-ph.GA].

Clarke, A. O. et al. (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. et al. (May 2000). 2.5-11 micron spectroscopy and imaging of AGNs. Implication for unification schemes. 357, pp. 839–849. arXiv: astro-ph/0003298 [astro-ph].

Coil, Alison L. et al. (Jan. 2008). The DEEP2 Galaxy Redshift Survey: Color and Luminosity Dependence of Galaxy Clustering at z ~1. 672.1, pp. 153–176. DOI: 10.1086/523639. arXiv: 0708.0004 [astro-ph].

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

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

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

Croton, Darren J. et al. (Apr. 2006a). Erratum: The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. 367.2, pp. 864–864. DOI: 10.1111/j.1365-2966.2006.09994.x. arXiv: astro-ph/0602065 [astro-ph].

— (Jan. 2006b). The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. 365.1, pp. 11–28. DOI: 10.1111/j.1365-2966.2005.09675.x. arXiv: astro-ph/0508046 [astro-ph].

D'Isanto, A. and K. L. Polsterer (Jan. 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. et al. (Jan. 2021). The Landscape of Galaxies Harboring Changing-look Active Galactic Nuclei in the Local Universe. 907.1, L21, p. L21. DOI: 10.3847/2041-8213/abd852. arXiv: 2010.10527 [astro-ph.GA].

Doi, Yasuo et al. (June 2015). The AKARI far-infrared all-sky survey maps. 67.3, 50, p. 50. DOI: 10.1093/pasj/psv022. arXiv: 1503.06421 [astro-ph.GA].

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

Elvis, Martin et al. (Nov. 1994). Atlas of Quasar Energy Distributions. 95, p. 1. DOI: 10.1086/192093.

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

Faber, S. M. et al. (Aug. 2007). Galaxy Luminosity Functions to z~1 from DEEP2 and COMBO-17: Implications for Red Galaxy Formation. 665.1, pp. 265–294. DOI: 10.1086/519294. arXiv: astro-ph/0506044 [astro-ph].

Faber, Sandra M. et al. (Mar. 2003). The DEIMOS spectrograph for the Keck II Telescope: integration and testing. *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*. Ed. by Masanori Iye and Alan F. M. Moorwood. Vol. 4841. SPIE Conference Series, pp. 1657–1669. DOI: 10.1117/12.460346.

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

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

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

Fernández, A. et al. (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, and E. Hatziminaoglou (Mar. 2006). Revisiting the infrared spectra of active galactic nuclei with a new torus emission model. 366.3, pp. 767–786. DOI: 10.1111/j.1365-2966.2006.09866.x. arXiv: astro-ph/0511428 [astro-ph].

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

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

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

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

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

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

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

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

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

Gorjian, V. et al. (June 2008). The Mid-Infrared Properties of X-Ray Sources. 679.2, pp. 1040–1046. DOI: 10.1086/587431. arXiv: 0803.0357 [astro-ph].

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

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

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

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

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

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

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

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

Hacking, Perry et al. (Apr. 1989). A Very Deep IRAS Survey. III. VLA Observations. 339, p. 12. DOI: 10.1086/167272.

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

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

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

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

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

Hasinger, G. et al. (Jan. 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, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

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

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

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

Henry, J. P. et al. (June 2001). Overview of the ROSAT North Ecliptic Pole Survey. 553.2, pp. L109–L113. DOI: 10.1086/320672.

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

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

Hinton, Geoffrey and 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. (Sept. 2008). Nuclear activity in nearby galaxies. 46, pp. 475–539. DOI: 10.1146/annurev.astro.45.051806.110546. arXiv: 0803.2268 [astro-ph].

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

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

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

Huang, Song et al. (Jan. 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 et al. (Oct. 2020). CFHT MegaPrime/MegaCam u-band source catalogue of the AKARI North Ecliptic Pole Wide field. 498.1, pp. 609–620. DOI: 10.1093/mnras/staa2459. arXiv: 2008.05224 [astro-ph.GA].

Huang, Ting-Chi et al. (Oct. 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, pp. 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, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

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

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

Ishihara, D. et al. (May 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. et al. (May 2004). The central dusty torus in the active nucleus of NGC 1068. 429.6987, pp. 47–49. DOI: 10.1038/nature02531.

Jarrett, T. H. et al. (July 2011). The Spitzer-WISE Survey of the Ecliptic Poles. 735.2, 112, p. 112. DOI: 10.1088/0004-637X/735/2/112.

Jeon, Yiseul et al. (Sept. 2010). Optical Images and Source Catalog of AKARI North Ecliptic Pole Wide Survey Field. 190.1, pp. 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, pp. 189–192.

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

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

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

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

Kawada, M. et al. (Oct. 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. et al. (Nov. 1996). The Infrared Space Observatory (ISO) mission. 500, pp. 493–497.

Kim, Eunbin et al. (Nov. 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, pp. 3113–3124. DOI: 10.1093/mnras/stab2090. arXiv: 2108.07125 [astro-ph.GA].

Kim, S. J. et al. (Dec. 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 et al. (Dec. 2015). Mid-infrared luminosity function of local star-forming galaxies in the North Ecliptic Pole-Wide survey field of AKARI. 454.2, pp. 1573–1584. DOI: 10.1093/mnras/stv2006. arXiv: 1509.04384 [astro-ph.GA].

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

Kim, Seong Jin et al. (Jan. 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, pp. 4078–4094. DOI: 10.1093/mnras/staa3359. arXiv: 2012.00750 [astro-ph.GA].

Kimura, Masahiko et al. (Oct. 2010). Fibre Multi-Object Spectrograph (FMOS) for the Subaru Telescope. 62, pp. 1135–1147. DOI: 10.1093/pasj/62.5.1135.

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

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

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

Krumpe, M. et al. (Jan. 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, pp. 911–931. DOI: 10.1093/mnras/stu2010. arXiv: 1409.7697 [astro-ph.HE].

LaMassa, Stephanie M. et al. (Jan. 2015a). Discovery of the First Changing-Look Quasar. *American Astronomical Society Meeting Abstracts #225*. Vol. 225. American Astronomical Society Meeting Abstracts, p. 204.01.

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

Lee, H. M. et al. (Oct. 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 et al. (Feb. 2009). North Ecliptic Pole Wide Field Survey of AKARI: Survey Strategy and Data Characteristics. 61, p. 375. DOI: 10.1093/pasj/61.2.375. arXiv: 0901.3256 [astro-ph.GA].

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

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

Lilly, Simon J. et al. (Aug. 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, p. 119. DOI: 10.1088/0004-637X/772/2/119. arXiv: 1303.5059 [astro-ph.CO].

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

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

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

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

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

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

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

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

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

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

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

Marin, F. et al. (May 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 et al. (Jan. 2005). The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. 619.1, pp. L1–L6. DOI: 10.1086/426387. arXiv: astro-ph/0411302 [astro-ph].

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

Matsuhara, Hideo et al. (Aug. 2006). Deep Extragalactic Surveys around the Ecliptic Poles with AKARI (ASTRO-F). 58, pp. 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*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

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

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

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

Meusinger, H. and N. Balafkan (Aug. 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 et al. (Sept. 2007). The XMM-Newton Wide-Field Survey in the COSMOS Field. V. Angular Clustering of the X-Ray Point Sources. 172.1, pp. 396–405. DOI: 10.1086/516579. arXiv: astro-ph/0612369 [astro-ph].

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

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

Mullis, C. R. et al. (June 2001). The North Ecliptic Pole Supercluster. 553.2, pp. L115–L118. DOI: 10.1086/320670. arXiv: astro-ph/0103202 [astro-ph].

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

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

Nayyeri, H. et al. (Feb. 2018). Spitzer Observations of the North Ecliptic Pole. 234.2, 38, p. 38. DOI: 10.3847/1538-4365/aaa07e. arXiv: 1712.01290 [astro-ph.GA].

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

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

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

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

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

Neugebauer, G. et al. (Mar. 1984). The Infrared Astronomical Satellite (IRAS) mission. 278, pp. L1–L6. DOI: 10.1086/184209.

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

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

Onaka, Takashi et al. (Oct. 2004). The infrared camera (IRC) on board the ASTRO-F: laboratory tests and expected performance. *Optical, Infrared, and Millimeter Space Telescopes*. Ed. by John C. Mather. Vol. 5487. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, pp. 338–349. DOI: 10.1117/12.550857.

Oyabu, S. et al. (May 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. et al. (Aug. 2017). Active galactic nuclei: what's in a name? 25.1, 2, p. 2. DOI: 10.1007/s00159-017-0102-9. arXiv: 1707.07134 [astro-ph.GA].

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

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

Pearson, Chris et al. (Jan. 2019). The Herschel-PACS North Ecliptic Pole Survey. 71.1, 13, p. 13. DOI: 10.1093/pasj/psy107. arXiv: 1809.03990 [astro-ph.GA].

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

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

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

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

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

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

Pilbratt, G. L. et al. (July 2010). Herschel Space Observatory. An ESA facility for far-infrared and submillimetre astronomy. 518, L1, p. 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, pp. 61–74.

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

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

Poliszczuk, Artem et al. (July 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. and Joel N. Bregman (Feb. 2020). SZ Scaling Relations of Galaxy Groups and Clusters Near the North Ecliptic Pole. 890.2, 156, p. 156. DOI: 10.3847/1538-4357/ab6e6c. arXiv: 2001.07802 [astro-ph.CO].

Predehl, P. et al. (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, and 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. et al. (June 2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. 123.6, pp. 2945–2975. DOI: 10.1086/340187. arXiv: astro-ph/0202251 [astro-ph].

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

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

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

Rousseeuw, Peter J. and Katrien Van Driessen (Aug. 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. and I. F. Mirabel (Jan. 1996). Luminous Infrared Galaxies. 34, p. 749. DOI: 10.1146/annurev.astro.34.1.749.

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

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

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

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

Seo, Hyunjong et al. (Oct. 2019). Clustering of extremely red objects in the AKARI NEP-deep field. 71.5, 96, p. 96. DOI: 10.1093/pasj/psz079.

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

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

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

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

Shim, Hyunjin et al. (Nov. 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, pp. 5065–5079. DOI: 10.1093/mnras/staa2621.

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

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

Smith, Paul S. et al. (Apr. 2002). The Optical Polarization of Near-Infrared-selected Quasi-Stellar Objects. 569.1, pp. 23–35. DOI: 10.1086/339208. arXiv: astro-ph/0112334 [astro-ph].

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

Solarz, A. et al. (May 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. et al. (Oct. 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. et al. (Jan. 2007). Mid-Infrared Galaxy Classification Based on Silicate Obscuration and PAH Equivalent Width. 654.1, pp. L49–L52. DOI: 10.1086/511268. arXiv: astro-ph/0611918 [astro-ph].

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

Stern, Daniel et al. (Sept. 2005). Mid-Infrared Selection of Active Galaxies. 631.1, pp. 163–168. DOI: 10.1086/432523. arXiv: astro-ph/0410523 [astro-ph].

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

Stern, Daniel et al. (July 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, p. 30. DOI: 10.1088/0004-637X/753/1/30. arXiv: 1205.0811 [astro-ph.CO].

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

Takagi, T. et al. (Jan. 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 (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

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

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

Tran, Hien D., Joseph S. Miller, and Laura E. Kay (Oct. 1992). Detection of Obscured Broad-Line Regions in Four Seyfert 2 Galaxies. 397, p. 452. DOI: 10.1086/171801.

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

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

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

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

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

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

Victoria-Ceballos, César Ivan et al. (Feb. 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, p. 192. DOI: 10.3847/1538-4357/ac441a. arXiv: 2201.11869 [astro-ph.GA].

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

Voges, W. et al. (June 2001). The ROSAT North Ecliptic Pole Survey X-Ray Data. 553.2, pp. L119–L123. DOI: 10.1086/320673.

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

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

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

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

Weisskopf, Martin C. et al. (July 2000). Chandra X-ray Observatory (CXO): overview. *X-Ray Optics, Instruments, and Missions III*. Ed. by Joachim E. Truemper and Bernd Aschenbach. Vol. 4012. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, pp. 2–16. DOI: 10.1117/12.391545. arXiv: astro-ph/0004127 [astro-ph].

Werner, M. W. et al. (Sept. 2004). The Spitzer Space Telescope Mission. 154.1, pp. 1–9. DOI: 10.1086/422992. arXiv: astro-ph/0406223 [astro-ph].

White, G. J. et al. (July 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, and R. McCray (Oct. 2000). On the Absorption of X-Rays in the Interstellar Medium. 542.2, pp. 914–924. DOI: 10.1086/317016. arXiv: astro-ph/0008425 [astro-ph].

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

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

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

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

Zakamska, Nadia L. et al. (Nov. 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, pp. 2125–2144. DOI: 10.1086/378610. arXiv: astro-ph/0309551 [astro-ph].

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

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

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

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

A

Appendix: Software

This work was using several Python packages. The creation of machine learning pipeline as well as analysis of obtained results was done wtih SciPy Virtanen et al., 2020, NumPy Harris et al., 2020, Pandas McKinney, 2010; team, 2020, Scikit-learn Pedregosa et al., 2011, and XGBoost Chen and Guestrin, 2016 packages. Visualization of the results was done with Matplotlib Hunter, 2007 and Seaborn Waskom and team, 2020 packages. Most codes, training data, AGN catalog as well as additional SED fitting performed by the members of the NEP team can be found on the GitHub page: https://github.com/ArtemPoliszczuk/NEPWide_AGN

B

Appendix: Metric values

This appendix presents detailed metric results obtained during training. Tab. B.1 and B.2 contain results of the main training process. Tab. B.3 and B.4 show results from the second iteration experiment.

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

TABLE B.1: Metrics for the main classificatoin. Part 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

TABLE B.2: Metrics for the main classification. Part 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

TABLE B.3: Metrics for the extrapolation experiment. Part1/2.

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

TABLE B.4: Metrics for the extrapolation experiment. Part 2/2.