NATIONAL CENTRE FOR NUCLEAR RESEARCH

DOCTORAL THESIS

Machine learning based catalogs of quasars and galaxies for cosmological studies

Author:
Szymon NAKONECZNY

Supervisor:
Agnieszka POLLO
Auxiliary supervisor:
Maciej BILICKI

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

in the

Department of Astrophysics

NATIONAL
CENTRE
FOR NUCLEAR
RESEARCH
ŚWIERK

June 24, 2022

Declaration of Authorship

I, Szymon NAKONECZNY, declare that this thesis titled, “Machine learning based catalogs of quasars and galaxies for cosmological studies” 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

Machine learning based catalogs of quasars and galaxies for cosmological studies

Szymon Nakoneczny

We present two catalogs of quasars derived from the photometric Kilo-Degree Survey (KiDS) Data Release 3 and 4 (DR3, DR4). We present an approach to classification in KiDS DR3, and a more complex methodology to KiDS DR4, in which we add near-infrared imaging, estimate photometric redshifts, and extrapolate to magnitudes fainter than available in the spectroscopy. Our approach to KiDS DR4 produces the final quasar catalog, which we describe below. We build the catalog by training machine learning (ML) models, using optical *ugri* and near-infrared $ZYJHK_s$ bands, on objects known from Sloan Digital Sky Survey (SDSS) spectroscopy. We define inference subsets from the 45 million objects of the KiDS photometric data limited to 9-band detections, based on a feature space built from magnitudes and their combinations. We show that projections of the high-dimensional feature space on two dimensions can be successfully used, instead of the standard color-color plots, to investigate the photometric estimations, compare them with spectroscopic data, and efficiently support the process of building a catalog. The model selection and fine-tuning employs two subsets of objects: those randomly selected and the faintest ones, which allowed us to properly fit the bias versus variance trade-off. We tested three ML models: random forest (RF), XGBoost (XGB), and artificial neural network (ANN). We find that XGB is the most robust and straightforward model for classification, while ANN performs the best for combined classification and redshift. The ANN inference results are tested using number counts, Gaia parallaxes, and other quasar catalogs that are external to the training set. Based on these tests, we derived the minimum classification probability for quasar candidates which provides the best purity versus completeness trade-off: $p(\mathrm{QSO_{cand}}) > 0.9$ for $r < 22$ and $p(\mathrm{QSO_{cand}}) > 0.98$ for $22 < r < 23.5$. We find 158,000 quasar candidates in the safe inference subset ($r < 22$) and an additional 185,000 candidates in the reliable extrapolation regime ($22 < r < 23.5$). Test-data purity equals 97% and completeness is 94%; the latter drops by 3% in the extrapolation to data fainter by one magnitude than the training set. The photometric redshifts were derived with ANN and modeled with Gaussian uncertainties. The test-data redshift error (mean and scatter) equals 0.009 ± 0.12 in the safe subset and -0.0004 ± 0.19 in the extrapolation, averaged over a redshift range of $0.14 < z < 3.63$ (first and 99th percentiles). Our success of the extrapolation challenges the way that models are optimized and applied at the faint data end. The resulting catalog is ready for cosmology and active galactic nucleus (AGN) studies, and we perform an early study on constraining the quasar bias function, using its cross-correlation with the CMB lensing. We obtain a bias function $b_q(z) = 0.57^{+0.03}_{-0.03}(1+z)^2 + 0.07^{+0.06}_{-0.13}$, which at redshift $z = 1.5$ gives a bias value $3.63^{+0.25}_{-0.85}$. Finally, we report 15σ significance of the cross-correlation with the CMB lensing, which for quasars is one of the highest detections of this signal. We publicly release the catalogs at
http://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php (KiDS DR4),
http://kids.strw.leidenuniv.nl/DR3/quasarcatalog.php (KiDS DR3).

NATIONAL CENTRE FOR NUCLEAR RESEARCH

Streszczenie

Wspomagane metodami uczenia maszynowego budowanie katalogów kwazarów oraz galaktyk na potrzeby badań kosmologicznych

Szymon Nakoneczny

Prezentujemy dwa katalogi kwazarów zbudowane na podstawie danych z trzeciego oraz czwartego wydania (DR3, DR4) przeglądu Kilo-Degree Survey (KiDS). Prezentujemy dwie metodologie, pierwszą do klasyfikacji źródeł w KiDS DR3, oraz drugą, bardziej zaawansowaną, w której dodajemy dane z bliskiej podczerwieni, estymujemy fotometryczne przesunięcia ku czerwieni, oraz ekstrapolujemy modele uczenia maszynowego (ML) na magnitudy ciemniejsze niż te znane z przeglądów spektroskopowych. Analiza danych z KiDS DR4 produkuje ostatecznych katalog kwazarów, którego wyniki opisuję na tej stronie. Katalog oparty jest o optyczne filtry *ugri*, oraz bliską podczerwień $ZYJHK_s$, a modele ML trenowane są na danych znanych ze spektroskopowego przeglądu Sloan Digital Sky Survey (SDSS). W 45 milionach obiektów ograniczonych do detekcji we wszystkich 9 pasmach definiujemy podzbiory inferencyjne na podstawie przestrzeni cech zbudowanej z magnitud oraz ich kombinacji. Pokazujemy, że projekcje wysoko wymiarowej przestrzeni cech na dwa wymiary mogą być użyte, zamiast standardowych wykresów kolor-kolor, w celu wizualizacji wyników. Kalibracji modeli uczenia maszynowego dokonujemy za pomocą dwóch podzbiorów walidacyjnych: losowo wybranego z całego zakresu magnitud, oraz ekstrapolacyjnego, do którego wybieramy najciemniejsze magnitudy, które celowo wyłączamy z treningu. Takie podejście pozwala nam na prawidłowe dopasowanie kompromisu między obciążeniem a wariancją. Testujemy trzy modele ML: las losowy (RF), XGBoost (XGB), oraz sztuczna sieć neuronowa (ANN). Pokazujemy, że XGB dostarcza najlepszych wyników klasyfikacji, zaś ANN pozwala osiągnąć najlepsze wyniki dla klasyfikacji i przesunięcia ku czerwieni. Ostateczny katalog, zbudowany za pomocą ANN, testujemy metodą zliczeń, paralaksami z przeglądu Gaia, oraz zewnętrznymi katalogami kwazarów. Na podstawie tych testów, znajdujemy minimalne prawdopodobieństwo klasyfikacji, które kalibruje kompromis między czystością oraz kompletnością, i wynosi: $p(\mathrm{QSO_{cand}}) > 0.9$ dla $r < 22$ oraz $p(\mathrm{QSO_{cand}}) > 0.98$ dla $22 < r < 23.5$. W katalogu znajduje się 158,000 kandydatów na kwazary w bezpiecznym zakresie $r < 22$, oraz kolejne 185,000 kandydatów w zasięgu ekstrapolacji $22 < r < 23.5$. Na podstawie danych testowych szacujemy czystość oraz kompletność katalogu na odpowiednio 97% oraz 94%, a w wyniku ekstrapolacji o jedną magnitudę, kompletność maleje o 3%. Przesunięcia ku czerwieni modelujemy za pomocą rozkładu gaussowskiego, i błąd tej estymacji (średnia oraz rozrzut) wynosi 0.009 ± 0.12 w bezpiecznym zakresie, oraz -0.0004 ± 0.19 w ekstrapolacji, policzone jako średnia w dla $0.14 < z < 3.63$. Sukces ekstrapolacji stawia wyzwanie temu jak modele uczenia maszynowego są kalibrowane oraz stosowane do ciemnych obiektów. Katalog wynikowy jest gotowy do badań kosmologicznych oraz nad aktywnymi jądrami galaktyk (AGN), i w niniejszej pracy przedstawiamy pierwsze wyniki oszacowania funkcji biasu z użyciem korelacji krzyżowej z soczewkowaniem mikrofalowego promieniowania tła (CMB). Otrzymujemy wynik w postaci $b_q(z) = 0.57^{+0.03}_{-0.03}(1+z)^2 + 0.07^{+0.06}_{-0.13}$, co dla $z = 1.5$ daje wartość biasu $3.63^{+0.25}_{-0.85}$. Detekcja korelacji krzyżowej wynosi 15σ, co jest obecnie jedną z najsilniejszych detekcji tego sygnału.

Acknowledgements

I would like to express my gratitude to my supervisor, prof. Agnieszka Pollo, who had faith in me when admitting me to the PhD studies, supported me in many different ways, and taught me a great deal about science. I would also like to thank my co-supervisor, prof. Maciej Bilicki, who helped me greatly, with scrutiny supervised my scientific endeavors, and from whom I learned a lot. During my studies, I met many students and workers from the Department of Astrophysics, the National Center for Nuclear Research (NCBJ), and other institutes in Poland and abroad, with whom I shared the most interesting scientific talks, as well as crazy laughs. I would like to thank Dorota Dobrowolska for continuous support in many formal matters, as well as numerous employes of NCBJ, from whom I received a lot of kindness. I would like to thank the Kilo-Degree Survey and Low-Frequency Array cosmological collaborations. I would like to thank my friends, family, and many other people who matter greatly and were there with me.

The work is based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017 and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD and the KiDS production team, on behalf of the KiDS consortium. Funding for SDSS-III was provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science.

I was financially supported by NCBJ, Polish National Center for Science (NCN), and Polish Ministry of Science and Higher Education (MNiSW), through grants DIR/WK/2018/12 (MNiSW), 212727/E-78/M/2018 (MNiSW), and UMO-2018/31/N/ST9/03975 (NCN).

Contents

To the young boy who never stops dreaming.
To the man who challenges fear to achieve them.

1 Introduction

One of the key goals of the ongoing and planned wide-angle sky surveys is to map the large-scale structure (LSS) of the Universe and derive various cosmological constraints, using different probes such as galaxy clustering or gravitational lensing. The building blocks of the LSS are galaxies, and among them, quasars (QSOs) stand out as some of the most distant objects we can observe. Unlike regular galaxies, these extragalactic sources cannot be easily identified based on their angular sizes because similarly to stars, they are mostly point-like. We observe QSOs up to very high redshifts because of the accretion of matter on supermassive black holes (Kormendy and Ho, 2013), which leads to enormous amounts of energy being radiated out. Quasars are important for LSS studies as they reside in dark matter halos of masses above $10^{12} M_{\odot}$ (Eftekharzadeh et al., 2015; DiPompeo, Hickox, and Myers, 2016), which makes them highly biased tracers of the LSS (DiPompeo et al., 2014; Laurent et al., 2017). Possible applications of QSOs in cosmology include tomographic angular clustering (Leistedt, Peiris, and Roth, 2014; Ho et al., 2015), the analysis of cosmic magnification (Scranton et al., 2005), bias estimation and measurement of halo masses (Shen et al., 2009; Oogi et al., 2016; DiPompeo et al., 2017; Laurent et al., 2017), cross-correlations with various cosmological backgrounds (Sherwin et al., 2012; Cuoco et al., 2017; Stölzner et al., 2018), and even the calibration of the reference frames for Galactic studies (Lindegren et al., 2018).

At any cosmic epoch, QSOs are sparsely distributed in comparison to inactive galaxies. Therefore, wide-angle surveys are essential to obtain catalogs containing a sufficient number of QSOs to be useful for studies where good statistics are important. Previous spectroscopic surveys, such as the 2dF QSO Redshift Survey (2QZ, Croom et al., 2004) or the Sloan Digital Sky Survey (SDSS, York et al., 2000; Lyke et al., 2020), provided $\sim$ 10^4-10^5 QSOs. In spectroscopy, QSO detection and redshift measurement are based on broad emission lines such as [OIII]λ5007/Hβ, [NII]λ6584/Hα (Kauffmann et al., 2003; Kewley et al., 2013). Many surveys exploit this approach, including: 2QZ, 2dF-SDSS LRG, and QSO (2SLAQ, Croom et al., 2009), SDSS, or the forthcoming DESI (DESI Collaboration et al., 2016) and 4MOST (de Jong et al., 2019; Merloni et al., 2019; Richard et al., 2019).

Object type and redshift are more difficult to extract from photometric broad-band surveys. However, photometric surveys are often the only feasible approach, particularly for large-scale structure (LSS) studies, which require a high number density and completeness as well as samples of millions of objects. Upcoming large photometric surveys, such as the Vera Rubin Observatory Legacy Survey of Space and Time (LSST, Ivezić et al., 2019), will provide an unprecedented number of objects and depth of observations.

Spectral energy distribution (SED) fitting is a standard approach to analyze photometry of galaxies with active galactic nuclei (AGN), which include QSOs in particular. It allows one to derive physical properties (Ciesla et al., 2015; Stalevski et al., 2016; Calistro Rivera et al., 2016; Yang et al., 2020; Małek et al., 2020) and estimate photo-zs (Salvato et al., 2009; Salvato et al., 2011; Fotopoulou et al., 2016; Fotopoulou and Paltani, 2018). The QSO selection in photometry is commonly based on color-color cuts (Warren, Hewett, and Foltz, 2000; Maddox et al., 2008; Edelson and Malkan, 2012; Stern et al., 2012; Wu et al., 2012; Secrest et al., 2015; Assef et al., 2018). More sophisticated and arguably more robust approaches to QSO selection are the probabilistic methods (Richards et al., 2004; Richards et al., 2009a; Richards et al., 2009b; Bovy et al., 2011; Bovy et al., 2012; DiPompeo et al., 2015; Richards et al., 2015), while machine learning (ML) has been gaining popularity in this respect as well (Brescia, Cavuoti, and Longo, 2015; Carrasco et al., 2015; Kurcz et al., 2016; Nakoneczny et al., 2019; Logan and Fotopoulou, 2020). Machine learning models have also been applied to derive QSO photometric redshifts (photo-zs, Brescia et al., 2013; Yang et al., 2017; Pasquet-Itam and Pasquet, 2018; Curran, 2020).

In the context of the Kilo-Degree Survey (KiDS, de Jong et al., 2013), which is the focus of our work, the QSO-related studies have so far dealt with high-redshift ($z \sim 6$) QSOs (Venemans et al., 2015), heavily reddened QSOs (Heintz et al., 2018), and selecting QSOs to search for strong-lensing systems (Spiniello et al., 2018; Khramtsov et al., 2019). We note that, in general, every QSO present in KiDS multiband catalogs has a redshift estimate derived with the Bayesian Photometric Redshift code (BPZ, Benítez 2000), as such photo-zs are computed by default for each cataloged object. However, these redshifts are usually not correct for QSOs as their derivation is optimized at galaxies used for weak lensing studies (Kuijken et al., 2015) and in particular proper AGN templates are not used in the BPZ implementation. Similarly, the KiDS database does not offer any direct indication of which sources could potentially be QSOs.

Here, we create quasar catalog from KiDS Data Release 3 (DR3, de Jong et al., 2017) with an machine learning (ML) approach based on optical data, and we focus on validating the resulting catalog (Nakoneczny et al., 2019). In case of the KiDS DR4 (Kuijken et al., 2019), we employ a more complex approach (Nakoneczny et al., 2021), where additionally to classification, we also estimate the photometric redshifts based on optical and near-infrared (near-IR) broad-bands available from the partner VISTA Kilo-degree Infrared Galaxy (VIKING, Edge et al., 2013). Additionally, we propose a methodology which allows to test and optimize the models for extrapolation, and test the resulting catalog at fainter magnitudes. Finally, we analyse the correlation function of the catalog resulting from the KiDS DR4, and estimate the quasar bias function using cross-correlation with the CMB lensing map (Planck Collaboration et al., 2020b).

In our approach to the KiDS DR3, we make the first step towards systematic studies of the KiDS quasar population by presenting automated detection of QSOs in the KiDS survey. For that purpose we employ one of the most widely used supervised machine learning algorithms, random forest, to detect QSOs in KiDS imaging in an automated way. The model is trained and validated on spectroscopic quasar samples which overlap with the KiDS DR3 footprint. We put special emphasis on selecting the most informative features for the classification task, as well as on appropriate trimming of the target dataset to match the training feature space and avoid unreliable extrapolation. This is also aided by analysis of two-dimensional projection of the high-dimensional feature space. The trained algorithm is then applied on the photometric KiDS data, and the robustness of the resulting quasar selection is verified against various external catalogs: point sources from the Gaia survey, as well as spectroscopic and photometric quasar catalogs, which were not used for training or validation. We also verify the number counts as well as mid-IR colors of the final QSO catalog.

In our approach to KiDS DR4, our main goal is to create a catalog of QSOs with robust photometric redshift estimates. We test what near-IR imaging brings to classification in terms of separating QSOs from stars. We aim to fit ML models for the best bias versus variance trade-off in order to achieve reliable results at the faint data end, not represented well by the spectroscopic data used in training. We verify whether randomly selected subsets of spectroscopic objects used to test ML models lead to the proper bias-variance trade-off, or if it is better to also validate based on the faintest objects, which are never seen during training. This is necessary to assess the level of overfitting, address the problem of extrapolation in the feature space (a space of n-dimensional feature vectors consisting of, for instance, magnitudes and colors), and provide reliable estimates at the faint data end. We test different strategies of building features from broad-band magnitudes, find which of the most popular ML models perform best for classification and redshifts, and model QSO photometric redshift uncertainties with a Gaussian output layer in an Artificial Neural Network (ANN).

The thesis is organized as follows. In Chapter 2 we describe the data and the methodology for QSO selection, redshift estimation, extrapolation in the feature space, bias-variance tuning, correlation analysis, and quasar bias estimation; in Chapters 3 and 4, we provide results of experiments done on a cross-match with spectroscopic data, properties of the final catalog, and purity-completeness calibration for KiDS DR3 and DR4 catalogs respectively; in Chapter 5 we provide results of correlation functions measurement and quasar bias constraints; in Chapter 6 we discuss the main findings, strengths, and weaknesses of the approach, and we outline the possible extensions. In Chapter 4, in order to calculate spatial densities, we use the flat ΛCDM cosmology based on the Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP9, Hinshaw et al., 2013) with $H_0 = 69.3$ km/s/Mpc and $\Omega_m = 0.287$, while in Chapter 5, where we perform the correlation analysis, we use flat ΛCDM cosmology based on the Planck Collaboration et al., 2020a with $H_0 = 67.4 \mathrm{km/s/Mpc}$ and $\Omega_m = 0.315$.

2

Data and methodology

This chapter is based on the publications Nakoneczny et al., 2021; Nakoneczny et al., 2019, and all the work was done by the thesis autor, unless stated otherwise. In this chapter we describe the methodology which we use to create quasar catalogs from the KiDS DR3 and DR4, and perform their auto- and cross- correlation with the CMB lensing in order to constraint the quasar bias function. We create the quasar catalog from KiDS DR3 with a simpler approach based on optical data, and we focus on validating the resulting catalog. In case of the KiDS DR4, we employ a more complex approach, where additionally to classification, we also estimate the photometric redshifts based on optical and near-IR data available starting with the KiDS DR4. Additionally, we propose a methodology which allows to test and optimize the models for extrapolation, and test the resulting catalog at fainter magnitudes. The chapter is organised as follows. In §2.1, we describe the KiDS survey, as well as inference and training sets. We explain the machine learning related methodology in §2.2, while machine learning pipelines for KiDS DR3 and DR4 are given in the §2.3 and §2.4 respectively. Finally, in §2.5, we provide methodology related to the correlation analysis and quasar bias estimation.

2.1 Data

2.1.1 Kilo-Degree Survey

KiDS[1] is an optical wide-field imaging survey with the OmegaCAM camera (Kuijken, 2011) at the VLT Survey Telescope (VST, Capaccioli et al., 2012), specifically designed for measuring weak gravitational lensing by galaxies and large-scale structure (Joudaki et al., 2017; van Uitert et al., 2018; Asgari et al., 2020; Heymans et al., 2020; Hildebrandt et al., 2020; Wright et al., 2020). It consists of 1350 square degrees imaged in four broad-band *ugri* filters. The mean limiting AB magnitude (5 σ in a 2 arcsec. aperture) of KiDS is ~ 25 in the r band. The Data Release 3 (de Jong et al., 2017) covers ~ 447 deg^2 and includes almost 49 million sources in its multiband catalog. The current fourth data release (Kuijken et al., 2019) is the penultimate one; it covers a total of 1006 deg^2 and provides a list of ~ 100 million (100M) objects based on the r-band detections. It also includes $ZYJHK_s$ photometry from the partner VIKING. The optical depth, wide sky coverage, and multiwavelength imaging make this survey an ideal resource for QSO science.

Additional discriminatory power for QSO selection and photo-zs could be provided from mid-infrared bands, such as from the Wide-field infrared Survey Explorer (WISE, Wright et al., 2010), as shown for instance in Logan and Fotopoulou, 2020. However, in this work we decide to limit the selection to the 9-band KiDS+VIKING only, as adding the WISE data

[1] http://kids.strw.leidenuniv.nl

would severely limit our dataset. For instance, at the $r < 23.5$ KiDS limit, which we find as a magnitude limit for our quasar catalog (§3.4), only ~19% of the sources have a counterpart in WISE within a 3" matching radius. This fraction decreases even further with an increased KiDS depth.

2.1.2 Inference set in the optical KiDS DR3

The main features used in the classification process come directly from the KiDS catalog and consist of the *ugri* magnitudes and the corresponding colors. As detailed in de Jong et al., 2017, KiDS data processing provides various photometric measurements of detected sources. For our purpose we need robust measurements of point source photometry, we therefore use the GAaP magnitudes (Gaussian Aperture and PSF, Kuijken, 2008) which are designed to compensate for seeing variations among different filters. Together with additional photometric homogenization across the survey area, these measurements provide precisely calibrated fluxes and colors (Kuijken et al., 2015). Combining the four magnitudes and six colors (one for every magnitude pair) results in ten basic features. Although using both magnitudes and colors seems redundant with respect to using for example only magnitudes or only colors, such redundancy does improve classification results. As detailed later in §2.2.2, we also tested ratios of magnitudes as additional parameters for the classification. Together with the stellarity index `CLASS_STAR` (see below), the magnitudes, colors and magnitude ratios constitute the eventual 17D feature space for classification, in which only the most relevant features are used.

We note that although a significant fraction of KiDS sources are stars, we always use magnitudes and colors corrected for Galactic extinction, as our focus is to detect extragalactic sources. The measurements are also corrected for the "zero-point offset" to ensure flux uniformity over the entire DR3 area[2].

Another parameter used in the classification process, which turns out to be very important for the performance (§3.1), is `CLASS_STAR`. This is a continuous stellarity index derived within the KiDS data processing pipeline using SExtractor (Bertin and Arnouts, 1996), which describes the degree to which a source is extended. It takes values between 1 (point-like, a star) and 0 (extended, typically a galaxy). Most quasars, except for the rare ones with a clearly visible extended host, are point-like and therefore have high values of `CLASS_STAR`. The reliability of this parameter for separation of point sources from extended ones depends on the signal-to-noise (S/N) of the objects and resolution of the imaging, therefore it may fail to identify faint and small galaxies (in terms of apparent values). However, the data used for the quasar detection in this approach are limited to a relatively bright and high S/N subsample of KiDS DR3 (as we explain later in this section), where `CLASS_STAR` is robust in separating these two main source classes. Indeed, we have verified that for the training set with known labels (§2.1.4)[3], separation at `CLASS_STAR` of 0.5 would leave a negligible fraction (0.5%) of galaxies marked as "point-like" and a small number (1.9%) of stars identified as "extended". As far as the training-set QSOs are concerned, the vast majority of them have `CLASS_STAR` > 0.6. Some quasars with resolved hosts may still be detected as extended, especially in a survey with such excellent angular resolution as KiDS. We therefore decided to test the usefulness of `CLASS_STAR` in the quasar automated selection and indeed found it very helpful. This is discussed in more detail in §3.1. We would like to emphasize, however, that in this approach no a priori cut on `CLASS_STAR` is made in either the training or inference sample. The QSO classification algorithm filters out most of the

[2] These offsets are on average ~ 0.03 mag in the u and i bands, and < 0.01 mag in the g and r bands. See de Jong et al., 2017 for details.

[3] This applies equally to validation and test sets, as they are chosen randomly from the general SDSS labeled sample.

TABLE 2.1: Numbers of objects left in the KiDS DR3 inference data after the subsequent preprocessing steps. See text for details of these cuts.

	objects left	% of all data
all KiDS DR3 sources	49M	100%
keep only four-band detections	40M	80%
cut at limiting magnitudes & remove errors > 1 mag	11M	22%
clean up image flags	9M	18%
cut at $r < 22$	3.4M	6.8%

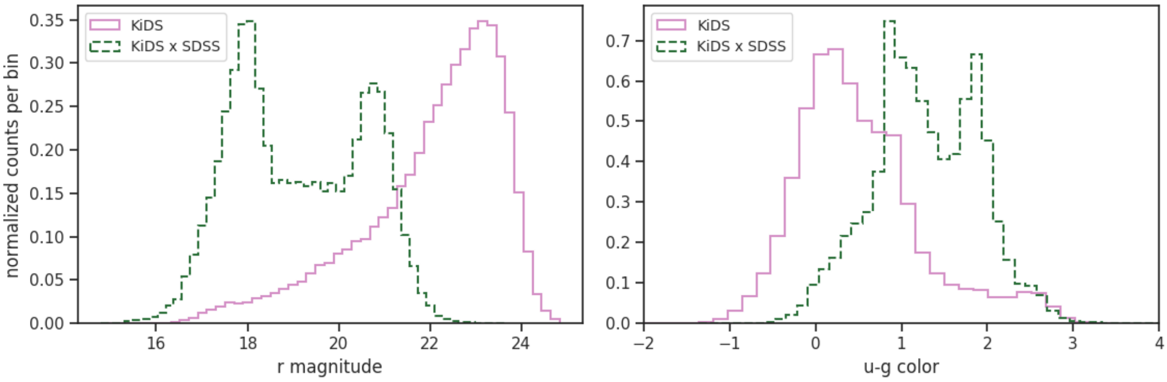

FIGURE 2.1: Normalized histograms of the r-band magnitude (left) and the $u - g$ color (right) for the KiDS DR3 inference dataset (pink solid) and the KiDS DR3 × SDSS training set (green dashed), both before applying the $r < 22$ mag cut. The bimodality of the matched sample is related to SDSS spectroscopic target preselections, preserved by the cross-match with the much deeper KiDS.

extended sources, but does identify a fraction of them as quasars. Roughly 1% of the objects with CLASS_STAR < 0.5 are assigned a QSO label (20k out of 2M). This means that ~11% of all our quasar candidates were classified by KiDS as extended.

The KiDS data processing pipeline provides also another star/galaxy separator, SG2DPHOT (de Jong et al., 2015), which uses the r-band source morphology and generally is more robust than CLASS_STAR. We have however found that in this particular approach, using SG2DPHOT instead of CLASS_STAR gives slightly worse classification results. Part of the reason might be that SG2DPHOT is a discrete parameter, and it provides less information to the model than the continuous CLASS_STAR.

An additional potentially useful feature in the classification process could be photometric redshifts (photo-zs). However, KiDS DR3 photo-zs were optimized for galaxies (de Jong et al., 2017; Bilicki et al., 2018) and are unreliable for quasars, as we indeed verified for overlapping spectroscopic QSOs. In our approach for KiDS DR4, we address the issue by deriving more robust QSO photo-zs (see section 2.4 and chapter 4, also see e.g., Yèche et al., 2010).

Machine-learning methods require data of adequate quality in order to perform reliably. In addition, in supervised learning, it might be desirable to avoid extrapolation beyond the feature space covered by the available training set. For those reasons, we apply appropriate cleaning and cuts on KiDS DR3 data as specified below, to ensure reliable quasar classification.

1. *Keep only four-band detections*

A fraction of KiDS DR3 sources do not have all the four bands measured. As Machine-learning models require all employed features to bear a numerical representation, using sources with missing features would require assigning some artificial values to relevant magnitudes and colors. As already one missing magnitude means three colors are not available, even in such a minimal scenario it would significantly reduce the available information. As we show in §3.1, colors are in fact among the most important features in our classification procedure, and we cannot afford to lose them. Therefore, we only use objects with all the four *ugri* bands measured. This step removes about 20% of the catalog entries, leaving roughly 40 million sources (Table 2.1).

2. *Ensure sufficient signal-to-noise ratio*

   To avoid working with excessively noisy data which could significantly affect classification performance, we first cut the KiDS DR3 data at the nominal limiting magnitude levels, which are 24.3, 25.1, 24.9, 23.8 in *ugri*, respectively. Additionally, we require the photometric errors in each band to be smaller than 1 mag, which roughly corresponds to S/N of 1. These two cuts applied simultaneously on all the bands (i.e., as a joint condition) are the most strict among other cleaning requirements and together with the above item #1, leave ∼ 11 million KiDS DR3 objects. However, we note that due to training set limitations, a further and stricter cut on the r band magnitude is applied below.

3. *Remove flagged objects*

   The KiDS data processing pipeline provides various flags for detected objects, indicating possible issues with photometry (see de Jong et al., 2017 for details). In this work, we take into account `IMAFLAGS_ISO_band` which are flags delivered by the KiDS pipeline and include information about critical areas in the images, which likely corrupt single source photometry. Issues such as star halos and spikes are automatically detected using an algorithm that first finds star positions from the saturation map, and then builds models of star halos and spikes taking into account the telescope orientation and the position in the focal plane (the Pullecenella mask procedure, de Jong et al., 2015). We have carefully examined several images of objects with these flags set and decided to remove sources indicated by any of the bits[4] except for the one for manual masking, which was valid only for KiDS DR1 and DR2 (de Jong et al., 2015). We have also considered `FLAG_band` (a SExtractor flag indicating possible issues in the extraction of the object) but found no significant deterioration in classification quality after including sources marked by this flag in the training and inference process. This cleanup step removes a non-negligible number of sources, about 2 million out of 11 million that were left after the previous step #2.

4. *Trim target data to match the training set*

   The previous steps of data cleaning were of a general nature to ensure adequate KiDS data quality for classification purposes. A final step is however required and it is specific to the training sample used. Namely, the SDSS DR14 spectroscopic training set (§2.1.4) does not reach beyond $r \sim 22$ mag (see Fig. 2.1 left panel). Using significantly deeper inference data than for the training set would require extrapolation, which for supervised ML might be unreliable and more importantly, its performance would be

[4] These are: 1 - saturation spike, 2 - saturation core, 4 - diffraction spike, 8 - primary reflection halo, 16 - secondary reflection halo, 32 - tertiary reflection halo, 64 - bad pixel.

> difficult to evaluate. We address this problem in the approach to KiDS DR4, but here, we trim the KiDS DR3 data to limit the feature space to ranges covered by training. As the default cut we adopt $r < 22$, although we have also experimented with a more permissive cut in the $u - g$ color, as matching the training and inference sets in this parameter leads to the removal of fewer objects (see Fig. 2.1 right panel). We provide more discussion on feature space limitation and our final choices in §2.3.2.
>
> This particular cut significantly limits the size of usable data for the classification, leaving us with 3.4 million KiDS DR3 objects. Such a cut would in fact make unnecessary the condition on the sufficient S/N described in item #2 above, as KiDS sources with $r < 22$ mag typically have very high S/N in all the bands. However, we keep this condition separately, as it is related to the characteristics of the specific training set used.
>
> We note that with future deeper QSO training sets, such as DESI (DESI Collaboration et al., 2016) and 4MOST (de Jong et al., 2019), it will be possible to extend the range of the usable feature space and therefore increase the number density of robustly identified quasars in KiDS.

Table 2.1 summarizes all the preprocessing steps which led to the creation of the final dataset on which our quasar catalog is based. We denote this dataset as the target or inference sample.

2.1.3 Inference set in the optical and near-IR KiDS DR4

In the case of the fourth data release of KiDS survey, which includes the near-IR imaging, we solved the problem of QSO detection with classification models and derived photo-zs with regression models. We created one feature set for both classification and regression to keep the models consistent in predictions. We limited the KiDS data to 9-band detections (sources which have all the nine bands measured) in order to provide the most reliable set of features (Section 4.2). In this approach, we do not exclude objects based on image flags or noise, as those are mostly not present in the bright training data, and we want to make inference on them to test the extrapolation capabilities of our models in the fainter KiDS data. During the experiments, we calculated all of the scores, including completeness, in the limited set of 9-band detections, but the number counts of the final catalog compare completeness with respect to all possible QSOs. The feature set includes nine GAaP magnitudes, 36 colors, 36 ratios of every magnitude pair, and the following two morphological classifiers: CLASS_STAR, and the third bit of SG2DPHOT[5] (de Jong et al., 2015; de Jong et al., 2017). In this approach, we divided the SG2DPHOT into separate bits, and found its third bit improving the classification results. In Section 4.2 we describe the experiments which led to this final set of 83 features. The 9-band detection requirement reduces the number of objects from ~100M to ~ 45M, which creates the inference set.

2.1.4 Training sets from the Sloan Digital Sky Survey

To learn object classification based on photometric data, our machine learning model needs ground-truth labels. In this work, the labels are taken from the SDSS, and in particular from the spectroscopic catalog of its Data Release 14 (Abolfathi et al., 2018)[6]. It includes over 4.8

[5] Flag values are: 1 (high-confidence star candidates), 2 (objects with FWHM smaller than stars in the stellar locus), 4 (stars according to S/G separation), and 0 otherwise (galaxies); flag values are summed. See sect. 4.5.1 of de Jong et al., 2015 for details.

[6] More recent SDSS DR16 does not provide additional overlap with KiDS with respect to DR14.

million sources with one of three labels assigned: star, galaxy, or quasar. To ensure the highest data quality and model performance, from that sample we use only sources with secure redshifts (velocities) by demanding the zWarning parameter to be null. The cross-match between the SDSS and KiDS inference dataset was done with the TOPCAT tool (Taylor, 2005) within a matching radius of 1”.

In KiDS DR3, the final training dataset consists of 12,144 stars, 7,061 quasars and 32,547 galaxies, which totals to 51,752 objects. These relatively low numbers are the consequence of the small sky overlap between SDSS and KiDS DR3, as the common area covers only the KiDS equatorial fields at $-3° < \delta < 3°$. In the test phase, from these labeled data we randomly extract the actual training, validation, and test sets, as detailed in §2.3.1. For training the final classification model, the entire spectroscopic cross-matched sample is used. Therefore, whenever parameter distributions or feature space properties are discussed for “training”, this applies equally to the general training data, as well as to the validation and test sets.

In Fig. 2.1 we present normalized distributions of the r-band magnitudes (left panel) and the $u - g$ color (right panel) for the KiDS DR3 inference dataset (pink solid) and the KiDS DR3 × SDSS training set (green dashed), both before applying the $r < 22$ mag cut. This shows clearly that the current SDSS data do not probe KiDS beyond the adopted r-band cut. On the other hand, the color space is better covered between the inference and training sample, although here we only showed one particular color as an example. The matching between the training and inference (target) set in the multidimensional feature space is discussed in more detail in Section 2.3.2. The bimodality of the KiDS × SDSS seen in both histograms is related to preselections of the SDSS spectroscopic targets at the various stages of the survey. In particular the flux-limited ($r < 17.77$) complete “SDSS Main” sample (Strauss et al., 2002) gives the first peak in the r-band histogram, while subsequent BOSS color selections used fainter magnitudes (Dawson et al., 2013). As KiDS is much deeper than any of the SDSS spectroscopic subsamples, the cross-match preserves these properties.

In the case of KiDS DR4, the training set was similarly derived from cross-matching the inference data with the SDSS DR14 spectroscopic observations. After removing objects flagged with warnings by SDSS, we obtained a training subset of 152k objects (69% galaxies, 11% QSOs, 20% stars). The training set is also limited to $r \sim 22$ by SDSS (99% of training is at $r < 21.98$), which is about three magnitudes brighter than the depth of the KiDS inference data.

2.2 Machine learning

2.2.1 Models

We tested three of the most popular ML models: random forest (RF, Breiman, 2001), XGBoost (XGB, Chen and Guestrin, 2016), and artificial neural network (ANN, Haykin, 1998). We used Python libraries: SCIKIT-LEARN, TENSORFLOW (Abadi et al., 2015), and KERAS (Chollet, 2015). The RF and XGB are ensemble models, in which classification or regression is performed using many decision trees. The RF randomizes the trees by choosing a subset of training data and/or features for each tree. The XGB introduces the boosting procedure which favors selection of data points for which the model has the highest errors. Additionally, it uses gradients to approximate and minimize an error function. The ANNs consists of stacked layers of neurons, with nonlinear activation function in each neuron.

We tested two redshift estimation strategies: one model for all the classes and two specialized models trained separately for quasars and galaxies. In case of the specialized models, we assigned zero redshift to stars. We also tested a neural network model with multiple outputs for classification and redshifts, which allowed us to solve both problems with only one model.

2.2.2 Feature engineering

Feature engineering is one of the ways to tune model complexity, and it is widely used in ML practice (see Bishop, 2006, chap. 6). Already in simpler models, such as linear regression, it allows for increased nonlinearity by applying a kernel. In more complicated models, it leads to better adaptation to the given training data and allows the models to extract the true patterns from the inference data. Feature engineering should not be considered as a limitation of ML models, but it is a consequence of adjusting the bias versus variance trade-off. However, excessive feature engineering can lead to overfitting; therefore, a reliable testing method is required for this approach to work properly and to match its strategy with proper model regularization.

In applications with hundreds of features, the process of choosing their best subset can be complex. Here, we use a method of backward elimination (Harrell, 2001). We start with all the available features, even those that, according to our knowledge, may not be very important, and apply a model which calculates feature importance (such as for instance the RF). After initial training, we sort all the features according to their importance and perform iterative removal of the least important ones (in some groups rather than individually). After each removal, we validate the performance of the new model. In this way, within a linear complexity with respect to the number of features, we find the best feature set and optimize model performance.

Feature importance for a single decision tree (DT) is calculated by simply summing all the information gain (IG) from a given feature over the whole tree. For the full RF, this value is averaged between all the DTs for each feature. From this, relative importance of features can be calculated as percentages, providing quantitative information on their usefulness in solving a problem.

Machine learning algorithms work more efficiently if they are provided not only with basic features but also their combinations, if those are correlated with the ground truth data like labels in case of a classification. A popular way to extend the feature set is to combine already existing features using simple algebraic operations (Piramuthu and Sikora, 2009). Colors (differences of magnitudes from various passbands) are one of such ways; another combination popular in ML is feature ratios, and we tested ratios of magnitudes as an extension of the feature space. The ratios are widely used in ML and also in astronomy (D'Isanto et al., 2018).

In the case of KiDS DR4, we also kept the feature engineering fairly simple by considering only the input features and, for magnitudes, their simplest combinations: differences (colors) and ratios. More complex feature engineering is possible, but we found this strategy sufficient to obtain good results, without a risk of overfitting according to our testing methods. In DR4, we rank features by their feature importance from XGB models, which was calculated as a sum of gain that a given feature provides to a model in all the splits which are made based on that feature. We chose XGB in this approach, as it provides better results in comparison to RF, in case of this more complicated problem which we state for the KiDS DR4 data. Additional complexity comes from available near-IR imaging, and more complex validation pipeline, as described below.

2.3 Classification pipeline for the optical KiDS DR3

Our goal is to detect quasars in KiDS DR3 data, however the training sample from SDSS provides labels for 3 types of sources: stars, galaxies, and QSOs. These objects usually populate different regions of the feature space which we use, and we have verified that the QSO identification is more robust if the model is formulated as a three-class rather than a binary (QSOs vs. the rest) problem. This is an expected result as in the three-class case we provide the model with more information.

2.3.1 Validation procedure

In ML methods it is desirable to separate out validation and test datasets from the training sample, in order to estimate model performance on observations which were not included in model creation. The validation set is used to select the best model and its parameters, while the test set is needed to report final scores and should never be employed to choose algorithm parameters, in order to eliminate the possibility of overfitting to a particular training sample. In practice, if validation and test scores are significantly different, more regularization should be applied to a model. In our application, as the test set we choose a random 20% subsample of the full training set described in Section 2.1.4. The remaining 80% of the training data are then used in a five-fold cross-validation procedure, in which they are divided into five separate equally-sized subsets, and four of them are used for the training, while validation scores are calculated on the fifth subsample. The training process is repeated five times, with a different subset used for the validation each time. This gives a total of five values for every metric used, which are then averaged to create the final validation results.

As far as the evaluation metrics are concerned, we use several of them in order to quantify model performance better than would be possible with individual scores. The basic metric used for three-class evaluation is the accuracy, which measures the fraction of correctly classified observations. Additionally, as the main goal of this work is to select quasars, we transform the classification output into a binary one. This is done by simply summing probabilities of stars and galaxies into a new class called `rest`, and evaluating the performance of the QSO vs. `rest` problem. To this aim, apart from accuracy applied on the binary problem, we use the purity (precision), completeness (recall), and F1, a harmonic mean of precision and recall. If the true positive (TP) is the number of correctly classified positives, false positive (FP) is the number of incorrectly classified positives, false negatives (FN) the number of incorrectly classified negatives, then the metrics are given by:

- $purity = TP/(TP+FP)$;
- $completeness = TP/(TP+FN)$;
- $F1 = 2 \cdot purity \cdot completeness/(purity + completeness)$.

In our case, the positive class consists of quasars, and the negative class of stars plus galaxies. The last binary metric we use is the area under the receiver operating characteristic curve (ROC AUC) based on the output probability for the quasar class only. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various probability threshold settings. TPR is the same value as completeness, while FPR is also known as the probability of false alarm and can be calculated as $FPR = 1 - specificity = 1 - TN/(TN+FP)$. The last validation tool used here is the confusion matrix, which shows relations between ground-truth and predicted labels for the multiclass problem. These metrics are used in our machine learning experiments both to select the most appropriate algorithm and set of features.

2.3.2 Feature space limitation

In addition to selecting the most relevant features, we need to make sure that the training data cover the feature space sufficiently well for the classification in the inference data to be robust. We already discussed in Section 2.1.4 that the SDSS training data are much shallower than the full KiDS, therefore in this approach the KiDS data is limited to $r < 22$ mag to avoid extrapolation. In this subsection we provide details on feature space limitation by analyzing its full multidimensional properties.

One can understand machine learning models as complex decision boundaries in the training feature space. The models are expected to learn true patterns, which should then extend their applicability to new datasets, such as the KiDS inference sample in our case. However, for the points which lie outside of the original region of feature space for which decision boundaries were created, model predictions may implement a classification function extrapolated from the training data, which may then not agree with the patterns outside of the training set. The most straightforward solution is to simply match the inference dataset to the training sample. In our case, the simplest approach is to limit the KiDS DR3 data to $r < 22$ mag. However, one could also work in color space only, without using magnitudes, and perform a cut of $u - g > 0$ instead (see right panel of Fig. 2.1), which would significantly extend the inference dataset, giving about twice as many entries than the 3.4 million in the fiducial sample limited to $r < 22$. Below, we show why a cut in the r-band magnitude is more appropriate for our model than a cut in the $u - g$ color.

Cuts performed in single features allow for a better match between the training and inference sets in these particular dimensions. However, as the final classification is performed in a space of significantly larger dimensionality, the usefulness of such an approach is rather limited. A match between individual features does not have to imply proper coverage of the full feature space. A simple counterexample is a 2D square covered by data points drawn from a 2D Gaussian distribution and separated into two subsets by a diagonal. In such a case, the histograms of single features show overlap of data in individual dimensions, while in fact there is no data from two subsets overlapping in 2D at all. Therefore, we look in more detail at coverage in the multidimensional feature space of the training and inference data. This is done by projecting the feature space onto two dimensions using the t-distributed stochastic neighbor embedding (t-SNE, Maaten and Hinton, 2008) method.

There are many ways of mapping N-dimensional feature spaces onto 2D projections. A popular one in astronomy is Self Organizing Map (SOM, Kohonen, 1997), and a relevant example of its usage is the mapping of multicolor space to visualize which regions are not covered by spectroscopic redshifts (Masters et al., 2015). Here, we use an another advanced visualization method, the t-SNE, which finds complex nonlinear structures and creates a projection onto an abstract low-dimensional space. Its biggest advantage over other methods is that t-SNE can be used on a feature space of even several thousand dimensions and still create a meaningful 2D embedding. Moreover, unlike in SOM where datapoints are mapped to cells gathering many observations each, in t-SNE every point from the N-dimensional feature space is represented as a single point of the low dimensional projection. This makes t-SNE much more precise, allowing it to plot the exact data point values over visualized points as different colors or shapes, making the algorithm output easier to interpret. Some disadvantages of using t-SNE are its relatively long computing time and its inability to map new sources added to a dataset after the transformation process, without running the algorithm again.

In case of ML methods which use many features at once during the calculations, it is useful to normalize every feature in order to avoid biases related to their very different numerical ranges (such as for magnitudes vs. colors). This does not apply to RF, which uses only one feature in each step of data splitting; however, it does affect the t-SNE algorithm

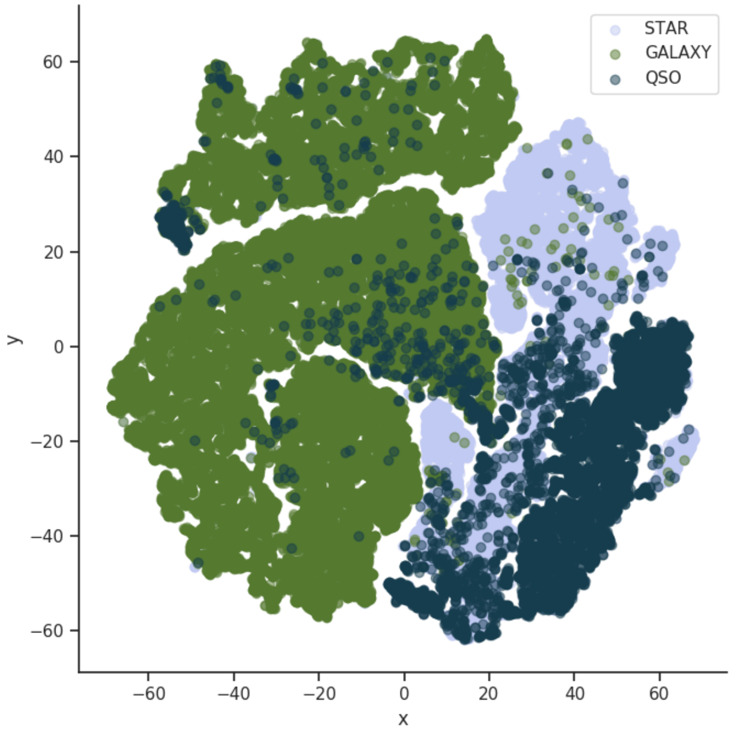

FIGURE 2.2: t-SNE visualization of the KiDS DR3 training dataset. The plot illustrates a projection of the multidimensional feature space onto a 2D plane, where *x* and *y* are arbitrary dimensions created during the visualization process. Labeled training data are mapped with three different colors as in the legend.

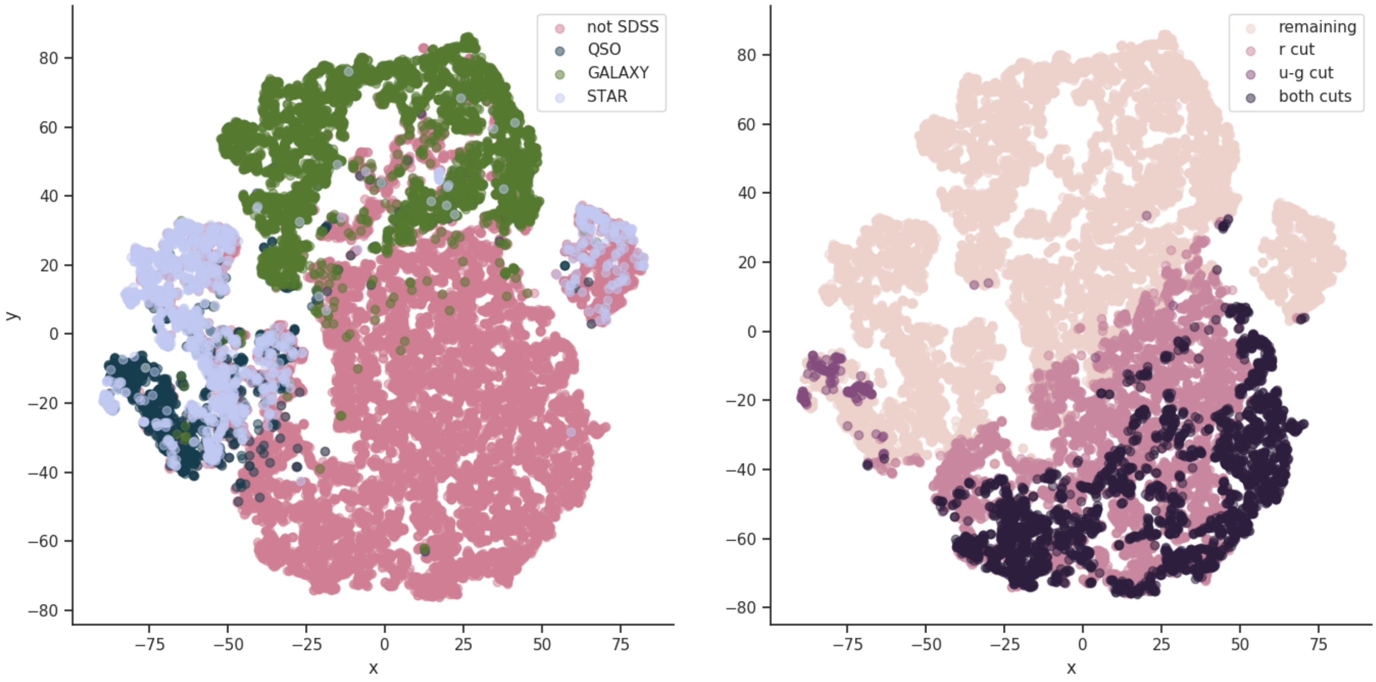

FIGURE 2.3: t-SNE visualization of the KiDS DR3 catalog feature space. *Left:* data points with SDSS labels from the training set (green and light and dark blue) together with those from the inference sample without training coverage (pink). *Right:* results of applying magnitude and color cuts on the inference sample. The darkest color shows objects removed by both $r < 22$ and $u - g > 0$ criteria simultaneously, colors in between stand for objects removed by only one of those cuts, and the lightest points are left after the cuts.

which calculates distances based on all available features. In order not to artificially increase the importance of the features with larger numerical values, we always scale every feature individually to the range $[0,1]$, as a preprocessing step in the visualization. The transformation of each feature is given by $F'_i = (F_i - F_{min})/(F_{max} - F_{min})$, where F stands for a given feature, F_i is its value for the i-th data point, and F_{min} and F_{max} represent the minimum and maximum values of this feature in a considered dataset.

Our first t-SNE visualization is applied to the KiDS DR3 training dataset, using the full 17D feature space selected in Section 2.2.2, and it gives important information whether the automated quasar detection can be performed at all in the feature space provided by KiDS DR3. As shown in Fig. 2.2, most of the quasars form their own cluster in the 2D projection, while some do indeed overlap with stars. This does not necessarily mean that those observations are not distinguishable by classification models which work in the original feature space, but it does point at a problem that perhaps additional features should be added, such as magnitudes and colors at other wavelengths than the currently used *ugri* ones. We study this issue in the extended approach for the KiDS+VIKING data in the fourth data release (section 2.4 and chapter 4).

We now turn to a comparison of the training and inference datasets. For that purpose we join the training set with a random subsample of the inference data of similar size as the training (this is to speed up the computation which for the full KiDS DR3 would be very demanding). In Fig. 2.3 we show projections of the full 17D feature space for the dataset constructed this way. Using a new dataset required creating a new visualization, meaning that x and y axis in this figure are independent of the ones present in the training set visualization (Fig. 2.2). The left panel includes SDSS labels for the training part of data (green and light and dark blue dots), and “not SDSS” (pink) standing for inference data which covers feature space outside of the training. Here, the inference data have no magnitude or color cuts applied except for those related to the basic data cleaning (items #1-#3 in Section 2.1.2). This visualization confirms that a large part of feature space in the inference dataset would not be covered by the training if no additional cuts were applied on the target KiDS sample.

In the right panel of Fig. 2.3 we illustrate the effect of magnitude and color cuts on the inference data. The darkest color indicates objects removed by both $r < 22$ and $u - g > 0$ criteria simultaneously, colors in between show objects clipped by demanding only a single cut, while the lightest points are left after applying any of the cuts. By comparison with the left panel, we clearly see that the r-band cut is much more efficient in removing the part of the feature space not covered by training than the cut in $u - g$. It is also worth noting that the color cut would also remove some quasar data points from the feature space covered by training, which is much less the case for the magnitude cut. This is related to the flux-limited character of the SDSS spectroscopic QSOs.

2.4 Classification and photo-z pipeline for the optical and near-IR KiDS DR4

2.4.1 Inference subsets

In this section we define the inference subsets for KiDS DR4 based on feature set considerations. The training set we use is a small subset of the KiDS inference data and does not fully cover the feature space. Inference on parts of the feature space not covered by the training data may result in the deterioration of results or a complete failure, due to new combinations of features or completely new feature values. For continuous features, such as magnitude,

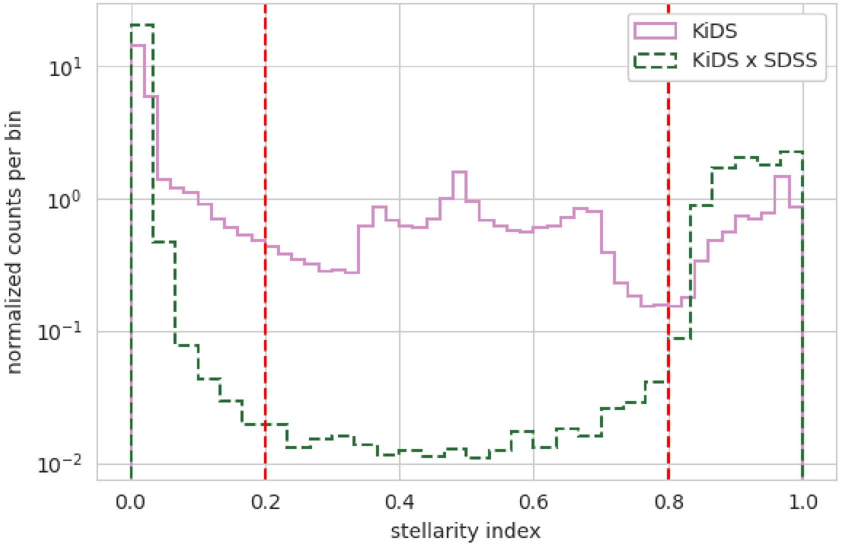

FIGURE 2.4: Normalized histograms of the CLASS_STAR stellarity index in the training (KiDS DR4 x SDSS) and inference (KiDS DR4) datasets. The intermediate values represent failures of the morphological classifier. Those values are not commonly present in the training data, thus we cannot expect the ML models to work correctly for objects with such index values. We consider the sources in between the red dashed lines as unsafe for the inference.

we may expect well-generalized models to extrapolate with deteriorating quality of the estimations. In case of discrete features, whose new values cannot be understood based on the ones available in training, supervised ML models may fail completely. We therefore define inference subsets based on how feature coverage changes from training to inference data and how this can affect the ML models.

The morphological classifiers tend to fail at the faint data end. We used them to achieve the highest accuracy at the bright end, and as a proxy for data quality at the faint end. The SExtractor-based stellarity index has a continuous distribution between zero and one, with intermediate values pointing to classifier failure. Because the failures are almost not present in the bright training data, ML models do not understand their meaning (Fig. 2.4). We therefore only consider the stellarity index ranges $(0, 0.2)$ and $(0.8, 1)$ covered by the training data as safe for the inference. Choosing cuts which admit more objects to the safe inference subset might increase completeness at the cost of purity, while stricter cuts do the opposite. We find empirically that using only the third bit of the SG2DPHOT provides the best improvement in our results. Cleaning the uncertain stellarity index values removes most of the SG2DPHOT failures.

The magnitude range $r < 22$ is covered by the training data, whereas for $r > 22$ we expect ML models to extrapolate with deteriorating quality. We define three inference subsets based on the feature space coverage, morphological classification quality, and the r-band depth of the survey. Firstly, the safe subset is $r < 22$ and CLASS_STAR $\notin (0.2, 0.8)$; secondly, the extrapolation subset is $r \in (22, 25)$ and CLASS_STAR $\notin (0.2, 0.8)$; and lastly, the unsafe subset is $r > 25$ or CLASS_STAR $\in (0.2, 0.8)$. It is a much more complex approach then the standard one which we employ for KiDS DR3.

We visualize the KiDS DR4 feature space and the inference subsets with t-SNE in Fig. 2.5. We created the visualization with the same set of 83 features that are used in classification and redshift estimation in order to visualize the same feature space. Due to the computational complexity of t-SNE, we took 8k random objects from KiDS data and merged them with 4k

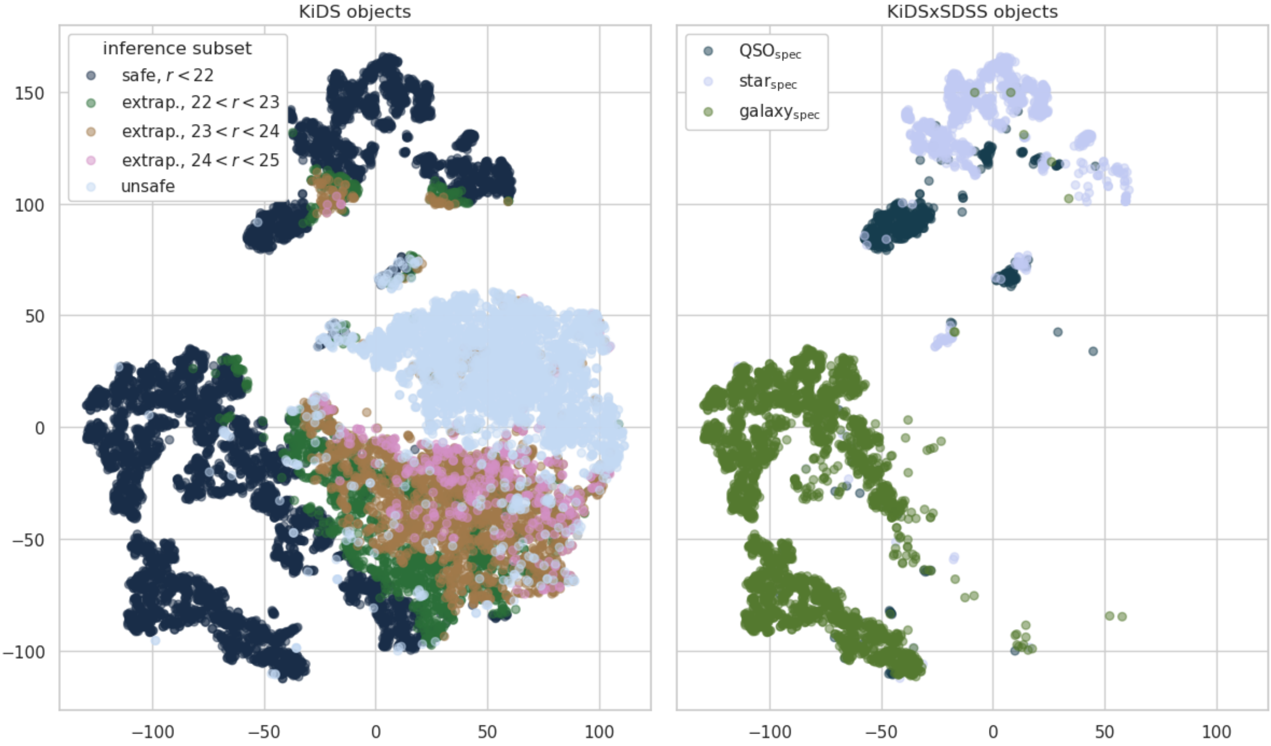

FIGURE 2.5: t-SNE projections. *Left:* KiDS DR4 inference subsets. *Right:* SDSS spectroscopic classification. The visualizations were made on subsets of 12k objects. The real density of objects at any part of the feature space is 3.8k times higher than visualized. We can see three main groups. The point-like objects cover the top part, extended ones are located at the bottom, and those with undetermined morphology are placed in the middle, in the unsafe subset. The spectroscopic data cover only the bright part of the photometric data; this illustrates the extrapolation problem to address with machine learning. The results of the inference are later investigated on similar plots (Section 4.5), which we consider a more robust approach than investigating color-color diagrams.

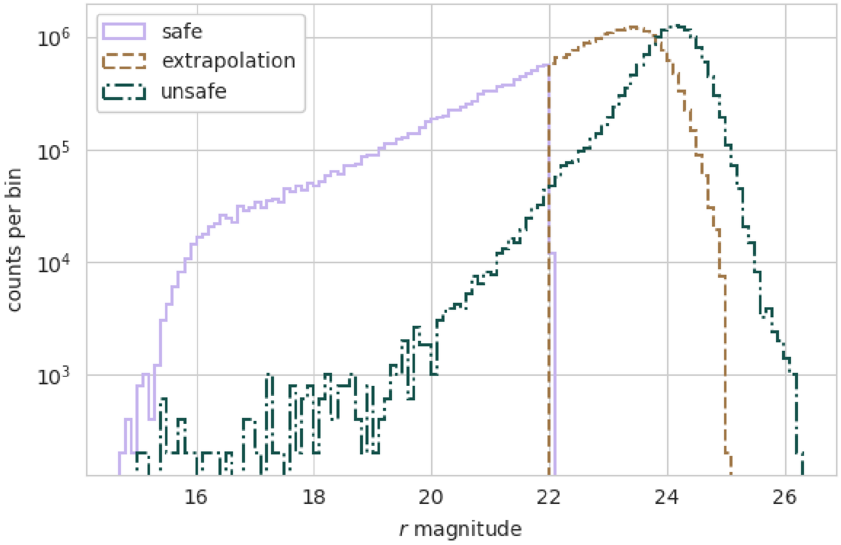

FIGURE 2.6: Distribution of the KiDS DR4 inference subsets over the r magnitude. The limit of the SDSS training data, $r = 22$, defines the lower limit of the extrapolation subset. The morphological classifier failure and sources beyond survey depth ($r > 25$) provide the unsafe subset (fig. 2.4). The extrapolation subset is complete up to $r < 23.5$. The safe subset covers 21% of data, extrapolation 45%, and unsafe 34%.

random objects from KiDSxSDSS cross-match to visualize the spectroscopic classes, which are sparse in the whole KiDS data, and put emphasis on the much fainter inference data. The plots show the main groups of spectroscopic classes and their placement over the whole feature space. The safe subset at $r < 22$ matches the part of the feature space covered by the training data, confirming that a single cut on the r magnitude assigns proper limits to the other magnitudes, colors, and ratios; the results are the same as in our simpler approach for KiDS DR3, where we matched only the *ugri* magnitudes, colors, and ratios. The main star and QSO groups are separated in the training data, but not in the whole KiDS data. The first part of the extrapolation subset at $r \in (22, 23)$ is located close to the training data, and it may provide reliable estimations. The rest of the extrapolation set covers fainter and more complicated parts of the feature space, such as the joining space between QSO and star groups at $23 < r < 24$, thus such objects have a lower chance of their classification predictions being correct.

We used the 2D visualization to investigate estimation performance of the ML models. The models work with highly dimensional data, which makes it difficult to visualize the decision boundaries. We did not investigate the color-color plots due to the large number of possible combinations and the required domain knowledge of how to interpret them. Instead, the manifold learning, such as t-SNE, visualizes nonlinear data structures and this allows us to understand the models as well as, or better than, it would be possible with the color-color plots. Additionally, we used the embedding to have insight into the extrapolation part of the feature space, which cannot be tested with methods based on ground-truth data.

Figure 2.6 shows r magnitude distributions for the inference subsets. The safe subset was cut at $r = 22$, while the extrapolation and unsafe subsets overlap in magnitudes. We can see that the extrapolation subset is complete to $r < 23.5$, which puts a completeness limit on our catalog. We expect that the number counts of QSOs identified using the currently available training sets would become incomplete at $r > 23.5$.

TABLE 2.2: Train and test subsets of the KiDS DR4 x SDSS data. We selected the faintest 10% of the data as the faint extrapolation test. This splits the training data at $r = 21.3$. We used the same amount of objects at $r < 21.3$ as in the faint-end for the random test.

		size	quasar	galaxy	star
train	r < 21.3	105k	11k (11%)	71k (68%)	23k (22%)
test random	r < 21.3	13k	1.5k (12%)	8.8k (67%)	2.8k (21%)
test faint	21.3 < r < 22	13k	3.3k (25%)	7.2k (55%)	2.6k (20%)

2.4.2 Validation procedure

Proper design of the testing methods is one of the main goals of this work so as to make sure we did not overfit the models. Validation data have to differ from the training set to ensure proper model generalization. A randomly chosen sample of data which densely covers the feature space might not fully show the overfitting effects, and this might have a very negative influence on the inference at the faint data end, both for classification and photo-zs. We used additional spectroscopic surveys to introduce some differences from the training data and tested the final predictions (Section 4.5.4). During the experiments, we used internal data characteristics to differentiate training from validation. The approach is similar to time series processing, where validation data should consist of dates later than the training ones. Similarly, we chose the faintest objects to test the regularization of the models. Another option would be to use highest-redshift objects, chosen separately for each class as they reside at different ranges of redshift, which would test the prediction of values not seen during training. However, radial velocities of stars measured by SDSS obviously do not correlate with photometry and we would not observe any variation in star colors between the training and validation data. As magnitude correlates with redshift in the case of QSOs and galaxies, we expect the faint test to evaluate the extrapolation accuracy of ML models with respect to the estimated redshift values. Figure 2.7 explains the whole methodology in blocks illustrating experiments as well as inference and catalog testing.

Table 2.2 summarizes the training and validation sets. We selected the faintest 10% of the training data as a faint extrapolation test, and the same amount of random objects from the rest of the training data as a random test. Both tests allowed us to correctly tune the models for a bias-variance trade-off and check how the estimations deteriorate when we extrapolate to fainter magnitudes. The faint extrapolation test has a higher contribution of QSOs, which adds to differences between the training and validation. The faint extrapolation test sample in the spectroscopic data, at $21.3 < r < 22$, should not be confused with the faint extrapolation inference data at $r > 22$.

We tested QSO redshifts on two subsets: the true spectroscopic QSOs from SDSS and QSO candidates from the output of an ML model. The QSO candidates may contain true stars and galaxies due to misclassification. As we are solving two distinct tasks, classification to identify QSOs and regression to estimate their redshifts, a test of QSO candidates evaluates the consistency between classification and redshift models and it requires both class and redshift to be assigned correctly. This test informs us about the robustness of the final catalog, and we consider redshift errors obtained in the set of QSO candidates as the most important metric for model selection.

We used the following classification metrics[7] (scikit-learn, Pedregosa et al., 2011): accuracy for the three-class classification problem (QSO, galaxy, and star) as well as purity and completeness for QSO detection. For redshifts, we used:

[7] https://scikit-learn.org/stable/modules/model_evaluation.html

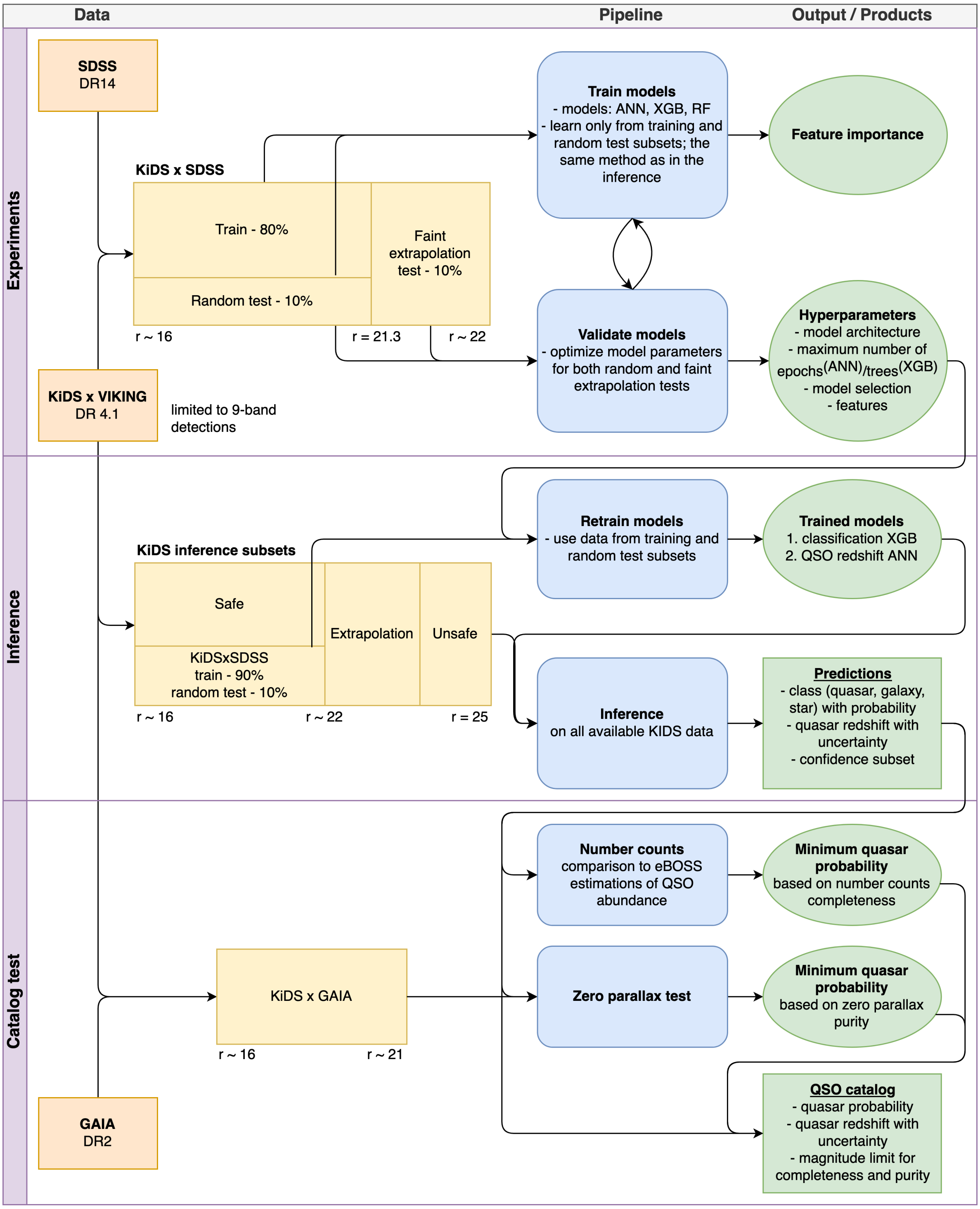
Data
Pipeline
Output / Products
Experiments
SDSS DR14
KiDS x SDSS
Train - 80%
Random test - 10%
Faint extrapolation test - 10%
r ~ 16
r = 21.3
r ~ 22
KiDS x VIKING DR 4.1
limited to 9-band detections
Train models
- models: ANN, XGB, RF
- learn only from training and random test subsets; the same method as in the inference
Feature importance
Validate models
- optimize model parameters for both random and faint extrapolation tests
Hyperparameters
- model architecture
- maximum number of epochs(ANN)/trees(XGB)
- model selection
- features
Inference
KiDS inference subsets
Safe
KiDSxSDSS train - 90% random test - 10%
Extrapolation
Unsafe
r ~ 16
r ~ 22
r = 25
Retrain models
- use data from training and random test subsets
Trained models
1. classification XGB
2. QSO redshift ANN
Inference
on all available KIDS data
Predictions
- class (quasar, galaxy, star) with probability
- quasar redshift with uncertainty
- confidence subset
Catalog test
Number counts
comparison to eBOSS estimations of QSO abundance
Minimum quasar probability
based on number counts completeness
KiDS x GAIA
r ~ 16
r ~ 21
Zero parallax test
Minimum quasar probability
based on zero parallax purity
GAIA DR2
QSO catalog
- quasar probability
- quasar redshift with uncertainty
- magnitude limit for completeness and purity

FIGURE 2.7: KiDS DR4 methodology diagram. The procedure consists of three main parts: experiments as well as inference and catalog tests. The experiments are based on the cross-match between KiDS and SDSS data, and they include the repeatable process of training and evaluating ML models. The training is based only on the train and random test subsets, while the hyper-parameter tuning uses both random and faint extrapolation tests. The best hyper-parameters found are used in the inference to train new models, now on the whole range of magnitudes available in the training data. The raw predictions were then tested with number counts and Gaia parallaxes to calibrate the final catalog with probability cuts for the optimal purity-completeness trade-off.

- the mean squared error

$$MSE = \frac{1}{N}\Sigma(z_{\mathrm{spec,i}} - z_{\mathrm{photo,i}})^2, \tag{2.1}$$

- R-squared

$$R^2 = 1 - \frac{SS_{\mathrm{RES}}}{SS_{\mathrm{TOT}}} = 1 - \frac{(z_{\mathrm{spec,i}} - z_{\mathrm{photo,i}})^2}{(z_{\mathrm{spec,i}} - \bar{z}_{\mathrm{spec}})^2}, \text{ and} \tag{2.2}$$

- the redshift error

$$\delta z = \frac{z_{\mathrm{photo}} - z_{\mathrm{spec}}}{1 + z_{\mathrm{spec}}}, \tag{2.3}$$

where SS_{RES} is the residual sum of squares, SS_{TOT} is the total sum of squares, z_{spec} is the true spectroscopic redshift, z_{photo} is the predicted photometric redshift, and $\bar{z}_{\mathrm{spec}}$ is the mean spectroscopic redshift of a given validation sample.

We performed 100 bootstrap samplings on random and faint extrapolation tests to make sure that the mean standard errors ($\sigma/\sqrt{100}$) are at about 3-4 decimal places depending on the metric. This gives statistical relevance to the precision with which we report the results.

Due to differences between the training and inference data, we used several methods to test the final catalog: number counts, spatial densities, GAIA parallaxes, and comparison with external quasar catalogs. This way we ensure that any decision on model parameters or feature engineering does not lead to issues in the final inference. Those testing methods allowed us to calibrate the purity versus completeness ratio of the final quasar catalog by setting the minimum classification probability. With calibrated classification, the photometric redshifts might be a good approximation of the real redshift distribution, but further calibrations are possible. We test the photometric redshift distribution for quasar bias constraints in the section 5.2.

2.5 Correlation analysis

The resulting catalogs of quasars are characterized by high surface density, $\sim 157\mathrm{deg}^{-2}$ at $r < 22$ and $\sim 309\mathrm{deg}^{-2}$ at $r < 23.5$ in the KiDS DR4 (see tab. 4.3), in comparison to, for instance, $\sim 80\mathrm{deg}^{-2}$ in the SDSS DR16 (Lyke et al., 2020). It provides an opportunity to study large scale structure at redshifts higher than 2, and our goals here are to perform an early study in order to validate the catalog through its correlation analysis and estimate the quasar bias function.

We use the publicly available CMB lensing convergence map provided by the Planck Collaboration et al., 2020b. The data are provided in the form of spherical harmonic coefficients κ_{lm}, which we transform into a HEALPix map (Górski et al., 2005) with the resolution parameter Nside=512, the same as for the galaxy overdensity map.

2.5.1 Theory

The structure of this section is based on Alonso et al., 2021. We define and calculate the map of sky-projected quasar overdensity as

$$\delta_q(\hat{n}) = \frac{N_q(\hat{n}) - \bar{N}_q}{\bar{N}_q}, \tag{2.4}$$

where $\hat{n}$ is position on the sky, N_q is the number of quasars, and $\bar{N}_q$ is the mean number of quasars. It relates to the three dimensional overdensity field Δ_g through

$$\delta_q(\hat{n}) = \int dz \frac{dp}{dz} \Delta_q(\chi(z)\hat{n}, z), \tag{2.5}$$

where χ is the comoving radial distance, and $\frac{dp}{dz}$ is the redshift distribution of the galaxy sample normalized to 1 when integrated over the z.

The lensing convergence $\kappa(\hat{n})$ quantifies the distortion in the trajectories of the CMB photons caused by the gravitational potential of the intervening matter structures, and is proportional to the divergence of the deflection in the photon arrival angle $\alpha : \kappa = -\nabla \cdot \alpha/2$. As such, κ is an unbiased tracer of the matter density fluctuations $\Delta_m(x,z)$, and is related to them through:

$$\kappa(\hat{n}) = \int_0^{\chi_{LSS}} d\chi \frac{3H_0^2\Omega_m}{2a} \chi \frac{\chi_{LSS} - \chi}{\chi_{LSS}} \Delta_m(\chi(\hat{n}), z(\chi)), \tag{2.6}$$

where Ω_m is the fractional matter density, $a = 1/(1+z)$ is the scale factor, H_0 is the Hubble constant, and χ_{LSS} is the comoving distance to the surface of last scattering.

Any three dimensional field (U) can be projected onto a sphere using a kernel W_u

$$u(\hat{n}) = \int d\chi W_u(\chi) U(\chi\hat{n}, z(\chi)). \tag{2.7}$$

Any such projected quantity can be decomposed in terms of its spherical harmonic coefficients $u_{\ell m}$, the covariance of which is the so-called angular power spectrum (C_ℓ^{uv}), which is a Fourier transform of a correlation function. The angular power spectrum can be related to the power spectrum of the 3D fields P_{UV} through

$$C_\ell^{uv} = \int \frac{d\chi}{\chi^2} W_u(\chi) W_v(\chi) P_{UV}(k = \frac{\ell + 1/2}{\chi}, z(\chi)), \tag{2.8}$$

where $P_{UV}(k,z)$ is the variance of the Fourier coefficients of U and V. In this formalism, for the two fields under consideration, the radial kernels are given by

$$W_q(\chi) = \frac{H(z)}{c} \frac{dp}{dz}, \tag{2.9}$$

$$W_\kappa(\chi) = f_\ell \frac{3H_0^2\Omega_m}{2a} \chi \frac{\chi_{LSS} - \chi}{\chi_{LSS}} \Theta(\chi_{LSS} - \chi), \tag{2.10}$$

where $H(z)$ is the expansion rate, $\Theta(x)$ is the Heaviside function, and f_ℓ is the scale-dependent prefactor given by

$$f_\ell = \frac{\ell(\ell+1)}{(\ell+1/2)^2} \simeq 1, \tag{2.11}$$

which is relevant only for $\ell \lesssim 10$, and accounts for the fact that κ is related to Δ_m through the angular Laplacian of the gravitational potential Φ.

Eq. 2.8 is only valid only in the Limber approximation, i.e. for the case of small angular scales and spatial distances (Limber, 1954), where the spherical Bessel functions can be approximated by

$$j_\ell(x) \simeq \sqrt{\frac{\pi}{2\ell+1}} \delta(x - \ell - 1/2). \tag{2.12}$$

This is accurate when the radial kernels W_u are broader than the typical correlation length of the density inhomogeneities, as is the case for both δ_q and κ.

In order to analyse the matter distribution using quasars as its tracers, we use bias function which relates the auto- and cross-correlation between the matter and quasar overdensity, $P_{gg}(k,z)$, $P_{gm}(k,z)$, $P_{mm}(k,z)$. We assume a scale independent bias function $b_q(z)$, such that:

$$P_{gm}(k,z) = b_q(z)P_{mm}(k,z), \tag{2.13}$$

$$P_{gg}(k,z) = b_q^2(z)P_{mm}(k,z). \tag{2.14}$$

We use the range of the multipoles ℓ smaller than 550, where the non-Gaussian contributions to the covariance matrix can be neglected (e.g. Alonso et al., 2021). Finally, we subtract the shot noise from the auto-correlation, due to discrete nature of quasar number counts, calculated as $\frac{4\pi f_{sky}}{N}$ where f_{sky} is the fraction of the observed sky and N stands for the total number of objects (e.g. Alonso et al., 2019).

To compare the observed and theoretical power spectra, we use a Gaussian likelihood given by

$$\chi^2 = -2\log p(d|q) = (d - t(q))^T Cov^{-1}(d - t(q)), \tag{2.15}$$

where d denotes all measured power spectra, and $t(q)$ are the theoretical predictions for a set of parameters q. We report significance of the $C_{q\kappa}$ detection as a square root of the difference in χ^2 between a null hypothesis and the best fit model ($\sqrt{\Delta\chi^2}$).

2.5.2 Quasar application

The key problems we are interested in are as follows:

1. Measuring the correlations C_{qq} and $C_{q\kappa}$ under assumptions of the perfect purity and completeness of the catalog
2. Testing if additional quasars from the extrapolation range of the ML models ($r > 22$) allow to obtain higher significance of the correlation functions
3. Putting constraints on the QSO bias function

In order to test the extrapolation of the catalog to magnitudes fainter than the SDSS limit ($r > 22$), we test the catalog limited to r lower than 21, 22, 23, and 23.5 (as suggested by the tests in §4.5). Additionally, we test different quasar selection probabilities, which control the misclassification rate with stars and change the shape of the auto-correlation power spectrum.

In order to constrain the bias function, we fit theoretical power spectra to C_{qq} and $C_{q\kappa}$ based on the redshift distribution resulting from the photo-z estimates, while using two degrees of freedom for the bias model, namely $b_q(z) = A(1+z)^2 + B$, as suggested in the literature (e.g. Sherwin et al., 2012).

We use the pseudo-Cl estimator (Peebles, 1973; Hivon et al., 2002) implemented in the NaMaster library (Alonso et al., 2019). For the theoretical computations, we use the Core Cosmology Library (CCL, Chisari et al., 2019). We model the matter power spectrum using the HALOFIT parameterization (Takahashi et al., 2012) as implemented in the CAMB Boltzmann code (Lewis, Challinor, and Lasenby, 2000). We run the Monte Carlo Markov Chains (MCMC) implemented in the emcee library (Foreman-Mackey et al., 2013). We assume a flat ΛCDM cosmology based on the Planck Collaboration et al., 2020a with $H_0 = 67.4\mathrm{km/s/Mpc}$ and $\Omega_m = 0.315$.

3

Quasar catalog in the optical KiDS DR3

This chapter is based on the publication Nakoneczny et al., 2019, and all the results were obtained by the thesis author, unless stated otherwise. In this chapter, we show the results of classification and the properties of the quasar catalog from the KiDS DR3 described in §2.1.2, using methodology outlined in §2.2 and §2.3. The chapter is organized as follows. §3.1 shows the feature selection results, §3.2 describes the model selection results, §3.3 gives the classification results on the internal tests, and, finally, §3.4 evaluates the final catalog using external testing methods and provides its properties.

3.1 Feature selection

We use methodology from §2.2.2 to perform feature selection. We verify the usefulness of a large number of features from the KiDS DR3 database, such as *ugri* magnitudes, their differences (colors), their ratios, and also fluxes in all available apertures, observation errors, ellipticity, as well as star/galaxy separators. By applying the method of feature importance evaluation, we create the final feature set which provided the best model performance. It consists of 17 features in total: four *ugri* magnitudes, six resulting colors, six magnitude ratios, and `CLASS_STAR`. Figure 3.1 quantifies the feature importance in percentages. The stellarity index is a very useful feature, and its role is to provide a nearly perfect separation between galaxies and point-like objects like stars and quasars. Importances of colors and ratios for the same magnitude pairs are similar, which means that they are similarly useful in providing information to the model.

The results illustrated in Fig. 3.1 show that magnitude values are of much less importance for the classification than colors and magnitude ratios. Therefore, in addition to the fiducial approach where all the listed features were used, we have experimented with a classification setup without magnitudes. In such a case, the purity and completeness of the quasar classification measured on the test data were worse by $\sim 1.5\%$, which is mostly due to increased confusion with stars. We also tested the no-magnitude model by generating its predictions on the inference set and comparing the results with those where the whole feature set was used. Only 67% of quasar candidates identified by the model which includes magnitudes were also classified as QSOs in the “color-only” case, and almost all of the rest were classified as stars. Based on these findings, we conclude that the model in which magnitudes are not used performs worse, and in particular leads to a higher rate of misclassification with stars. Our default approach is therefore to use all the features shown in Fig. 3.1.

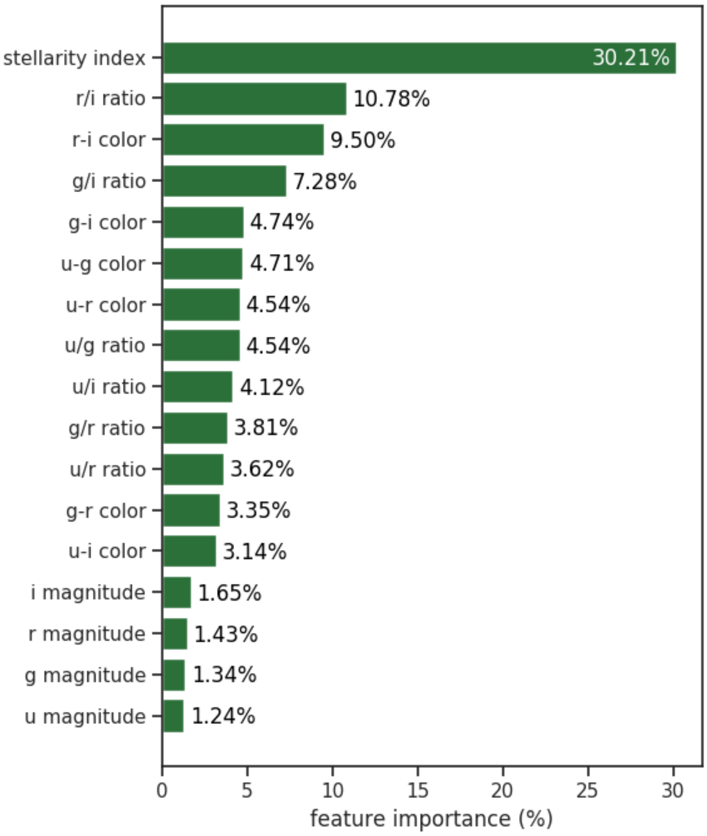

FIGURE 3.1: Features of the final model sorted according to their importance.

TABLE 3.1: A comparison of the test results for different models, achieved on the SDSS-based test set separated from the training and validation data. See Section 2.3.1 for details of the metrics.

	three-class: accuracy	QSO vs. rest: F1
RF	96.56%	88.67%
XGB	96.44%	88.12%
ANN	96.28%	87.63%

Figure 3.1 indicates also that among the four KiDS DR3 passbands, the *u* band is of least importance for our classification task. This is also the band which has the largest fraction of missing or excessively noisy observations, removed at the data preparation stage. One could therefore attempt classification based on only *gri* bands, which would give larger training and inference datasets than in our case (i.e., fewer sources would have been removed in the procedure described in Section 2.1.2). In the present application, this would however lead to significant limitation of the feature space, removing in total 7 of the 17 features.

3.2 Model selection

We use methodology from §2.2.1 to perform model selection. In the case of RF, the best results are usually achieved by building fully extended trees with leafs belonging only to one class, which also provided the best results for our work. We chose entropy as the function to measure the quality of a split, and 400 trees in the model as we did not observe any

TABLE 3.2: Evaluation metrics for KiDS DR3 calculated from the SDSS-based test set separate from the training and validation data.

classification type	metric	score
three-class	accuracy	96.6%
QSO vs. rest	accuracy	97.0%
	ROC AUC	98.5%
	purity	90.8%
	completeness	86.6%
	F1	88.7%

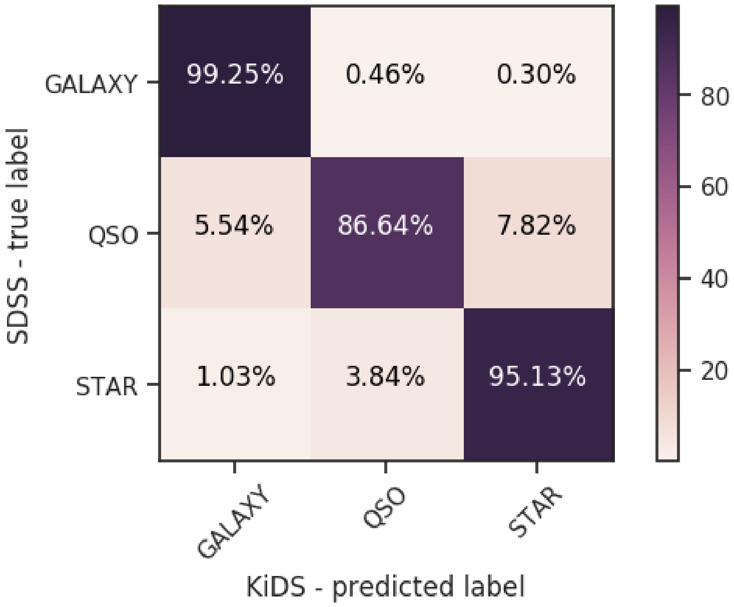

FIGURE 3.2: Normalized confusion matrix of the KiDS DR3 classification calculated for the SDSS test sample.

performance gain above this value. For XGB, we obtained good results when using 200 estimators of depth 7 and trained with a 0.1 learning rate, while artificial neural network was built with 2 hidden layers of 20 neurons each, using the rectified linear unit (ReLU) activation function. Table 3.1 shows a comparison between the model performances; for details of the model testing procedure and evaluation metrics see Section 2.3.1. We observed small differences between the performance of different models, with RF generally performing best. Such model hierarchy and small differences in scores are expected for this kind of a dataset with a rather low number of features and classes to predict.

We decided to choose random forest as the final classifier. This decision was based not only on the model performance, but also because RF does not require time-consuming selection of the best training parameters. It also provides a measure of feature importance, offering a relatively fast and straightforward way of choosing the most appropriate features.

3.3 Classification results

The classification results are tested as described in Section 2.3.1. As far as the evaluation metrics are concerned, we measure the accuracy for the three- and two-class (QSO vs. rest) problems, while the ROC AUC, precision, completeness and F1 are provided only for the

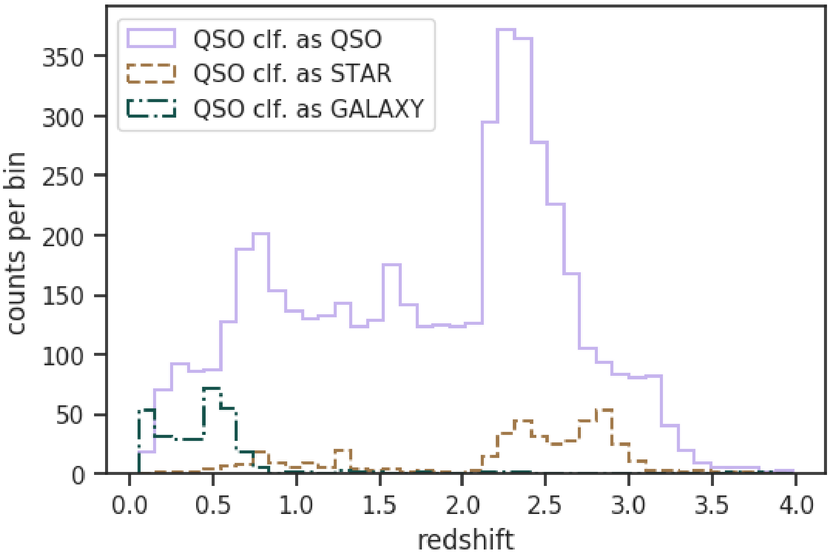

FIGURE 3.3: Redshift distribution of SDSS quasars present also in our KiDS DR3 inference sample, separated into the classes predicted by our model. The solid purple line shows correctly classified QSOs, while the orange dashed and green dot-dashed are for true quasars misclassified as stars and galaxies respectively.

binary case. All the scores are listed in Table 3.2. Accuracy of the algorithm is very similar in both two- and three-class cases and amounts to almost 97%. The ROC AUC provides a very high value of ~ 99%. Purity of the final catalog is estimated to be ~ 91%, as the quasar test sample is contaminated with ~ 7% stars and ~ 2% galaxies. The completeness is a little bit lower, ~ 87%. The $F1$ measure gives then ~ 89%.

Figure 3.2 visualizes the classification results in the form of a normalized confusion matrix (CM), from which more information can be extracted. In the case of a normalized CM, cell values are given as percentages which sum up to 100% in each row. This gives completeness values on the diagonal for each of the classes. From the top row, it is clear that almost all the galaxies are classified correctly, and in the case of stars, a small fraction is misclassified as quasars, which reduces the purity of the QSO catalog. Quasars themselves are the most prone to misclassification, as about 13% of them are assigned either star or galaxy labels, which translates to the incompleteness of final QSO sample.

In order to better understand the reasons for misclassification, we examine the redshift ranges at which the quasars are assigned incorrect classes. Figure 3.3 compares redshift distributions of the true SDSS quasars which were classified as QSOs (solid purple), or misclassified as stars (dashed orange) or as galaxies (dash-dotted green). As expected, the QSO-galaxy mismatch happens predominantly at very low redshifts, where the QSO host galaxies can have large apparent size and flux. The flux-limited nature of the spectroscopic training quasars means that at low redshifts, intrinsically less luminous QSOs are included, and the flux of the host galaxy can be comparable or even dominant over that from the AGN. An additional possible explanation is that although we do not explicitly use redshifts in the classification, fluxes and colors of galaxies and quasars are correlated with redshift, so the classification model indirectly learns this relation. Therefore, as the training data are dominated at low redshift by galaxies, this can be problematic for the model.

The results described in this paragraph were obtained by Aleksandra Solarz, one of the co-authors of the Nakoneczny et al., 2019. For better insight into the galaxy/QSO mismatch, we have inspected spectra of 100 randomly chosen sources which were labeled as QSOs by

SDSS but classified as galaxies by our algorithm. These objects show signs of AGN emission in the spectrum (broadened lines and prominent high ionization lines), however also the D4000 break and the calcium doublet are visible, characteristic of older stellar populations in the host galaxy. As our classification scheme does not use spectra, the shape of the continuum plays a crucial role in the performance. For that reason, when the emission of the host galaxy is detectable, the SDSS AGNs are often treated by the algorithm as galaxies, regardless of clear presence of AGN signatures in the spectrum.

Regarding QSOs incorrectly assigned with a star label, this happens at specific redshift ranges, such as $2.2 < z < 3.0$, where it is difficult to distinguish quasars from stars with spectral types spanning from late A to early F using broad-band optical filters (Richards et al., 2002; Richards et al., 2009a). This is exactly the redshift range where we observe the most of the star-QSO misclassification. The problem in not present if we add the near-IR imaging present in the KiDS DR4, as described in the next chapter.

The above analysis concerned the completeness of final QSO sample as a function of redshift. At present we cannot examine a relation between purity and redshift, as we would have to know the redshifts assigned to quasar candidates, including those which in fact are stars or galaxies. As already mentioned, presently in KiDS, the quasars do not have robust photo-zs (see e.g., Fotopoulou and Paltani, 2018). We address this problem in our approach to KiDS DR4. However, for the redshift estimation to be robust, additional near-IR data will be needed.

3.4 Catalog properties

In this Section we present and discuss the final quasar selection results in KiDS DR3. All the sources from our inference dataset of 3.4 million KiDS objects were assigned probabilities of belonging to the three training classes (star, galaxy or quasar). By selecting quasars as those objects which have $p_{\mathrm{QSO}} > \max(p_{\mathrm{star}}, p_{\mathrm{gal}})$, we have obtained 192,527 QSO candidates. Here we discuss the properties of this catalog. This is done by first calculating statistical measurements on the test set extracted from the general training sample but not seen by the classification algorithm. The final dataset is also cross-matched with data from the Gaia survey to examine stellar contamination, and with several external quasar catalogs to probe other properties of our sample. As a test for completeness, we analyze number counts in the final QSO catalog.

3.4.1 Gaia parallaxes

We validate the purity of our KiDS DR3 QSO catalog by analyzing parallaxes and proper motions of the contained sources. For this we use the second data release of the Gaia survey (Gaia Collaboration et al., 2018a), which is currently mapping the entire sky, focusing on stars in the Milky Way, but detecting also extragalactic objects like quasars (Gaia Collaboration et al., 2016). In Gaia DR2, over 1.3 billion Gaia-detected sources in the magnitude range $3 < G < 21$ have measurements of parallaxes and proper motions, therefore, a cross-match with that dataset can be used to test statistically if our QSO candidates are indeed extragalactic. In particular, the QSO candidates are expected to have negligible parallaxes and proper motions in the absence of systematics.

As Gaia is significantly shallower than KiDS ($G < 21$ corresponds to $r \lesssim 20$), practically all of the sources from Gaia over the common sky area have a counterpart in KiDS. The reverse of course does not hold, especially since Gaia does not store measurements of extended sources, and in particular only 32% of our inference sample is also matched to Gaia within 1” radius. In addition, due to considerable measurement errors in source motions at the faint end of Gaia, the test presented here cannot provide an unambiguous star/quasar division for

TABLE 3.3: Mean values of parallax (ϖ), right ascension and declination proper motions ($\mu_{\alpha*}$ and μ_{δ}), all in milli-arcsecond units, as derived from the Gaia high precision sample (see text for details). First three sets of rows show results for the ground-truth SDSS and KiDS DR3 × SDSS training objects. Next, acceptable quasar offsets based on model testing results are presented, while the last two rows show values for the KiDS quasar catalog and its probability-limited subset.

		size	ϖ	$\mu_{\alpha*}$	μ_{δ}
QSO	SDSS	138k	-0.02	-0.02	-0.03
	train	2.1k	-0.01	-0.02	-0.01
star	SDSS	560k	0.71	-1.81	-6.37
	train	7.3k	0.57	-6.12	-6.01
galaxy	SDSS	3.8k	0.16	-0.70	-2.50
	train	78	0.29	-3.60	-2.93
acceptable	SDSS	-	0.04	-0.14	-0.50
	train	-	0.05	-0.50	-0.48
QSO	KiDS	7.1k	0.21	-0.27	-1.08
	$p > 0.8$	5.8k	0.09	0.14	-0.42

our full inference sample. Moreover, as discussed in detail by Lindegren et al., 2018, the measurements of motions in Gaia DR2 have some non-negligible systematics. In particular, even stationary quasars have appreciable scatter in their measured parallaxes and proper motions. A special procedure is therefore needed to validate the contents of our quasar catalog using Gaia, as described below.

In order to analyze the systematics in Gaia DR2 parallaxes and proper motions, Lindegren et al., 2018 used a sample of quasars, which define a celestial reference frame, known as Gaia-CRF2 (Gaia Collaboration et al., 2018b), nominally aligned with the extragalactic International Celestial Reference System and non-rotating with respect to a distant universe. This allowed them to design a set of criteria applied to Gaia measurements to make sure that the selected sources are indeed stationary. As a result, Lindegren et al., 2018 determined a global mean offset in Gaia parallaxes of −0.029 mas. Detecting appreciably higher offsets in the parallax distribution for sources assumed to be quasars would then point to stellar contamination.

A cross-match of our inference catalog with Gaia DR2 gave almost 1.1 million common objects. Among these, 1 million were classified by the model as stars, 40k as galaxies, and 38k as quasars. As our goal is to evaluate the stellar contamination of the quasar catalog, we cannot directly apply the criteria of Gaia data cleaning from Lindegren et al., 2018, as the aim there was to reduce this contamination in a QSO sample. Instead, we define a Gaia high precision sample, by taking sources with parallax and proper motion errors smaller than 1 mas, and compare the measurements for the KiDS sources to those obtained in the same way from the SDSS and KiDS×SDSS training sample (where for the SDSS case we used the full DR14 spectroscopic dataset cross-matched with Gaia DR2). This reduces the number of quasars found in both KiDS and Gaia from 38k to 7.1k.

In order to properly measure the purity of quasar candidates, we have to take into consideration the test results (Section 3.3), according to which our QSO candidate catalog consists of 91% quasars, 7% stars and 2% galaxies. Therefore, we calculate "acceptable" parallax and proper motion offsets by taking a weighted mean of the respective astrometric quantities with weights of 0.91, 0.07 and 0.02 for ground truth quasars, stars and galaxies. Table 3.3 shows parallax and proper motion mean values for ground truth SDSS and training objects, together with the acceptable offsets and values derived for the KiDS quasar candidates.

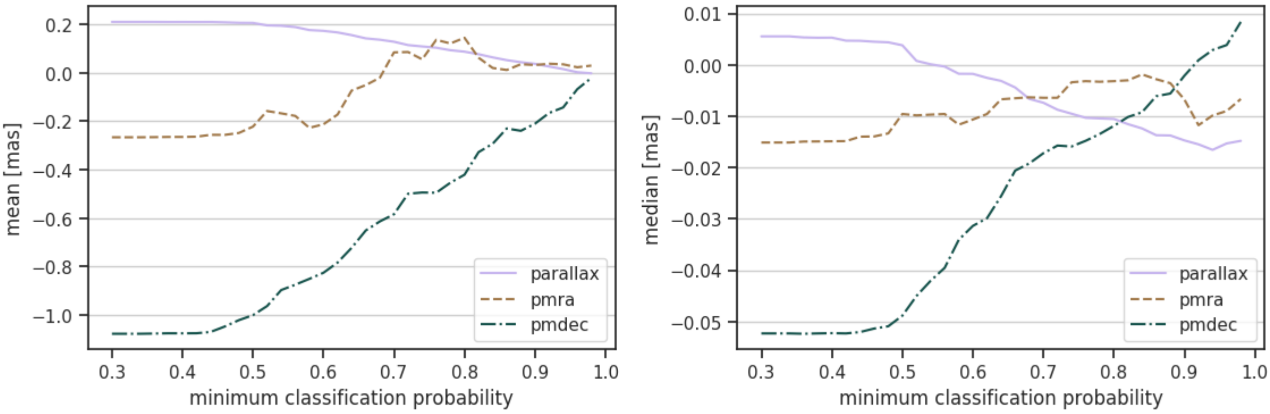

FIGURE 3.4: Mean (left panel) and median (right panel) of parallax (solid purple), right ascension (dashed orange) and declination (dash-dotted green) proper motions, derived from the Gaia high precision sample for the KiDS DR3 quasar candidates, as a function of minimum quasar probability limit.

The acceptable offset for parallax (ϖ) is about 0.05 mas for both the full SDSS and training QSO samples, and ~ -0.50 mas for the proper motion in declination (μ_δ). The right ascension proper motion ($\mu_{\alpha*}$) shows inconsistent results between the full SDSS QSO and training datasets. In addition, it varies much more than ϖ and μ_δ, and its mean even changes sign depending on the QSO threshold probability, as shown below in Fig. 3.4.

The full catalog of KiDS quasar candidates matched with Gaia shows mean offsets significantly higher than the acceptable levels, which must be an imprint of residual stellar and galaxy contamination. We note however that as we use unclipped means, a fraction of significant outliers can highly influence the means. Still, those measurements can be used to purify the catalog by limiting the quasars to higher probability values according to our classification model. This makes sense from an ML point of view, as our model was optimized for the training dataset whose properties may differ from the final inference sample. Moreover, we know that some quasars are not easily distinguishable from stars in the optical bands used here, and the two-classes may occasionally overlap in terms of their positions in the feature space (Fig. 2.3). Such objects may have lower classification probability as they are surrounded by sources from an opposite class. This fact can be used to reduce the problem of stellar contamination by simply applying a limit on quasar probability. As shown in Table 3.3, at $p_{\mathrm{QSO}} > 0.8$ we obtain an acceptable offset for the mean value of μ_δ, and close to acceptable for ϖ. The absolute value of $\mu_{\alpha*}$ is also at an acceptable level for this probability limit, and in fact its mean oscillates around 0 for $p_{\mathrm{QSO}} \gtrsim 0.7$ (Fig. 3.4).

Figure 3.4 shows how the mean and median values of parallax and proper motions change as we increase the QSO probability limit. Mean values converge to 0 mas, while median values at this level of precision are required to stay within the QSO mean offsets shown in the first row of Table 3.3. Both mean and median values of the astrometric measurements decrease (in terms of their absolute values) for $p_{\mathrm{QSO}} > 0.5$. For higher QSO probability levels of 0.7 – 0.8 they are sufficiently close to the acceptable offsets for mean measurements, or 0 mas in case of median values, that at these p_{QSO} the quasar candidates can be considered reliable. An exception is the parallax, whose median changes sign at $p_{\mathrm{QSO}} \sim 0.5$ and continues decreasing to almost -0.02 mas at $p_{\mathrm{QSO}} \sim 1$. This is however expected from the offset calculated by Lindegren et al., 2018 which equals -0.029 mas for parallax measurements.

We consider these results a considerable success of our model, especially since this analysis of the KiDS DR3 objects which are also present in Gaia focuses on the star/quasar separation, which is the most difficult task to solve. Moreover, Gaia measurements may be

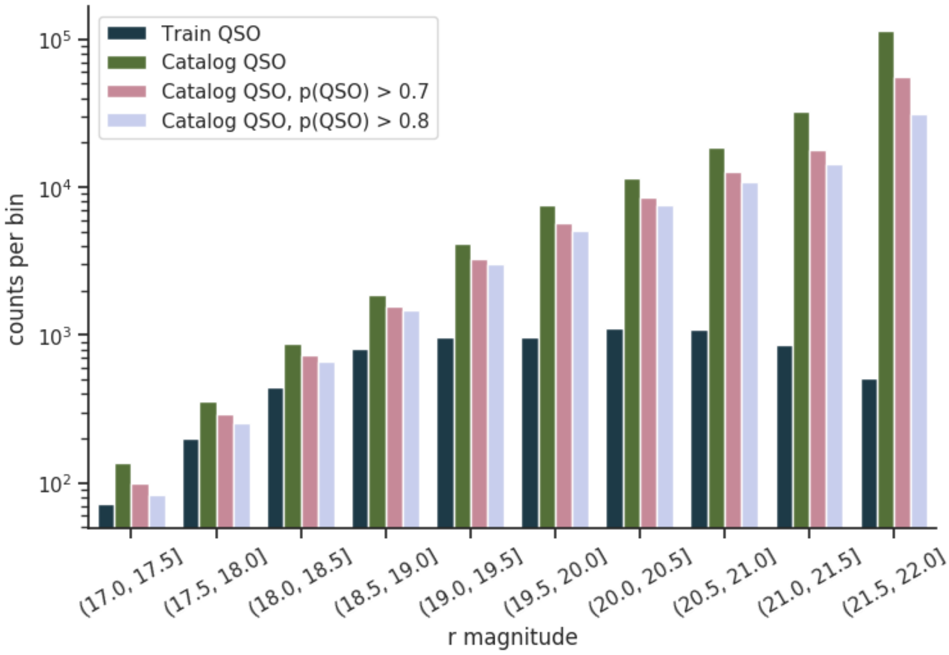

FIGURE 3.5: Number counts of SDSS quasars and KiDS quasars classified in this work. Together with the full QSO candidate sample, we also show samples limited to quasar probabilities above 0.7 and 0.8, which are the cuts we suggest applying to improve the purity of the sample.

strongly contaminated with large positive or negative values, resulting from an inconsistent matching of the observations to different physical sources. This may especially affect quasar measurements which require higher resolution than other objects, and therefore can show larger offsets in our catalog than in the training data. Based on this analysis, we suggest to limit the catalog to a minimum classification probability of $p > 0.8$ which favors purity and gives a sample of ~75k quasars, or use a cut of $p > 0.7$ which gives better completeness for a sample of ~100k quasars.

3.4.2 Number counts

As another test of completeness, we now compare the number counts of the quasars used for training and those in our final sample. This is done for the r band on which KiDS detections in the multiband catalog are based. The SDSS DR14 quasars used here as the training sample were preselected based on several color cuts and other criteria (e.g., Blanton et al., 2017), which results in various levels of incompleteness as a function of redshift and magnitude. On the other hand, a QSO sample selected from imaging data as the one resulting from our work is expected to provide a much more complete sample, ideally volume-limited. In Fig. 3.5 we compare $N(r)$ for the input SDSS spectroscopic quasars with results from our classification. In the latter case, we show number counts for the general QSO candidate sample and for the cut at $p_{\mathrm{QSO}} > 0.7$ and $p_{\mathrm{QSO}} > 0.8$, determined above as optimal for improving the purity of this dataset.

The counts shown give the total number of sources per bin for the full respective samples, that is not normalized by area. Therefore, the comparison has mostly qualitative character. It serves as a verification that the number counts in our photometrically-selected quasar sample steadily rise up to the limiting magnitude of the sample, also for the cases of minimum probability thresholds applied. In other words, the incompleteness visible for the SDSS

TABLE 3.4: Contributions of the classes predicted by our model, starting with the whole inference dataset and then moving to its cross-matches with external quasar catalogs: one spectroscopic (2QZ), providing ground truth, and 3 photometric, which are probabilistic. In those cross-matches, the highest quasar contribution is expected.

	size	star	quasar	galaxy
KiDS DR3 inference dataset	3.4M	35%	6%	59%
× 2QZ	5.4k	2%	97%	1%
× R09	17k	9%	86%	5%
× R15	18k	6%	91%	3%
× DP15	43k	15%	74%	11%

training sample at the faint end is not propagated to the final selection of the KiDS DR3 QSOs.

3.4.3 External quasar catalogs

Another method of examining the properties of our quasar catalog is by matching it with other QSO datasets overlapping with KiDS DR3. We use four external samples for this purpose: the spectroscopic 2dF QSO Redshift Survey (Croom et al., 2004, hereafter 2QZ), and three photometric samples (Richards et al., 2009a; Richards et al., 2015; DiPompeo et al., 2015, hereafter R09, R15 and DP15 respectively). 2QZ includes confirmed quasars, stars and galaxies, while the photometric catalogs are probabilistic, based on selection from SDSS (R09) and SDSS+WISE (R15 & DP15). In addition, only 2QZ significantly overlaps with the KiDS footprint, while the others (R09, R15 and DP15) cover the SDSS area[1]. They also have different depths, as illustrated in Fig. 3.6 which shows r-band distributions of cross-matches between the full KiDS DR3 and the four discussed catalogs. Among these, 2QZ is considerably shallower ($r < 21$) than our inference dataset, which is the main reason why we have not included it in our training set. As a result, the number of cross-matches between our inference catalog and the external datasets is not expected to be very large. Indeed, taking from the comparison catalogs sources which are labeled as quasars, we find respectively 5.4k objects of our inference sample in 2QZ, 17k in R09, 18k in R15 and 43k in DP15. Of these, respectively 5.2k (97%), 14.6k (86%), 16.4k (91%) and 31.8k (74%) have quasar labels in our catalog (see Table 3.4). A relatively lower consistency between our QSOs and DP15 might be related to the fact that in the DP15 some quasar candidates have probabilities as low as $p_{\rm QSO} > 0.2$. It should be stressed that the probabilistic character of the R09, R15 and DP15 datasets means that they can be only used for qualitative rather than quantitative comparisons. Unlike the spectroscopic 2QZ, these 3 photometric QSO datasets cannot be treated as ground truth and we will use them mainly to test the consistency between our model and the external approaches, and to further validate the minimum QSO probability at which our quasar catalog is robust.

The availability of three-class spectroscopic labels in 2QZ allows us to calculate the same metrics as for the SDSS test sample discussed in Section 3.3. The cross-match with KiDS reduces 2QZ to 7.8k objects which, in terms of spectroscopic 2QZ labels, consists of 5.4k QSOs, 2.4k stars and only 15 galaxies. We obtain high metric values in this case: three-class accuracy of 95%, QSO purity of 95% and completeness of 97%. This is summarized in Table 3.5, and Fig. 3.7 which shows the relevant confusion matrix. These results give an independent confirmation of the very good performance of our QSO classification also at the

[1] Another QSO sample that could be used is 2SLAQ (Croom et al., 2009) but it has much less overlap with KiDS DR3 than those considered here.

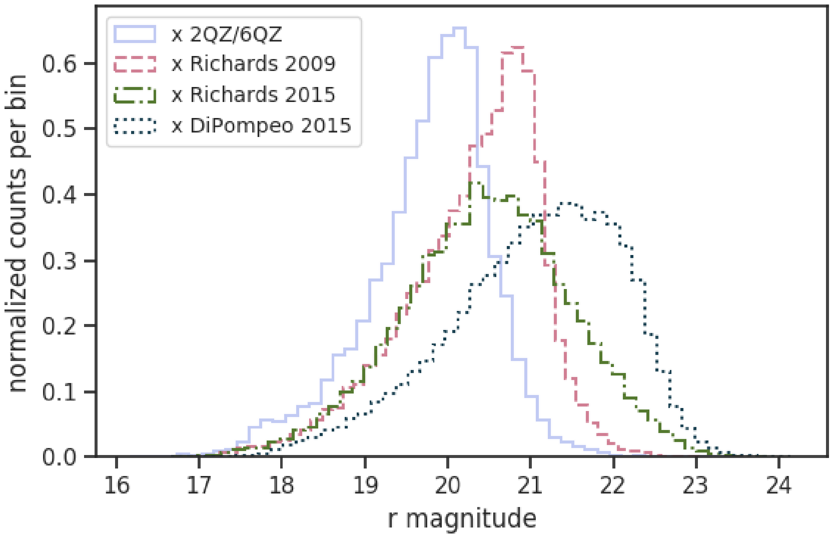

FIGURE 3.6: Normalized distributions of the r-band magnitude for cross-matches between the KiDS DR3 and four overlapping quasar catalogs, as indicated in the legend.

TABLE 3.5: Evaluation metrics for KiDS DR3 quasar classification, calculated from the 2QZ test set.

classification type	metric	score
three-class	accuracy	94.5%
QSO vs. rest	accuracy	94.9%
	ROC AUC	96.4%
	purity	95.3%
	completeness	97.4%
	F1	96.3%

bright end, here evaluated on a truly "blind" test set which was not part of the general training data. In particular, that comparison sample had been preselected from different input imaging and according to different criteria than SDSS, although we note that some 2QZ quasars are now included in the SDSS database.

As already mentioned, the remaining QSO catalogs used here for validation are probabilistic, therefore matching with them leads to more qualitative than quantitative evaluation. Still, we observe good consistency between our detections and those external models, especially for the R09 and R15 cases. Together with 2QZ, we use the overlapping QSO samples to further improve the purity of our quasar catalog. For this, we employ the probabilities delivered by the RF model, as was already discussed in Section 3.4.1. Except for the 2QZ case, this serves mostly to improve the consistency between our model and those external methods. Fig. 3.8 shows how the consistency between the models rises as we increase the minimum probability at which we accept the KiDS QSO classification. For 2QZ, we see excellent consistency for all the probability values and almost perfect (i.e., KiDS QSO contribution ∼ 100%) for $p_{\mathrm{QSO}} > 0.8$. For the probabilistic catalogs, we observe that many external quasars are classified by our method with $p_{\mathrm{QSO}} > 0.8$, which we deduce from the increase in consistency above this value. Those results fully agree with the conclusion from Section 3.4.1 which states that it is a good option to limit our identifications to $p_{\mathrm{QSO}} > 0.8$ when optimizing the purity of the quasar catalog.

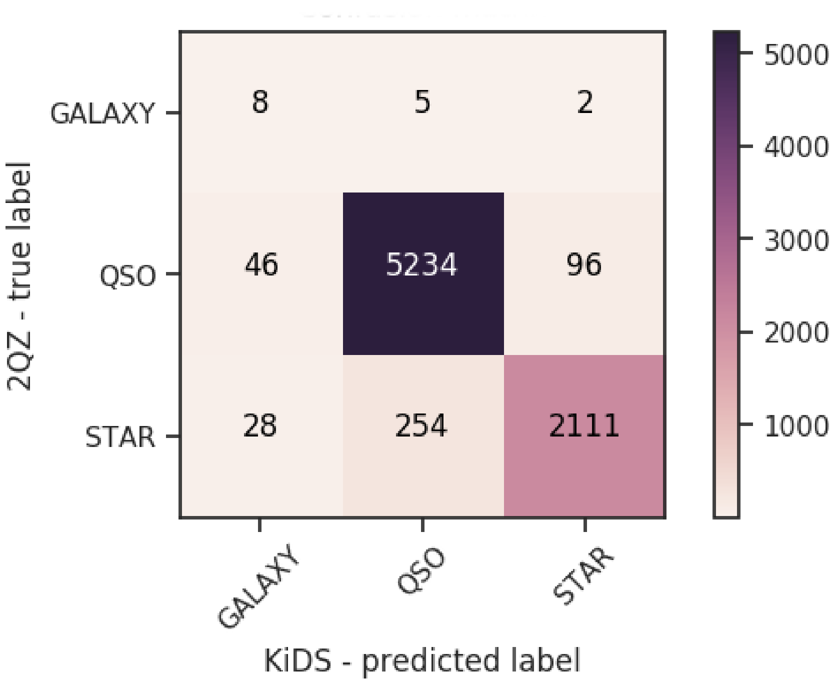

FIGURE 3.7: Confusion matrix of the KiDS DR3 classification calculated for the overlapping 2QZ sources.

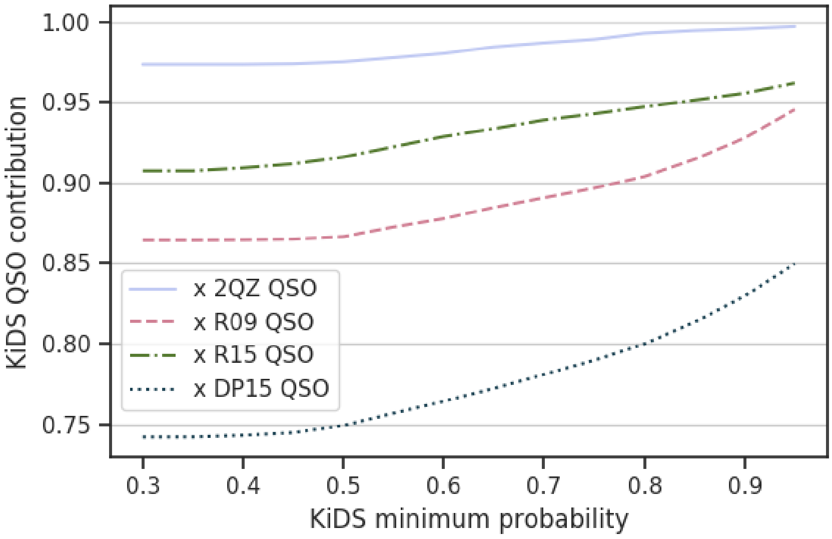

FIGURE 3.8: Proportion of KiDS DR3 QSOs in cross-matches with external quasar samples as a function of KiDS minimal classification probability. See text for details of the datasets.

3.4.4 WISE photometric data

We also validate our KiDS QSO catalog using mid-IR data from the full-sky Wide-field Infrared Survey Explorer (WISE, Wright et al., 2010). Despite being relatively shallow (∼ 17 mag (Vega) in the 3.4 μm channel), WISE is very efficient in detecting quasars at various redshifts. In particular, QSOs in WISE stand out as having very "red" mid-IR $W1 - W2$ ([3.4μm] – [4.6μm]) color (e.g., Wright et al., 2010; Jarrett et al., 2011; Jarrett et al., 2017). The general rule-of-thumb for QSO selection in WISE is $W1 - W2 > 0.8$ (Stern et al., 2012), but more refined criteria are needed to obtain pure and complete quasar samples from WISE (e.g., Assef et al., 2013; Assef et al., 2018). In particular, a non-negligible number of optically selected QSOs have $W1 - W2$ significantly lower than the 0.8 limit (e.g., Kurcz et al., 2016). That being said, QSOs are generally well separated from galaxies and stars in the $W1 - W2$ color with some minimal overlap for $W1 - W2 < 0.5$.

Although there exist QSO or AGN catalogs selected from WISE only (Secrest et al., 2015; Assef et al., 2018), here we use the entire AllWISE data release (Cutri and al., 2013) for the cross-match, as our goal is to derive the mid-IR $W1 - W2$ color of all the KiDS quasar candidates. We have cross-matched both our training set and the output catalog with AllWISE using a 2" matching radius (a compromise between KiDS sub-arcsecond resolution and the ∼ 6" PSF of WISE). We first note that ∼ 81% of our training set have counterparts in AllWISE, while for the inference sample this percentage is lower, ∼ 45%, mostly due to WISE being considerably shallower than KiDS in general. We also confirm the observation from Kurcz et al., 2016 that a large fraction of SDSS-selected quasars have $W1 - W2 < 0.8$ (∼ 22% in the cross-match of our training set quasars with WISE detections).

For the output catalog, the distribution of the $W1 - W2$ color for QSO candidates in the matched sample is in good agreement with that of the training set, with a slight preference to "bluer" colors which might actually reflect true properties of these optically-selected quasars rather than problems with our algorithm. Interestingly, this distribution shifts towards redder values of $W1 - W2$ when cuts on higher pQSO are applied. This is illustrated in Fig. 3.9, which shows that for $p_{\mathrm{QSO}} > 0.9$, the distribution of $W1 - W2$ for the KiDS QSO candidates is very similar to that of the SDSS spectroscopic quasars matched to WISE. This is remarkable given that nowhere in our classification procedure any mid-IR information was used, which additionally confirms the purity of KiDS quasar catalog.

3.5 Summary

In this Chapter, we present a catalog of quasars selected from broad-band photometric *ugri* data of the KiDS DR3. The QSOs are identified by the RF supervised machine learning model, trained on SDSS DR14 spectroscopic data. We first cleaned the input KiDS data of entries with excessively noisy, missing or otherwise problematic measurements. Applying a feature importance analysis, we then tune the algorithm and identify in the KiDS multi-band catalog the 17 most useful features for the classification, namely magnitudes, colors, magnitude ratios, and the stellarity index. We used the t-SNE algorithm to map the multi-dimensional photometric data onto 2D planes and compare the coverage of the training and inference sets. We limited the inference set to $r < 22$ to avoid extrapolation beyond the feature space covered by training, as the SDSS spectroscopic sample is considerably shallower than KiDS. This gives 3.4 million objects in the final inference sample, from which the random forest identified 190,000 quasar candidates. Accuracy of 97% (percentage of correctly classified objects), purity of 91% (percentage of true quasars within the objects classified as such), and completeness of 87% (detection ratio of all true quasars), as derived from a test set extracted from SDSS and not used in the training, are confirmed by comparison with external spectroscopic and photometric QSO catalogs overlapping with the KiDS footprint. The

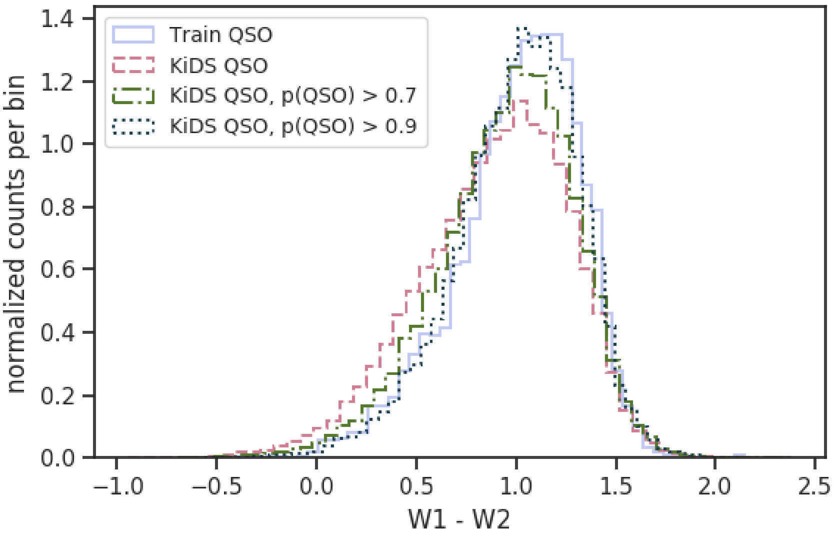

FIGURE 3.9: Distribution of the mid-infrared $W1 - W2$ color (3.4 μm - 4.6 μm, Vega) of quasar candidates in our KiDS sample, derived from a cross-match with all-sky WISE data. We show histograms for all KiDS QSOs matched with WISE, as well as for two examples of the probability cut: $p_{\mathrm{QSO}} > 0.7$, which is recommended to increase the purity of the sample, and $p_{\mathrm{QSO}} > 0.9$ to illustrate how the resulting $W1 - W2$ changes when the minimum probability considerably increases.

robustness of our results is strengthened by number counts of the quasar candidates in the r band, as well as by their mid-infrared colors available from the Wide-field Infrared Survey Explorer (WISE). An analysis of parallaxes and proper motions of our QSO candidates found also in Gaia DR2 suggests that a probability cut of $p_{\mathrm{QSO}} > 0.8$ is optimal for purity, whereas $p_{\mathrm{QSO}} > 0.7$ is preferable for better completeness. This study presents the first comprehensive quasar selection from deep high-quality KiDS data and will serve as the basis for versatile studies of the QSO population detected by this survey.

4

Quasar catalog in the optical and near-IR KiDS DR4

This chapter is based on the publication Nakoneczny et al., 2021, and all the results were obtained by the thesis author, unless stated otherwise. In this chapter, we show the results of classification, photometric redshift estimation, and the properties of the quasar catalog from the KiDS DR4 data described in §2.1.3, using methodology outlined in §2.2 and §2.4. The chapter is organized as follows. §4.1 compares the results of evaluation using the faint extrapolation test, §4.2 shows the feature selection results, §4.3 describes the model selection results, §4.4 gives the classification and photo-z results on the internal tests, and, finally, §4.5 evaluates the final catalog using external testing methods and provides its properties. The quality of the catalog allows for its applications in cosmology and quasar clustering analysis, which we perform in the next Chapter.

4.1 Training histories

We use methodology from §2.4.2 to compare different evaluation strategies. Figure 4.1 compares XGB training histories (number of trees used) for random and extrapolation tests. The random test is a good tracer of model quality for a broader range of magnitudes, and the faint extrapolation test is more sensitive to overfitting. During the model training, both testing methods should be taken into consideration. In the case of the classification, which achieves high accuracy, the faint extrapolation test can be given more importance. For redshifts, which are more difficult to fit at the faint data end, the extrapolation test might not show the full learning process, as illustrated by early minimums in QSO and galaxy redshift performance. When training the final inference models, we have to use the full magnitude ranges for training, so the extrapolation test is not available at that point, and we stop the model training based only on the results from the random test. Therefore, the best optimization approach during the experiments is to aim not only for the lowest error in a random test, but also for the lowest error in the extrapolation in the moment when the random error achieves its global minimum. This way, we can make sure that the final inference models, whose training is stopped based only on the random test, will also achieve good results at the faint data end.

4.2 Feature selection

We use methodology from §2.2.2 to perform feature selection. The final set of features consists of 83 values: optical *ugri* and near-IR $ZYJHK_s$ magnitudes, differences (colors) and

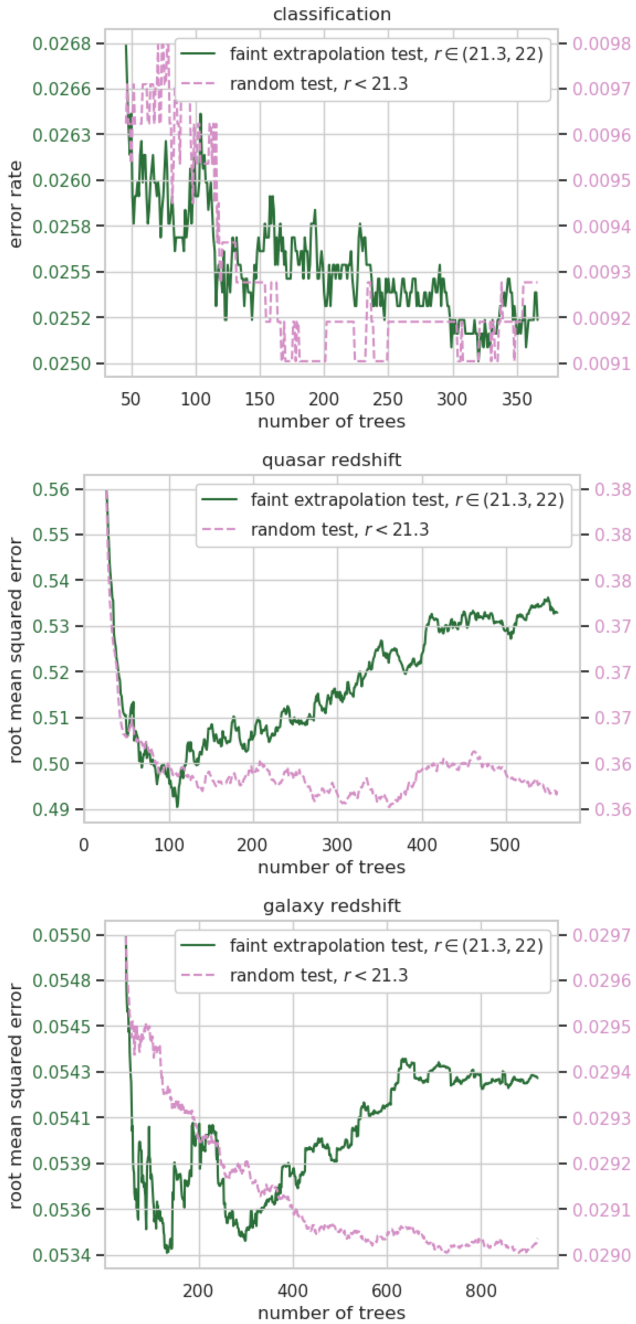

FIGURE 4.1: Learning histories for the XGB models. *Left:* Classification. *Center:* QSO redshift. *Right:* Galaxy redshift. The x-axis shows the number of trees created iteratively during the model training, and the y-axis shows the classification error rate and redshift root mean square error on two different scales for the random and faint extrapolation tests. We obtain the galaxy redshifts through the same pipeline as for the quasars. The errors in the faint test are higher than in the random tests due to extrapolation and higher noise. The models were stopped if the results on the faint test did not improve for 200 consecutive trees. For classification, which is easier to solve than redshift regression, the random test shows minimums sooner, followed by oscillations, while the faint test suggests longer training. For redshifts, which is a more complicated problem, the faint test achieves minimum quickly and then shows overfitting, while the random test suggests longer training.

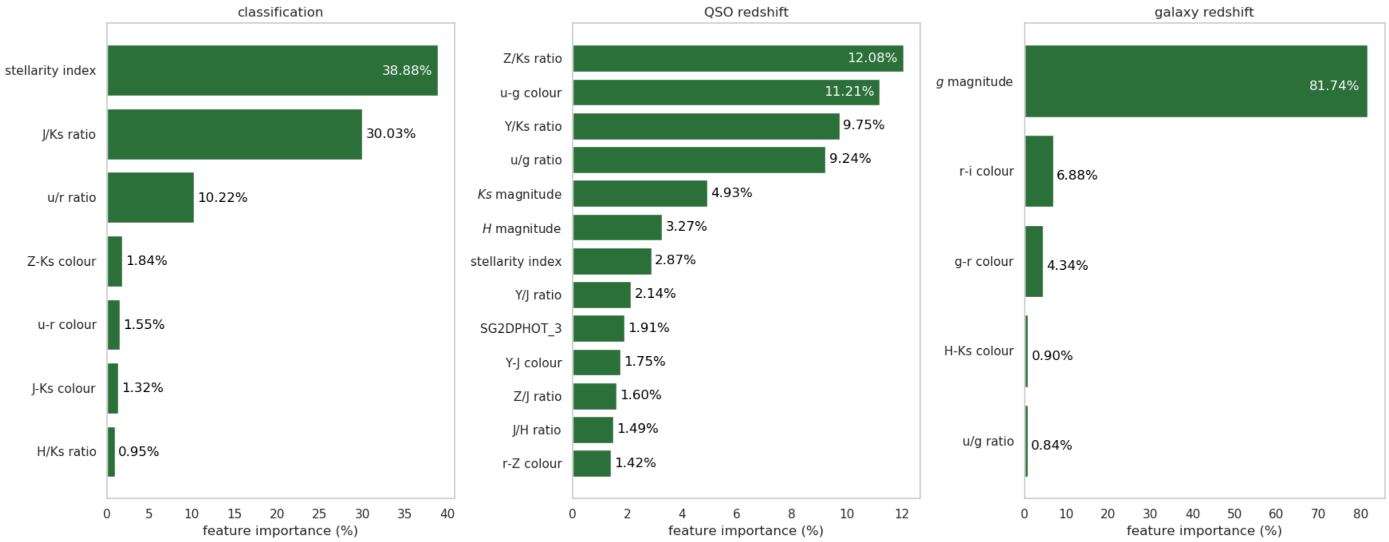

FIGURE 4.2: Feature rankings from the XGB models. *Left:* Classification. *Center:* QSO redshift. *Right:* galaxy redshift. We used the total gain across all splits in which the feature is used. The classification is mostly based on the stellarity index, near-IR JK_s, and optical ur bands. The QSO redshifts use all the NIR bands and most of the optical ones, but also the morphological parameters. The galaxy redshifts are based practically only on the optical gri magnitudes. Colors and ratios of the same magnitude pairs have a different importance.

ratios of every pair of magnitudes, and two morphological classifiers: the stellarity index from SExtractor and the third bit of SG2DPHOT from KiDS. We tested other bits of the SG2DPHOT without an observable improvement in the results. Ellipticity and other apertures were tested in the previous chapter and no significant increase in performance was seen.

Figure 4.2 shows the most important features for the classification and redshift estimation. We observe the importance of near-IR imaging, which is less affected by dust than the optical bands. The classification is mostly based on colors and magnitude ratios, but the redshift models also use the magnitude values, which is expected due to correlation between apparent magnitude and redshift. Quasar redshifts require more features than galaxy photo-zs, which confirms that they are more challenging to estimate. The most important magnitudes for QSO redshifts, the near-IR ZK_s, are the two extreme bands in this range. We observe only one feature of relatively low importance, which mixes the optical and near-IR, the $r-Z$ color. The stellarity indices were also used for QSO redshift, allowing models to distinguish extended low-redshift AGNs.

Feature importance suggests using ratios of magnitudes, however, the importance is based only on the training set and might fail in showing the effects of overfitting. Table 4.1 compares the full set of 83 features, with a limited set of 47 features excluding the magnitude ratios. We fine-tuned the models to the full set of 83 features as suggested by the feature importance, but we note that this approach might underestimate the performance of the no-ratio feature set. We observe that the differences between the two feature sets are significant for a faint extrapolation test of photometric redshifts. The ANN trained on the full set of features achieves the best results overall. Due to underestimated performance of the no-ratio feature sets, two scenarios are still possible: the magnitude ratios provide better results on both tests, or lead to overfitting in which case the random test results might be better, but the faint extrapolation results would be worse. It is very important that, to be able to properly

TABLE 4.1: Comparison of two feature sets: all 83 features and a limited set of 47 features excluding magnitude ratios, reported on the random ($r < 21.3$) and faint extrapolation ($r \in (21.3, 22)$) tests. The redshifts were tested in two subsets of QSOs: true spectroscopic ones and photometric candidates. The candidates include misclassified sources, e.g., a true star assigned a QSO class and redshift. We mark with bold font the best results, independently for random and faint extrapolation tests. The asterisk (*) marks the approach used to create the final catalog. Recall is the same as completeness, and MSE, R^2, and δz are given by equations 2.1, 2.2, and 2.3, respectively.

			classification			redshift for true QSOs			redshift for QSO candidates		
test	model	features	accuracy	purity	recall	MSE	R^2	δz	MSE	R^2	δz
random	RF	all features	99.01%	97.39%	94.43%	0.12	85%	0.02 ± 0.14	0.12	84%	0.03 ± **0.21**
		no ratios	98.98%	97.20%	94.57%	0.11	86%	0.02 ± 0.14	0.14	82%	0.04 ± 0.25
	XGB	all features	**99.09%**	**97.80%**	**95.82%**	0.13	84%	0.02 ± 0.15	0.13	83%	0.03 ± **0.21**
		no ratios	99.07%	**97.80%**	94.66%	0.12	84%	**0.01** ± 0.15	0.12	83%	0.03 ± 0.20
	ANN*	all features*	98.98%	96.90%	94.72%	**0.09**	**88%**	**0.01** ± **0.12**	**0.11**	85%	**0.02** ± 0.23
		no ratios	98.99%	97.58%	94.30%	**0.09**	**88%**	**0.01** ± **0.12**	**0.11**	**86%**	**0.02** ± **0.21**
faint extrap.	RF	all features	97.41%	96.07%	**92.30%**	0.31	31%	0.02 ± 0.25	0.33	31%	0.05 ± 0.38
		no ratios	97.39%	96.12%	92.22%	0.29	35%	0.01 ± 0.24	0.31	35%	0.04 ± 0.38
	XGB	all features	97.41%	96.44%	92.07%	0.27	39%	0.04 ± 0.23	0.34	29%	0.08 ± 0.41
		no ratios	**97.45%**	96.55%	91.94%	0.27	39%	0.04 ± 0.22	0.33	29%	0.08 ± 0.40
	ANN*	all features*	97.24%	96.54%	90.82%	**0.22**	**51%**	**0.00** ± **0.19**	**0.28**	**38%**	0.04 ± **0.37**
		no ratios	97.26%	**96.94%**	90.88%	0.24	46%	**0.00** ± 0.20	0.30	32%	**0.03** ± 0.38

assess the possibility of overfitting while using the ratios, the faint extrapolation test is necessary, as the random test already fails to show differences between the two feature sets. In this work, we decided to use the full set of 83 features, suggested by the feature importance. The approach we chose may not be optimal, as more experiments with feature and model engineering are possible. However, as the results we achieved are already very good, more extensive experiments are beyond the scope of this work.

Additionally, we experimented with reducing the feature set by removing, not whole groups of features (magnitudes, colors, or ratios), but single and least important features used for classification to minimize possible overfitting and increase model interpretability. This provided stable results for classification, but worsened the redshift estimates in the subset of QSO candidates due to lower consistency between the classification and redshift models. The inconsistency between the models results in more objects with either one of the classes or a redshift assigned incorrectly, while the redshifts of the QSO candidates require both the class and the redshift to be assigned correctly.

4.3 Model selection

We use methodology from §2.2.1 to perform model selection. Machine learning models can be modified in many ways which control the bias versus variance trade-off, in addition to the number of trees investigated in Fig. 4.1. In the case of ANNs, we tuned the number and size of layers, regularization, dropout, and learning rate. Some attempts at model optimization showed improvement in the results for both the tests, while the increased regularization usually led to better results only in the faint extrapolation case. For instance, once we reached the optimal network size for classification, using more layers or nodes per layer did not show any change in the random test, but led to deterioration in the faint extrapolation. Using only the randomly chosen subset may lead to a different set of parameters than when an extrapolation subset is also incorporated, and uncontrolled failure of estimation for the faint end. In case of incorrectly regularized models, such a failure can happen not only in extrapolation data, but also for the faintest magnitudes covered by the spectroscopic training data ($r \sim 22$

in our case). Thanks to both tests, we have the full picture of the bias versus variance trade-off and we can tune the models so that they perform well on both bright and faint data, and extrapolate to magnitudes fainter than available from the spectroscopy. We consider this an important success of our approach.

We tested several ML strategies, and we conclude that two ANNs, one for classification and one for QSO redshifts, provide the best results overall[1]. We find that a neural network model with multiple outputs for classification and redshifts, which would allow us to solve both problems at once, can be tuned to provide some improvement either for detection or redshift over two separate networks, but we did not manage to tune the network to simultaneously achieve the best results for both problems. It is due to both problems requiring different parameters. The specialized redshift models, trained either on galaxies or QSOs, are necessary for the best results, due to the differences between the two classes in the optimal model parameters, such as ANN size or regularization.

Table 4.1 shows the results of the specialized redshift models. The redshift metrics (Section 2.4.2) were calculated on two subsets of QSOs: true spectroscopic ones and our QSO candidates from photometric classification, as explained in Section 2.4.2. In our previous approach to KIDS DR3, which dealt with classification only, we did not observe a significant difference between RF and XGB performance. In this work, we find distinct results between all the tested models, due to a more complex validation method and larger feature space, now extended by near-IR bands. In the random test, XGB performs best in classification, and ANN performs best in redshifts. The faint extrapolation test shows less agreement on which model is the best for classification, but the superiority of ANN for redshifts is more prominent. We find that XGBoost is the most robust and straightforward model for classification, while ANN is the best for a combined classification and redshift.

A mixed approach, where classification is performed with XGB and redshifts with ANN gives the best results on the random test, but worse results for QSO candidates in the faint test, due to different characteristics of both models resulting in fewer objects with both class and redshift assigned correctly. In the case of the faint extrapolation test and the subset of QSO candidates, the R^2 deteriorates by 3 percentage points, while the standard deviation of δz is higher by 0.03.

Artificial neural networks provide good extrapolation results for both classification and redshifts. The classification deteriorates by 3 percentage points in the faint extrapolation test, while the standard deviation of δz is higher by 0.07 than in the random test.

4.4 Classification and photo-z results

We use methodology from §2.4.2 to evaluate the performance of classification and photo-z estimation. Quasar misclassification occurs mostly at low redshift (Fig. 4.3), with AGNs which have extended hosts and are generally labeled as QSO by SDSS. This affects the completeness more than purity, as in broad-band optical and NIR photometry those AGNs are more similar to galaxies than to quasars. It is due to the spectra taken through fibers in the SDSS, and in case of galaxies with AGN, the fiber is centered on the nucleus. This allows resolved galaxies to be matched with a QSO template by SDSS, and be spectroscopically classified as quasars. The KiDS photometry, however, picks up the host galaxy light and does not allow one to see the emission lines, therefore such AGNs are classified as galaxies from imaging. We define quasars as all the objects labeled as QSO by the SDSS, and this misclassification is a consequence of a mismatch in the QSO definitions between the spectroscopic and imaging surveys. Quasars at low redshifts with a low value of the stellarity

[1] The final model parameters and ANN architecture can be found in the script *models.py* in the github repository `https://github.com/snakoneczny/kids-quasars`.

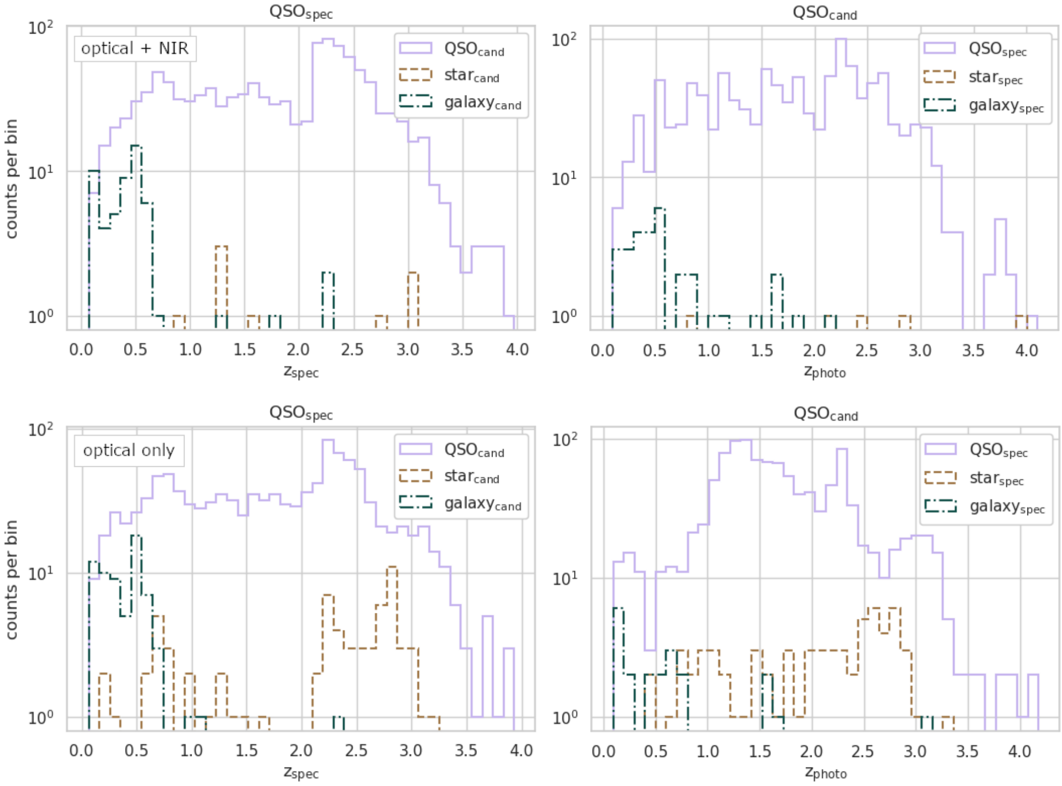

FIGURE 4.3: QSO misclassification as a function of redshift. *Top:* Using optical KiDS and near-IR VIKING features. *Bottom:* Using only optical KiDS features. *Left:* Spectroscopic QSOs and redshifts – a test for completeness. *Right:* QSO candidates and redshifts – a test for purity.

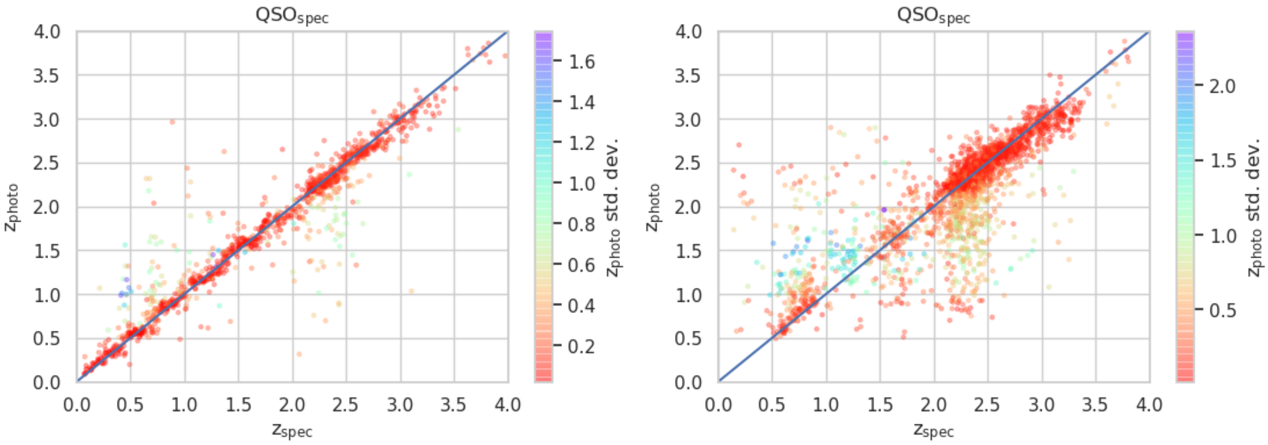

FIGURE 4.4: Comparison of the spectroscopic and photometric redshifts for SDSS test-set quasars. *Left:* Random test ($r < 21.3$). *Right:* Faint extrapolation ($21.3 < r < 22$). The mean photo-z error for the random and faint test equals 0.009 ± 0.12 and -0.0004 ± 0.19, respectively. Every redshift estimate is a Gaussian probability density function, the standard deviation of which represents the uncertainty (color coded).

index may additionally look more similar to extended galaxies for ML models. The quasar candidates consist of 96.9% true quasars, 2.6% galaxies, and 0.4% stars. The bottom plots of Fig. 4.3 show results obtained using only the optical *ugri* broad-bands (analogous to the Fig. 3.3). We observe misclassification with stars at the QSO redshift of $2 < z < 3$ (bottom left), and worse redshift estimates (bottom right), when only KiDS optical imaging is used, as studied in the previous approach.

Figure 4.4 compares spectroscopic and photometric redshifts on the random and faint tests. The random test shows a well-fitted distribution and thus the modeled uncertainty increases for objects further from the diagonal. We observe some clustering of redshifts around several values in the random test, but we did not manage to establish whether it is due to the ML model or internal data characteristics. The outliers behave similarly also in spectroscopic measurements due to confusion between pairs of emission lines (e.g., Croom et al., 2009, Fig. 10). The faint extrapolation test shows more scatter and more outliers. The aleatoric uncertainty, which we model with a Gaussian output layer, is related to the fact that objects which appear similar in photometry may have different redshifts. This model does not include the situation in which part of the feature space is not covered by data, and we would expect higher uncertainty for such estimations – this case would relate to epistemic uncertainties. After several iterations of tuning the model with random and faint extrapolation tests, we managed to achieve useful uncertainties also for the faint extrapolation test, not covered by the training data.

As already mentioned, KiDS provides photometric redshifts for all cataloged galaxies, including quasars, and they are stored in the Z_B column (Kuijken et al., 2019). As these photo-z estimates were optimized for galaxies used for weak lensing studies, they are not expected to perform well for quasars in general. For comparison with our results, the mean error of the BPZ estimates for the QSOs in the random test is $\delta z = -0.38 \pm 0.43$, while in the extrapolation[2], $\delta z = -0.45 \pm 0.32$. The BPZ redshifts for QSOs are significantly underestimated and much less precise than our estimates: Their scatter is 3.5 times higher in the random test, and 1.7 times higher in the faint extrapolation test, in comparison to our results.

[2] We note that as BPZ is a template-fitting approach, its photo-z derivations are independent of the properties of training sets.

TABLE 4.2: ANN results on MASK flagged objects in the random ($r < 21.3$) and faint extrapolation ($21.3 < r < 22$) tests. Brackets show differences to corresponding ANN results from Table 4.1.

test	purity	recall	δz for true QSOs	δz for QSO candidates
random	96.41% (-0.49%)	93.98% (-0.74%)	0.004 (-0.005) ± 0.13 (+0.01)	0.019 (+0.000) ± 0.25 (+0.06)
faint extrap.	94.36% (-2.18%)	88.28% (-2.55%)	0.01 (+0.01) ± 0.22 (+0.03)	0.07 (+0.03) ± 0.41 (+0.04)

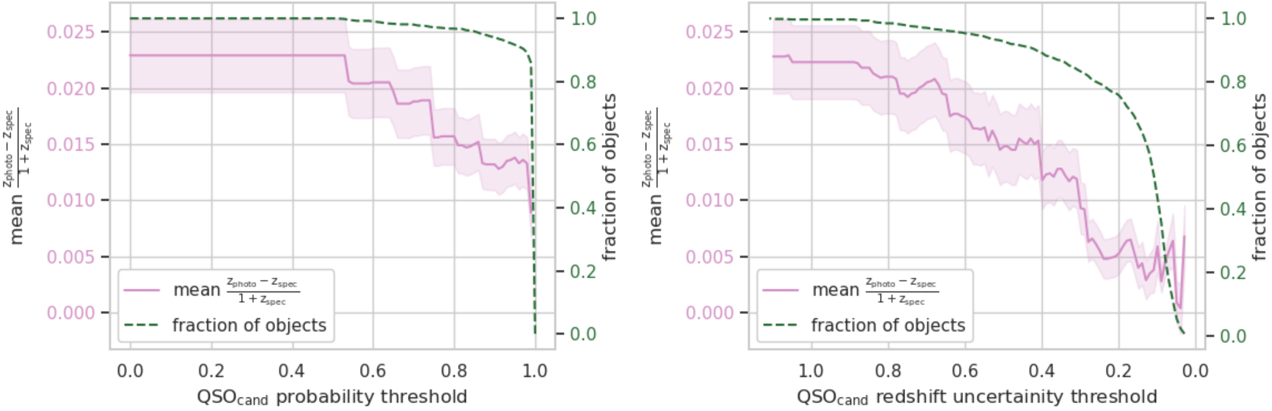

FIGURE 4.5: QSO photometric redshift errors as a function of thresholds in QSO probability (*left panel*) and model photo-z uncertainty (*right panel*). An increasing minimum classification probability yields better redshift estimations at a small cost in completeness. Low uncertainty estimations further increase redshift reliability at a cost of removing more objects.

The KiDS DR4 catalog provides a MASK flag indicating possible flux contamination from issues such as star halo, globular clusters, ISS, etc. We observe stability of the estimations in a random test on objects with such contamination. To verify this, we evaluated ANNs on the objects flagged with any MASK bit (Table 4.2). The results are stable in the random test and show some deterioration in the extrapolation test. We always include all masked objects in the training, so the models can learn how to process them, and the associated additional noise helps in regularization.

Classification and redshift results can be improved by limiting the sample to objects with higher classification probabilities or lower redshift uncertainties (Fig. 4.5). Similarly to our approach for KiDS DR3 described in the previous chapter, we consider the classification probability limits as the primary way to calibrate the catalog's purity-completeness trade-off, while the uncertainties can be used to achieve the necessary redshift precision.

4.5 Catalog properties

4.5.1 Number counts

In this Section we present and discuss the final quasar selection results in KiDS DR4. We applied the trained ML models to 45M objects of the KiDS DR4 inference data, and we found a total of 3M QSO candidates, excluding the unsafe inference subset. In the final model training, we used the whole range of magnitudes of the training set, as well as a randomly selected validation sample. We employed the same set of values of hyper-parameters as determined in the experiments which included the faint extrapolation test, and we only picked a new number of epochs based on new learning histories with a randomly selected test sample.

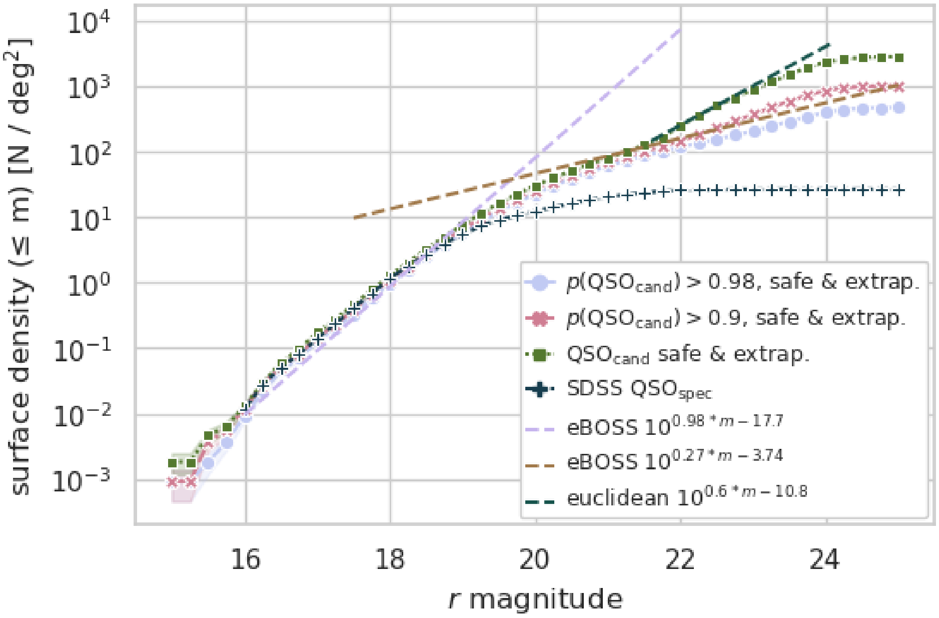

FIGURE 4.6: QSO number counts of SDSS spectroscopic QSOs and KiDS DR4 QSO candidates ($\mathrm{QSO_{cand}}$) at progressing classification probability cuts, excluding the unsafe inference subset. The dashed lines show eBOSS predictions fitted with a broken power law. The SDSS spectroscopic QSOs are complete to $r < 19$. KiDS QSO candidates without a probability cut are too numerous at $r > 21.5$ due to misclassification, and they follow standard Euclidean number counts. A cut at $p(\mathrm{QSO_{cand}}) > 0.9$ gives a complete catalog in the safe subset ($r < 22$). A cut at $p(\mathrm{QSO_{cand}}) > 0.98$ provides expected number counts up to $r \lesssim 24$.

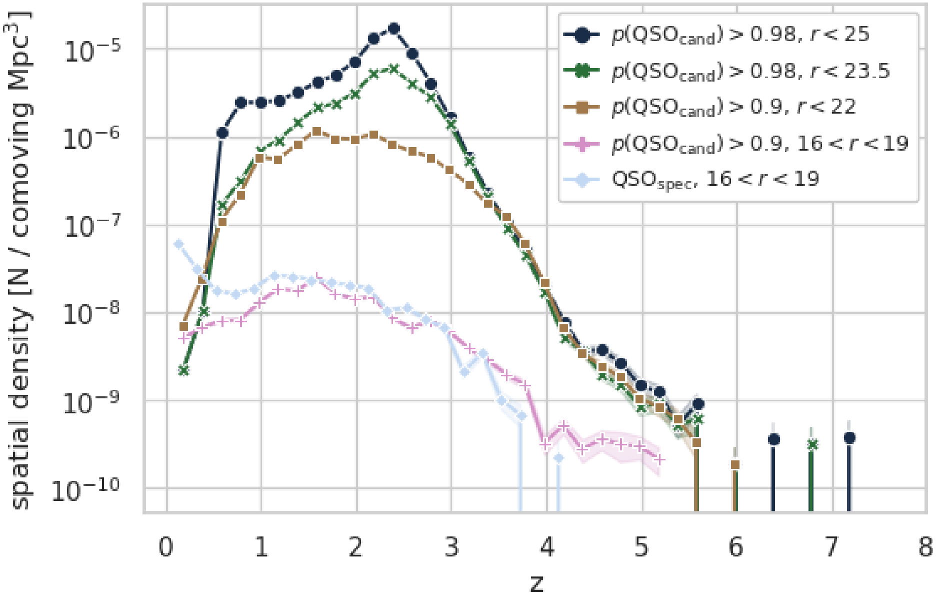

FIGURE 4.7: Spatial number densities, excluding the unsafe inference subset, for KiDS DR4 QSO candidates. Two bottom lines compare the KiDS QSO candidates to the SDSS spectroscopic QSOs at the SDSS completeness range of $16 < r < 19$. The three upper lines show the final QSO catalog at progressing magnitude limits with the suggested probability cuts. We chose a magnitude limit for the middle line at $r < 23.5$, as above this limit the distribution of QSO candidates gains another peak at redshift $z < 1.5$.

In Figure 4.6 we compare the number counts of QSO candidates (QSO_{cand}) in the safe and extrapolation subsets to the predictions from the eBOSS survey (table 7 from Palanque-Delabrouille et al., 2016). We fit the eBOSS predictions with a broken power law. Our analysis suggests that two cuts on the photometric QSO probability match the expected numbers: $p(\text{QSO}_{\text{cand}}) > 0.9$ for the safe magnitude range ($r < 22$) and $p(\text{QSO}_{\text{cand}}) > 0.98$ for the extrapolation. The fit of the QSO number counts to eBOSS predictions is reliable for $r < 23.5$, where the extrapolation subset is complete (Fig. 2.6). We do not observe the expected decrease in the QSO number counts at $r > 23.5$, which should result from reaching the completeness limit of the extrapolation subset. This suggests increased impurity of the QSO candidates in that range. The possible unreliability of the classification at $r > 23.5$ was already suggested by the t-SNE visualization in Fig. 2.5.

4.5.2 Spatial density

Figure 4.7 shows spatial number densities for KiDS QSO candidates based on the photometric redshifts and for SDSS spectroscopic QSOs based on the spectroscopic redshifts. We accounted for the V_{max} correction, taking the KiDS magnitude limit $r = 25$ and assuming the WMAP9 (Hinshaw et al., 2013) cosmology. The distribution is expected to peak at $z \sim 2$ - 3 and then follow an exponential decrease (Fan, 2006). Based on the SDSS spectroscopic QSO number counts (Fig. 4.6), we estimated its completeness to be $r < 19$. We observe some differences between KiDS photometric and SDSS spectroscopic QSO densities at this limit. The QSOs missing at low redshifts are due to the previously discussed misclassification with galaxies (Fig. 4.3). At the faintest end ($r > 23.5$), on the other hand, the photo-z-based density displays an additional peak at $z < 1$ for the suggested $p(\text{QSO}_{\text{cand}}) > 0.98$. This is due to

TABLE 4.3: Number of photometrically selected QSOs in our catalog at progressing magnitudes with the suggested probability cuts (bold), excluding the unsafe inference subset. At fainter magnitudes, a higher probability threshold is required for robustness. The cuts give smaller subsets of QSO candidates and increase the purity.

	safe $r < 22$	safe & extrap. $r < 23.5$	safe & extrap. $r < 25$
$\mathrm{QSO}_{\mathrm{cand}}$	266k (100%)	1.6M (100%)	3M (100%)
$p(\mathrm{QSO}_{\mathrm{cand}}) > 0.90$	**158k (59%)**	637k (39%)	1.1M (36%)
$p(\mathrm{QSO}_{\mathrm{cand}}) > 0.98$	127k (48%)	**311k (19%)**	507k (17%)

apparently faint galaxies classified by our model as QSOs and assigned redshifts lower than one. This conclusion agrees with the number counts indicating a QSO impurity at $r > 23.5$.

Table 4.3 summarizes the number of QSOs in the final catalog at progressing magnitude limits – thus reliability limits – and the suggested probability cuts. According to the number counts and spatial number densities, the QSO classification and redshift estimations should be reliable up to $r < 23.5$. At $r > 23.5$, the classification provides excessive number counts, and the photometric redshifts suggest misclassification with galaxies. The forthcoming DESI and planned 4MOST QSO surveys could help verify these finding, as they will include QSOs fainter than SDSS and will overlap with KiDS.

We visualized the outputs from the ML models, compared it to the spectroscopic information, and show the final catalog properties for the inference subsets and suggested probability cuts using t-SNE in Fig. 4.8. The main spectroscopic QSO group is accurately covered with photometric classification and redshifts. In the close extrapolation, the predictions appear as a regular extension of the main QSO group, which qualitatively confirms the success of our approach. The decision to separate out the unsafe inference subset is confirmed, as we observe the distributions of all three classes overlapping in the corresponding part of the feature space. The estimations for fainter magnitudes could be used to look for QSOs at the highest redshifts or to select candidates for follow-up spectroscopy.

4.5.3 Gaia parallaxes

We cross-matched the QSO candidates identified here with Gaia DR2 (Gaia Collaboration et al., 2018a) to estimate the star contamination. As described in the previous chapter (§3.4.1), a clean set of QSOs is expected to have a global mean parallax offset of −0.029 mas (Lindegren et al., 2018). This value was calculated by removing incorrectly measured high parallaxes for SDSS QSOs. Following the same procedure for KiDS, QSO candidates would remove the star contamination, which we want to measure. Instead, we calculated a less precise mean offset for SDSS QSOs in a high precision sample with parallax and proper motion errors smaller than 1 mas. This offset equals −0.017 mas, which is smaller in absolute terms than the official Gaia measurement.

The QSO candidates in the KiDS DR4 safe inference subset show a mean parallax offset of 0.003 mas, and this goes down at the progressing minimum classification probability (Fig. 4.9). This assessment is based on a cross-match between our catalog and the Gaia high precision sample mentioned above, which yields 1.63M objects: 1.61M (98.7%) classified photometrically as stars, 20k (1.2%) as QSOs, and 1k (0.1%) as galaxies. The test is limited to the Gaia magnitude $G < 21$, which corresponds to $r \lessapprox 20$. We then calculated an "acceptable offset" from a sample of the three spectroscopic classes, with the size of each class corresponding to the contamination of QSO candidates with stars and galaxies derived from the experiments: 96.9% QSOs, 2.6% galaxies, and 0.4% stars (Fig. 4.3). The minimum QSO photometric probability suggested by this test is $p(\mathrm{QSO}_{\mathrm{cand}}) = 0.9$. This cut, which

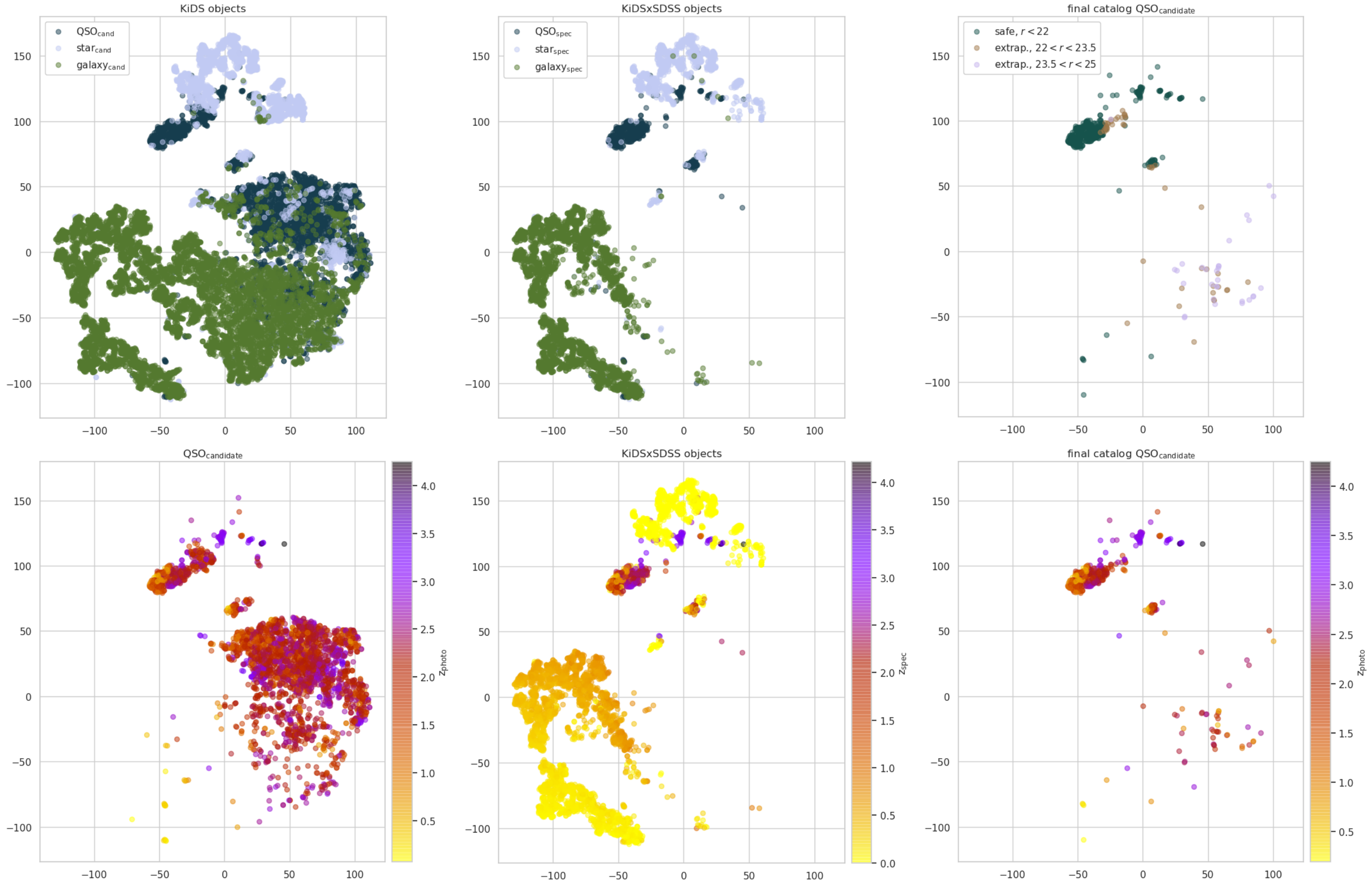

FIGURE 4.8: t-SNE projections. Top: Classification. Bottom: Redshifts. Left: Raw output from the ML models for all the inference subsets. Center: Spectroscopic SDSS distributions. Right: Final QSO catalog at progressing magnitudes with the corresponding probability cuts, excluding the unsafe inference. The visualizations were made on a subset of 12k objects, thus actual object density at any part of the feature space is 3.8k times higher.

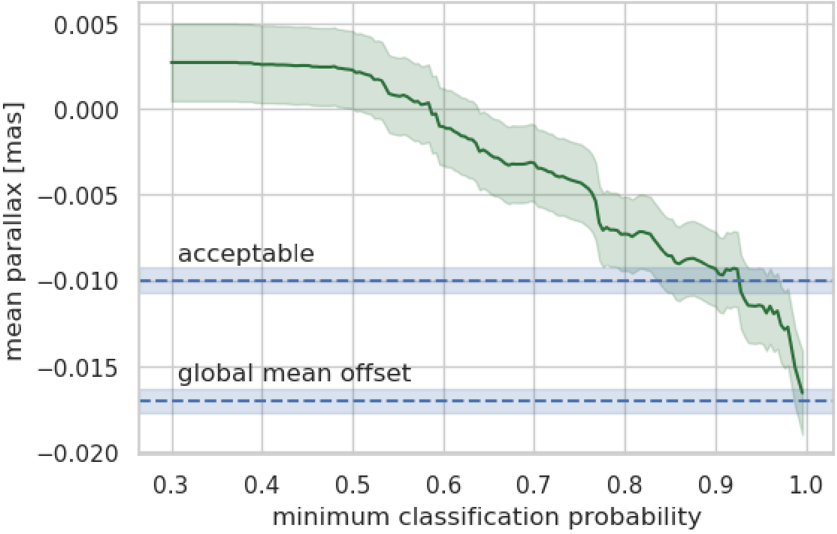

FIGURE 4.9: Mean parallax for KiDS DR4 QSO candidates as a function of minimum classification probability. The Gaia observations have a global mean offset, which is imprinted in the QSO mean parallax distribution. The offset for SDSS spectroscopic QSOs equals -0.017 ± 0.001 mas (standard error on the mean). We calculated the acceptable offset based on star and galaxy contamination estimated in the experiments. It equals -0.01 ± 0.0015 mas.

was obtained from the more precise test at $r \lessapprox 20$, agrees with the cut for the safe inference subset at $r < 22$ derived from the number counts.

4.5.4 External quasar catalogs

We find good agreement with other QSO catalogs overlapping with the KiDS DR4 footprint (Fig. 4.10). Additional ground-truth samples, which were not used in the training, provide a good test of ML estimations. We used the same QSO catalogs as in §3.4.3, built from different datasets and with different methodologies than ours. 2QZ, being spectroscopic, can be used as ground truth and confirms high QSO purity and completeness of our sample: 98.2% three class accuracy, 98.6% QSO purity, and 99.4% QSO completeness. We note, however, that as 2QZ sources are on average brighter than those from the SDSS QSO catalog, these numbers should not be taken as measurements of the overall performance of our classification.

4.6 Summary

In this Chapter, we presented a more complex approach in comparison to the one employed for KiDS DR3, in which we add near-infrared imaging, estimate photometric redshifts, and extrapolate to magnitudes fainter than available in the spectroscopy. We defined inference subsets from the 45 million objects of the KiDS photometric data limited to 9-band detections, based on a feature space built from magnitudes and their combinations. We show that projections of the high-dimensional feature space on two dimensions can be successfully used, instead of the standard color-color plots, to investigate the photometric estimations, compare them with spectroscopic data, and efficiently support the process of building a catalog. The model selection and fine-tuning employs two subsets of objects: those randomly selected and the faintest ones, which allowed us to properly fit the bias versus variance trade-off. We tested three ML models: random forest (RF), XGBoost (XGB), and artificial neural

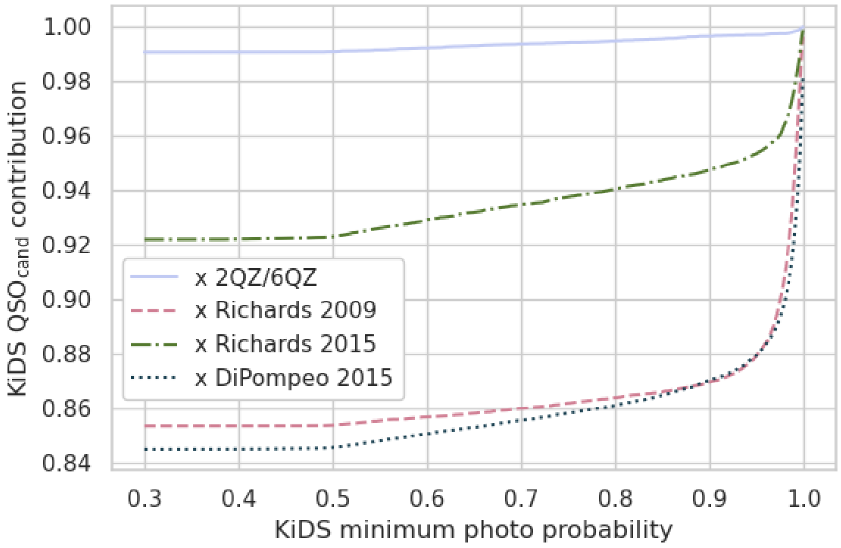

FIGURE 4.10: Proportion of KiDS DR4 QSO candidates in cross-matches with other QSO catalogs as a function of KiDS minimum photometric classification probability.

network (ANN). We find that XGB is the most robust and straightforward model for classification, while ANN performs the best for combined classification and redshift. The ANN inference results are tested using number counts, Gaia parallaxes, and other quasar catalogs that are external to the training set. We found 158k QSO candidates with a minimum classification probability of $p(\mathrm{QSO_{cand}}) > 0.9$ at $r < 22$, and a total of 311k QSO candidates with $p(\mathrm{QSO_{cand}}) > 0.98$ for $r < 23.5$, that is to say in the extension to the close extrapolation data. The far extrapolation at $r < 25$ provides a total of 507k QSO candidates at $p(\mathrm{QSO_{cand}}) > 0.98$. The catalog of QSOs is well designed for extrapolation, with the reliability regions derived from visualizations, and probability thresholds calibrated via a series of tests. Based on the SDSS QSO test sample, the purity of the catalog is 96.9%, and completeness is 94.7% for $r < 22$, which is better than in case of KiDS DR3. The extrapolation by ~0.7 magnitude lowers the purity by 0.4 percentage points and the completeness by 3.9 percentage points. The average redshift error in terms of $(z_{\mathrm{photo}} - z_{\mathrm{spec}})/(1 + z_{\mathrm{spec}})$ equals 0.009 ± 0.12 for $r < 22$, with its scatter increasing to -0.0001 ± 0.19 in the extrapolation ($r < 23.5$). The catalog can be empoyed for cosmological and clustering analysis, as we do in the next Chapter.

5

Quasar correlation functions and bias constraints

In this chapter, we show the results of the KiDS DR4 quasar catalog (§4.5) auto- and cross-correlation with CMB lensing, based on the methodology from §2.5. Our goals are to establish if the catalog of quasars detected with machine learning provides signals of the aforementioned observables which allows to measure the bias function, and test if fainter magnitudes accessible in our catalog allow to constrain bias also for the fainter galaxies. We use the definitions of auto- and cross-correlations as given in §2.5.1, and perform experiments for quasars as outlined in §2.5.2. The chapter is organized as follows. In §5.1 we chose the optimal quasar probability cuts for different magnitudes based on the auto-correlation measurement. In §5.2 we constraint the quasar bias based on the auto- and cross-correlations for different magnitude cuts and the optimal quasar probability cuts. All the results were obtained by the thesis author, unless stated otherwise.

5.1 Auto-correlation

We use photometric redshifts, as shown in Fig. 5.1, to obtain the theoretical prediction of the correlation functions. The distributions differ significantly, with the peak at $z \sim 2.2$ increasing with the fainter magnitudes. This is due to fainter magnitudes probing more the main quasar family located at $z > 2$. The photo-z distribution is limited at $z_{photo} > 0.5$, as our training sample quasars also had $z_{spec} > 0.5$. We obtain little unphysical peaks in the distribution (e.g. $z \sim 1.9$) as an artifact from from the neural network processing. It might affect the theoretical modelling of the auto-correlation, but should not have any vital effect on the modelling of the cross-correlation power spectrum with CMB lensing, which we use to estimate the bias of the sample. A more advanced approach might be to use uncertainties in redshift estimation from neural networks to assign errors to redshift distribution and marginalize over parametrization to include those uncertainties in the bias estimates, and smooth the redshift distribution.

We use the KiDS DR4 quasar catalog without the objects tagged as unsafe in the inference. We test a couple of magnitude cuts around the SDSS limit, namely r lower than 21, 22, 23, and 23.5. We pick several probability cuts, in order to optimize the purity vs completeness trade-off and obtain signals which are not contaminated with stars and provide the highest detection significance. The cuts at 90% and 98% are suggested by the number counts and parallax measurements (sec. 4.5), and here we add cuts at 99.8% and 99.9% in order to further test more aggressive selection which decrease contamination with stars to a possible minimum. Figure 5.2 shows a histogram of probabilities for the quasar candidates at different magnitude cuts.

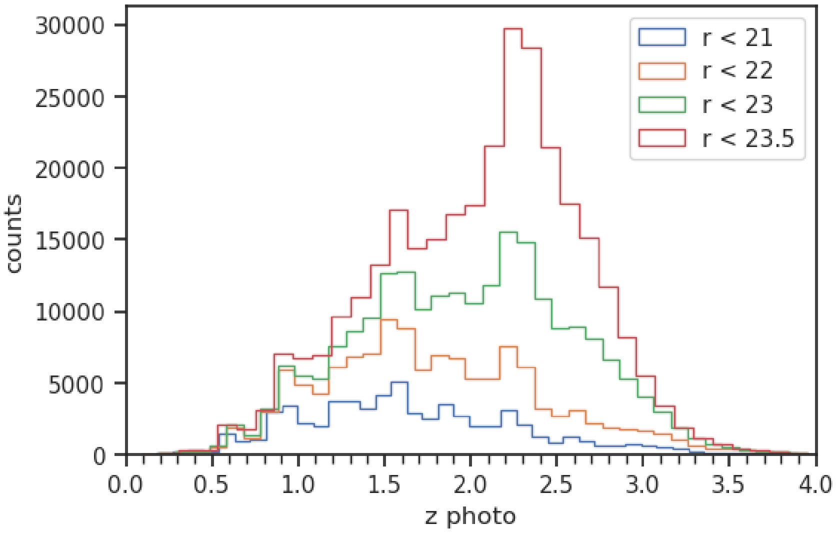

FIGURE 5.1: Photometric redshift distribution at different magnitude cuts for quasar catalog limited at the most aggressive cut as suggested by the number counts, the $p(\text{QSO}_{\text{cand}}) > 0.98$.

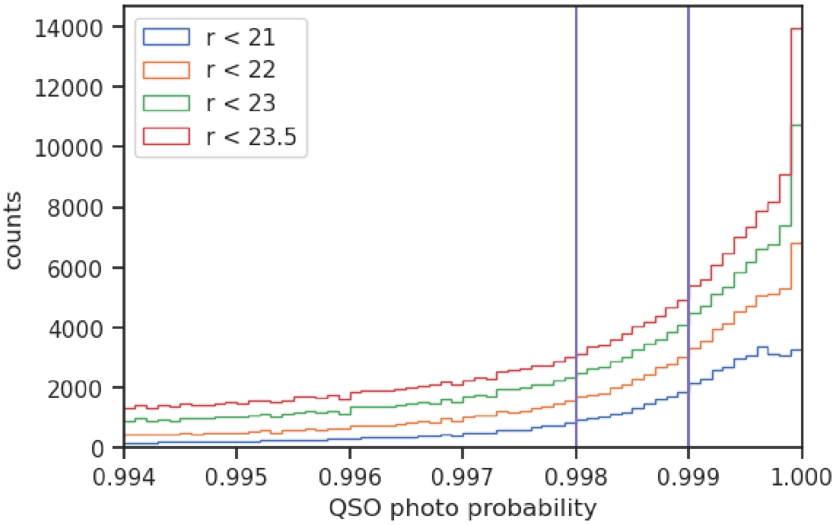

FIGURE 5.2: Histograms of quasar probability at the magnitude cuts of interest. The lines show the cuts at 99.8% and 99.9%.

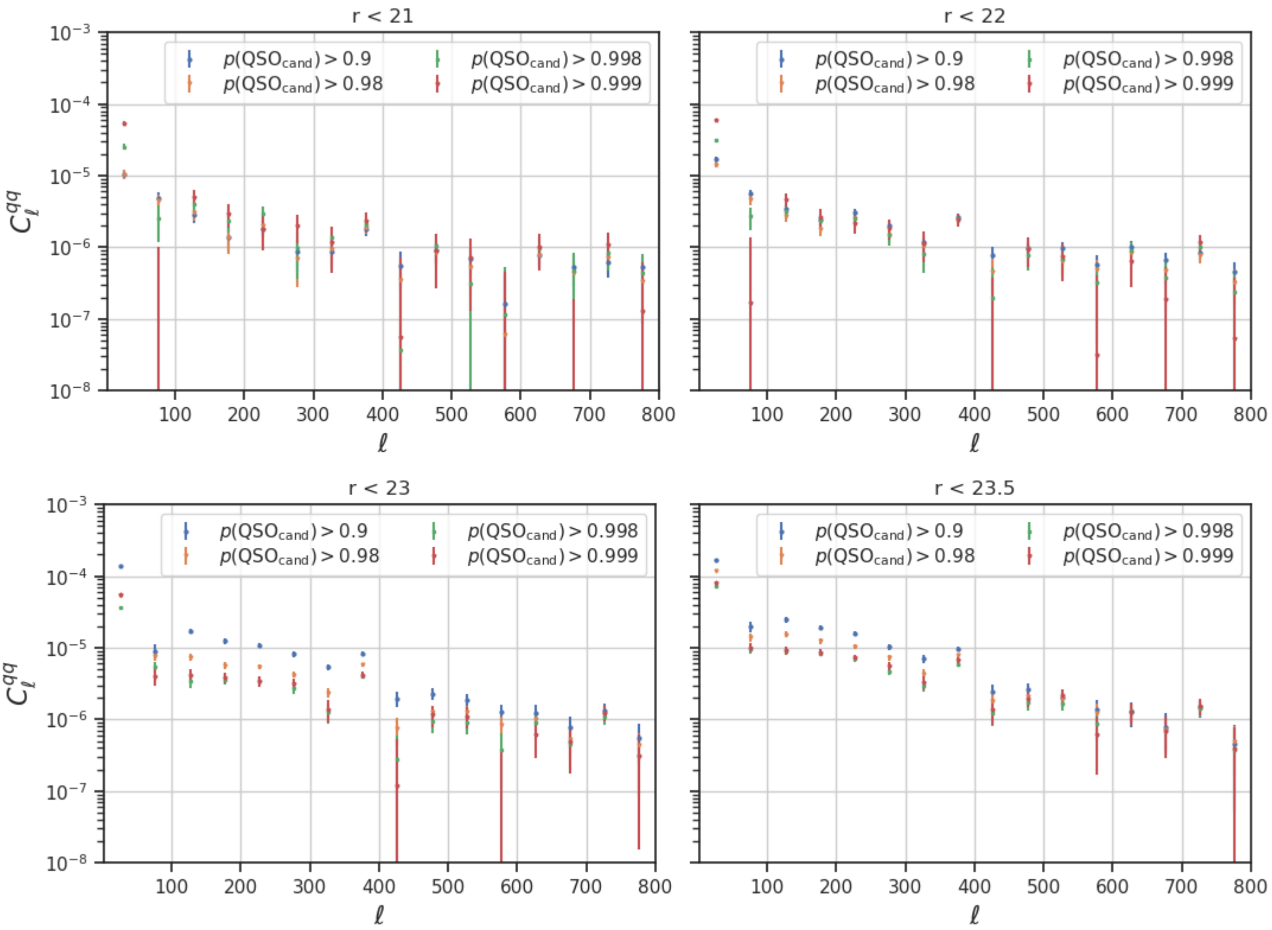

FIGURE 5.3: The comparison of probability cuts for different magnitude limits. We show here the range of the multipoles lower than 800 to highlight the differences at the scales between $100 < l < 400$. Figure 5.7 shows the auto-correlation at the higher, $l < 1500$, multipole range.

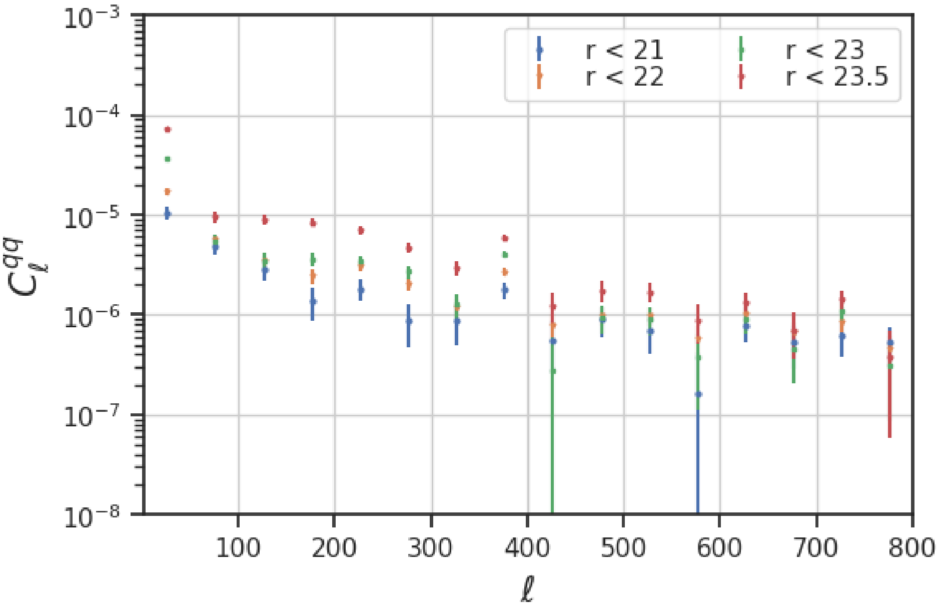

FIGURE 5.4: Comparison of the auto-correlation for different magnitude cuts. For each magnitude cut corresponds a different optimal quasar probability cut, as given in the table 5.1.

We chose the best probability cuts for each magnitude based on the auto-correlation power spectrum. Stars do not correlate with quasars, but their own power spectrum has different amplitude and shape than the C_{qq} and will contaminate the signal. Choosing higher probability cuts allows to remove stars from the catalog and clean the auto-correlation signal. If the removed objects were mostly quasars, we would expect to obtain higher errors of the C_{qq}, but agreement in amplitude and shape, as we are correcting for the shot noise which depends on the objects density. We tested it in this particular pipeline, that using a random sample of half quasars does provide the aforementioned effect. Figure 5.4 shows the effects of choosing higher probability cuts for the magnitude cuts which chose for the tests. Each pair of magnitude and probability cut gives different value of the shot noise, and we always calculate the shot noise independently for each sample of interest. For the magnitudes $r < 21$ and $r < 22$, we obtain agreement in amplitude and shape of the power spectra, which means that either the quasar sample is clean of any contaminants, or it is no longer possible to remove more stars based on the probability. It also means that number counts and parallax analysis from the section 4.5 were enough to remove the star contaminant. For the magnitudes $r < 23$ and $r < 23.5$ we can see that the amplitude and shape changes, but remains the same at above 99.8%, which is higher than the 98% suggested earlier. It shows that the auto-correlation is sensitive to star contamination, and it is important to use the auto-correlation as a catalog test at fainter magnitudes which include the problem of extrapolation. We chose 99.8% probability cut as the desired one for $r < 23$ and $r < 23.5$ magnitude cuts. Because we still cannot be sure about possible star contamination, we also use cross-correlation with CMB lensing in order to constraint the bias function of the quasars. In order to further test the catalog's purity, one can also measure cross-correlation between stars and quasars, or quasars at different redshift ranges, as in both cases we would expect zero correlation if the quasar catalog is free of any star contaminants. However, in this case we decide the performed tests as enough to properly measure the auto-correlation power spectrum.

Table 5.1 includes the desired probability cuts, as well as median redshift and surface density for those. For fainter magnitudes we obtain objects at higher median redshift, which

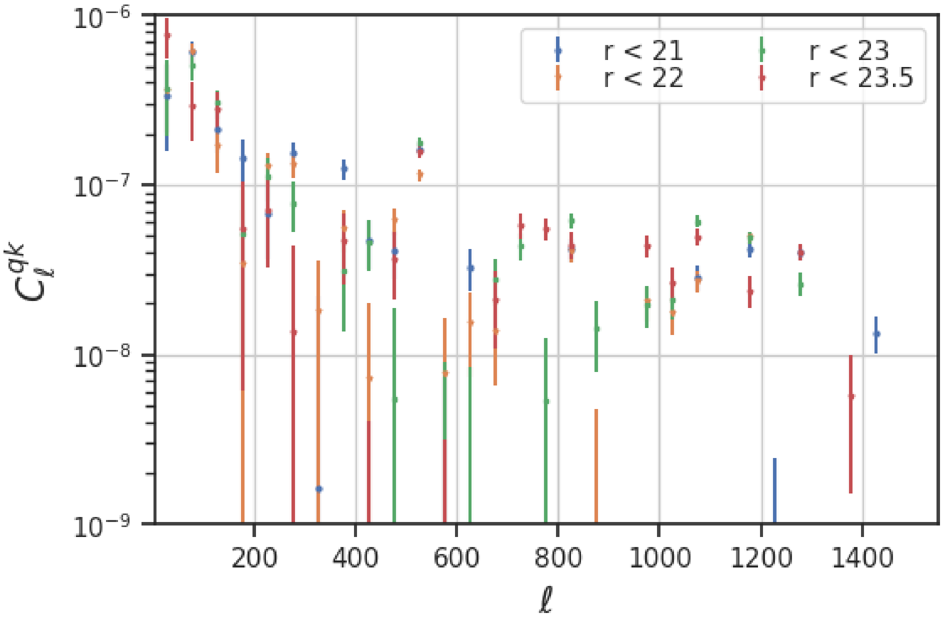

FIGURE 5.5: Comparison of the cross-correlation with the CMB lensing for different magnitude cuts and $p(\mathrm{QSO_{cand}})$ adapted to each magnitude limit (see Table 5.1).

TABLE 5.1: The probability cuts as suggested by the auto-correlation, median photo-z and surface density, the resulting bias constrains in the form of $b_q(z) = A(1+z)^2 + B$, and the χ^2 of the fit.

	p(QSO)	median photo-z	density (deg^{-2})	A	B	$\chi^2(C_{q\kappa})$
$r < 21$	90%	1.58	73	$0.49^{+0.05}_{-0.04}$	$0.26^{+0.19}_{-0.33}$	159
$r < 22$	90%	1.71	157	$0.62^{+0.03}_{-0.03}$	$0.07^{+0.05}_{-0.13}$	282
$r < 23$	99.8%	1.96	95	$0.57^{+0.03}_{-0.03}$	$0.07^{+0.06}_{-0.13}$	336
$r < 23.5$	99.8%	2.10	116	$0.84^{+0.02}_{-0.02}$	$0.05^{+0.04}_{-0.08}$	624

also means higher bias, and we would expect different amplitudes of the auto-correlation at different magnitude limits. Figure 5.4 compares the auto-correlation functions for the desired probability cuts. We obtain signals for each of the magnitude limits, which we will use to fit the theoretical models in the section 5.2. We can see that the fainter magnitudes give higher amplitude, which is reasonable considering that those quasars reside at higher redshift and have higher bias values.

5.2 Bias constraints with the CMB lensing

Figure 5.5 compares cross-correlation with the CMB lensing at different magnitude cuts. We can see that the brighter magnitudes provide higher amplitude at larger scales, i.e. $l < 500$, while higher quasar density available at fainter magnitudes allows to detect the signal at smaller scales, i.e. $l > 700$.

We constrain bias of the quasar sample, by assuming the redshift distribution as given by the photo-z, and using a bias model given by $b_q(z) = A(1+z)^2 + B$. We use two degrees of freedom, the A and B parameters, and cosmology fixed to the Planck Collaboration et al., 2020a values. Figure 5.6 shows the resulting bias fits for different magnitude cuts in comparison to the Sherwin et al., 2012. Out results at different magnitudes provide different

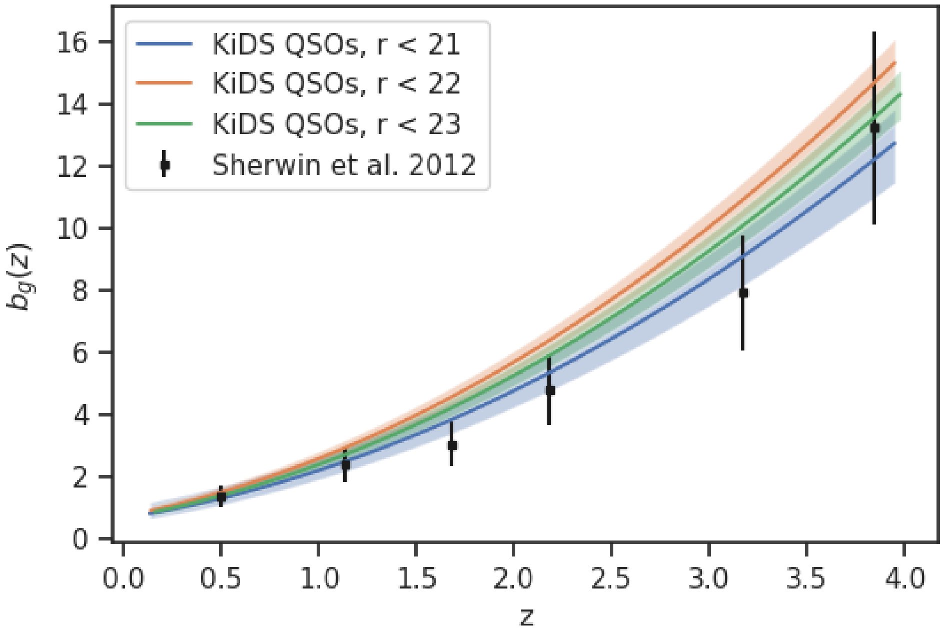

FIGURE 5.6: The resulting bias fits at different magnitude cuts, in comparison to the results from the Sherwin et al., 2012, based on auto- and cross-correlations between KiDS DR4 QSOs and CMB lensing map derived from the Planck Collaboration et al., 2020b observations, fitted with theoretical models assuming the Planck cosmology (Planck Collaboration et al., 2020a), photo-z redshift distributions, and bias functions modelled with $b_q(z) = A(1+z)^2 + B$.

amplitudes of the bias evolution. The quasar sample limited at $r < 21$ is in agreement with the Sherwin et al., 2012, the sample at $r < 23$ is at about 1 sigma difference, and the sample at $r < 23$ provides higher amplitude starting with $z > 1.5$. At redshift 0 we obtain bias lower than 1, but it is not a problem since our quasar sample is limited to $z > 0.5$ and those redshifts cannot be constrained with the given quasar catalog.

Table 5.1 shows the constraints of the A and B parameters, as well as the surface density and median photo-z of the chosen magnitude and probability cuts. The subset at $r < 23$ provides lower surface density due to the higher probability cut. It is not a problem, as the auto-correlation functions at $r < 22$ provide the same amplitude and shape at different probability cuts, hence, the lower number of objects at $r < 23$ due to higher probability cut applied also to data at $r < 22$ might increase the errors, but does not change the amplitude or shape of the correlation functions. Despite the lower surface density, fainter magnitudes probe higher redshifts which is shown with the median photo-z value. The sample limited at $r < 23.5$ gives the highest bias evolution, which does not agree with the Sherwin et al., 2012. At this magnitude range, it might be necessary to account for errors in the redshift distribution based on the photo-zs. The χ^2 of the C_{qk} increases with the magnitude, but at the $r < 21$ it provides the highest detection of the signal, with 15 sigma significance, in comparison to, for instance, 3.8σ reported by Sherwin et al., 2012. The final fits to the C_{qq} and C_{qk} are shown in the Fig. 5.7.

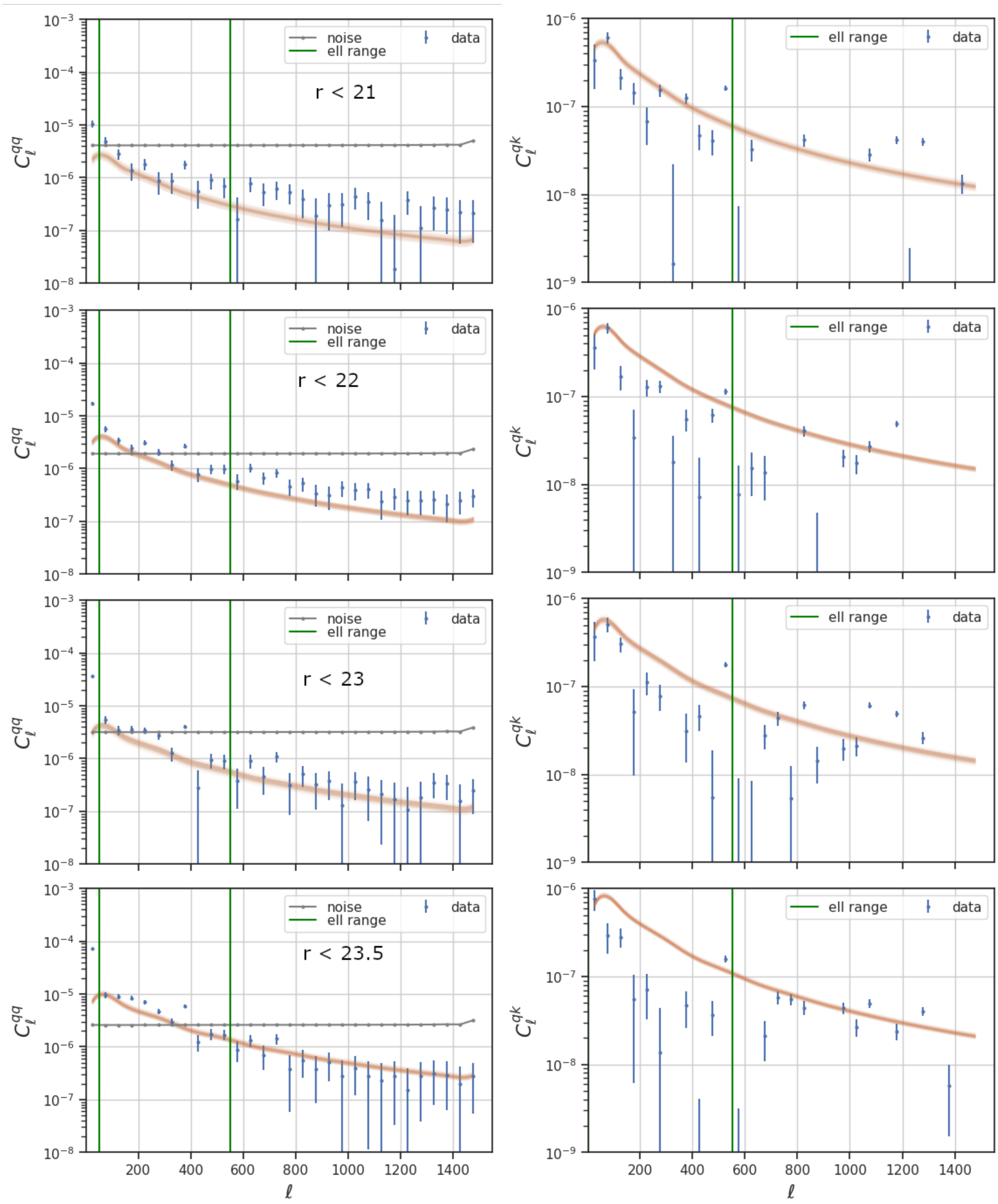

FIGURE 5.7: Auto- and cross-correlations between KiDS DR4 QSOs and CMB lensing map derived from the Planck Collaboration et al., 2020b observations, fitted with a theoretical model assuming the Planck cosmology (Planck Collaboration et al., 2020a), photo-z redshift distribution, and bias modelled with $b_q(z) = A(1+z)^2 + B$. The orange line shows the theoretical power spectra from the MCMC. The green line marks the range of the multipoles used to make the fit.

6

Discussion

This chapter is based on the publications Nakoneczny et al., 2021; Nakoneczny et al., 2019.

6.1 Main findings

In this work, we employed supervised ML models to identify QSOs in KiDS DR3 and DR4, and evaluate their redshifts in the case of DR4. In the conclusions, we focus on the catalog from the KiDS DR4, which is the most sophisticated one. In there, we found 158k QSO candidates with a minimum classification probability of $p(\mathrm{QSO_{cand}}) > 0.9$ at $r < 22$, and a total of 311k QSO candidates with $p(\mathrm{QSO_{cand}}) > 0.98$ for $r < 23.5$, that is to say in the extension to the close extrapolation data. The far extrapolation at $r < 25$ provides a total of 507k QSO candidates at $p(\mathrm{QSO_{cand}}) > 0.98$. The catalog of QSOs is well designed for extrapolation, with the reliability regions derived from visualizations, and probability thresholds calibrated via a series of tests. Based on the SDSS QSO test sample, the purity of the catalog is 96.9%, and completeness is 94.7% for $r < 22$. The extrapolation by ~0.7 magnitude lowers the purity by 0.4 percentage points and the completeness by 3.9 percentage points. The average redshift error in terms of $(z_{\mathrm{photo}} - z_{\mathrm{spec}})/(1 + z_{\mathrm{spec}})$ equals 0.009 ± 0.12 for $r < 22$, with its scatter increasing to -0.0001 ± 0.19 in the extrapolation ($r < 23.5$). Additionally, we measured the quasar bias function, which equals $b_q(z) = 0.57^{+0.03}_{-0.03}(1 + z)^2 + 0.07^{+0.06}_{-0.13}$, and at redshift $z = 1.5$ gives a value $3.63^{+0.25}_{-0.85}$. Finally, we report 15σ significance of the cross-correlation with the CMB lensing.

We found that the traditionally adopted testing method, based on randomly selected samples of objects, was insufficient to tune the bias versus variance trade-off. A faint-end test is necessary for the proper extrapolation of both classification and redshifts, but also important for appropriate tuning and inference on the bright end data. This approach towards ML model calibration and the satisfactory extrapolation results are the main novelty aspects of our work. Thanks to the faint extrapolation test, we also obtain useful redshift uncertainties in the extrapolation data, even though we used the Gaussian output layer to model aleatoric uncertainty. Otherwise, we would expect the aleatoric uncertainty to fail in the part of the feature space not covered by the training data.

The addition of the near-IR VIKING bands, which were not available in the KiDS DR3, provided crucial information for QSO redshifts and helped us to distinguish stars from QSOs at redshifts of $2 < z < 3$. The most important bands for QSO redshifts, according to our experiments, are the near-IR ZK_s, which are the two extreme bands covered by VIKING. This suggests that it is the span of the infrared wavelengths that is relevant here. We found it important to use both magnitude differences (colors) and magnitude ratios. Interestingly, colors and ratios constructed from the same magnitude pairs had a different importance for the ML models. What is more, the ratios were in fact more common than colors among the most important features used by XGBoost for classification and QSO redshifts. This experimental analysis could be further perfected with proper fine tuning of the models trained using a no-ratio feature set in order to draw the final conclusions. Additionally, possible further experiments may involve more custom feature engineering based on flux values in order to find the most robust photometric features.

The comparison of ML models also shows clear trends: XGB performs better at classification, while ANN provides a better redshift estimation, that is to say it works better for regression. Many astronomical papers report no such differences, which was also the case in our approach to the KiDS DR3. We uncovered these differences as more features are available from the VIKING imaging, which allowed us to obtain better results with more sophisticated classification models such as XGB. The superiority of ANN for regression is largely due to its better performance in extrapolation, not only in feature space, but also in higher values of the estimated photo-zs. The models tuned for both random and faint extrapolation tests are also less overfitted and show real differences between their characteristics.

We successfully supported our analysis with t-SNE projections of high-dimensional space onto 2D, instead of the standard color-color plots. The visualizations helped us to derive a reliable inference subset at close extrapolation, which was possible by verifying the location of these extrapolation data with respect to the feature space known from spectroscopic classification. We also used the projections to test different feature sets. The distribution of spectroscopic classes on the t-SNE plots allowed us to initially assess the reliability of feature engineering, without even training a supervised model. Last but not least, the visualizations helped us understand where the classification fails due to overlapping distributions between various object classes in the feature space.

We release the catalogs publicly at
http://kids.strw.leidenuniv.nl/DR3/quasarcatalog.php (DR4) and
http://kids.strw.leidenuniv.nl/DR3/quasarcatalog.php (DR3).
Description of the data format is provided in the webpages. The code, written in Python and Jupyter Notebook, is shared publicly at https://github.com/snakoneczny/kids-quasars.

6.2 Relation to other work

Most of the QSO classification and redshift estimation studies are not directly comparable due to the results depending on available bands, survey brightness, size of the training sample, and different definitions or detection schemes of AGNs and QSOs in spectroscopy and photometry. To ensure both high purity and completeness using color-color cuts, one has to model the data with many distributions or build a set of decision boundaries (e.g., Richards et al., 2002). On the other hand, ML allows us to build the most complicated decision boundaries in an automatic way, while simultaneously optimizing both purity and completeness. The power of ML approaches comes with the danger of possible overfitting. This problem is usually not addressed in the ML analyses, and the results on faint end data, which are most affected by overfitting, are rarely reported (e.g., Hausen and Robertson, 2020). As far as we know, our results for data fainter by one magnitude than the reach of the training data – completeness lower by 3% and redshift scatter increased by 0.07 in comparison to the regime covered by the training – are reported for the first time. This outcome challenges the way that ML models are usually optimized and applied on the faint data end. For other problems, other data characteristics can be used to obtain extrapolation tests, for example, the high and low mass end for galaxy cluster mass estimation, the number of objects in n-body problems, cosmological parameters not available during training in cosmological problems, etc.

Our work is the first in which the simultaneous selection of QSOs from photometry and evaluation of their photometric redshifts is performed for samples selected from the KiDS+VIKING catalog. In a recent study, Logan and Fotopoulou (2020, p. L20) performed a classification and redshift estimation in KiDS DR4, but on a smaller subset of 2.7M objects selected over 200 deg^2 with the additional requirement of available detections in the WISE mid-IR bands. That classification was done with unsupervised hierarchical density-based spatial clustering of applications with noise (HDBSCAN, McInnes, Healy, and Astels, 2017), redshift estimation with RF, and feature engineering with principal component analysis (PCA, Pearson, 1901). A quantitative comparison of our catalogs with respect to experimental results on SDSS data is not possible due to different train and validation strategies. We have, however, performed a qualitative comparison using the full training data from the L20 catalog. The classification results are different as L20 uses an unsupervised algorithm, which does not allow for a completeness that is as high as our supervised approach. We find our photo-zs to be more precise on average, but L20 photo-zs are more robust at the faint end.

In the study for KiDS DR3, we employed the RF algorithm and reported 91% purity and 87% completeness for QSOs. In the approach to the KiDS DR4, most of the improvement in classification comes from adding the NIR bands, which allowed us to correctly classify QSO at $2.5 < z < 3$, where they are similar to stars in the *ugri* broad-bands. Additionally, two significant improvements were made: we provide QSO photometric redshifts, and publish estimations for objects fainter than the training data, with models tuned for extrapolation.

Another related work is the KiDS Strongly lensed QUAsar Detection project (KiDS-SQuaD; Spiniello et al., 2018; Khramtsov et al., 2019), aimed at finding strongly gravitationally lensed quasars in the KiDS data. This latter paper in particular describes the KiDS Bright EXtraGalactic Objects catalog (KiDS-BEXGO), constructed from DR4 and including about 200k sources identified as QSOs based on an application of the CatBoost gradient boosting ensemble algorithm (Prokhorenkova et al., 2018). The BEXGO catalog is optimized for the lowest possible star contamination at a cost of reduced completeness, and it is limited to $r < 22$. The results of an ML QSO identification are not directly comparable between our work and that of Khramtsov et al., 2019, as in the latter the QSOs are defined as point-like objects, and any AGNs with a visible galaxy host had been removed from the training data, unlike in our case. We have kept QSOs, which appear extended in our training data, as such sources provide useful information on the relation between QSOs and galaxies at low redshifts. It might have a vital outcome on the final predictions and possibly makes both catalogs different.

Furthermore, the dataset constructed by Khramtsov et al., 2019 is aimed to carry out the specific purpose of QSO strong lensing, which requires the highest possible purity of the catalog. The approach that we have taken, on the other hand, is to obtain the most optimal purity-completeness trade-off, which requires ML models to be properly tuned to the given problem and data. A required level of purity or completeness can then be acquired a posteriori by properly calibrating the catalog, in particular by applying appropriate cuts on the probability that a given source is a QSO.

In this work, we trained the ML models to perform a full three-class classification on both extended and point-like objects. If instead one was not interested in AGNs with resolved galaxy hosts, but only point-like QSOs at higher redshifts, then based on the finding of our work, we suggest to train the ML classifier only on point-like objects – for example, those with the stellarity index higher than 0.8 – and apply only QSO versus star classification. Such a model is easier to train and interpret, and visualizations of the relevant data are simpler to understand than in the full three-class problem including both extended and point sources.

Since the publication, the catalogs have been used in various applications, including QSO target selection in SDSS DR17 (Abdurro’uf et al., 2022), training and validating classification models (Logan and Fotopoulou, 2020; Falocco, Carrera, and Larsson, 2022), testing QSO contamination (Bilicki et al., 2021; Stringer et al., 2021), and supporting feature selection decisions (Carvajal et al., 2021; Chan et al., 2022; Li et al., 2022).

We envisage that our catalog of QSOs can have versatile applications in studies related to AGNs or LSS, as it is optimized solely for QSO identification without outside requirements. The availability of robust photometric redshifts with uncertainty estimates for the QSOs contained in our catalog is expected to prove especially useful in approaches where "tomographic" dissection of the LSS is done, such as cross-correlations with various backgrounds.

Laurent et al., 2017 uses spectroscopic quasar from the SDSS-IV eBOSS over redshift range 0.9 to 2.2, and analyse their clustering. They report a precise bias value of 2.45 ± 0.05 at redshift $z = 1.55$. Sherwin et al., 2012 uses photometric quasars from the SDSS DR8 over redshift range of 0.5 to 3.5, in cross-correlation with the CMB lensing, and report the bias value of 2.5 ± 0.6 at $z = 1.4$, and obtain 3.8σ detection of the cross-correlation. In comparison, we analyse a redshift range spanning from 0.5 to 3.5, and report bias value $3.63^{+0.25}_{-0.85}$ at $z = 1.5$, and 15σ significance of the CMB lensing cross-correlation.

6.3 Limitations and possible improvements

We consider our approach towards the inference at the faint data end, which involves tuning the model based on a faint extrapolation test, as the most optimal as far as the current supervised ML models are concerned. However, a reliable test of our predictions outside of the magnitude coverage of spectroscopic samples is not possible, and at present KiDS does not overlap with any wide-angle samples providing sufficient numbers of spectroscopic QSOs beyond $r > 22$. This situation will likely improve in the coming years thanks to the already ongoing DESI (DESI Collaboration et al., 2016) and planned 4MOST (de Jong et al., 2019) QSO surveys, which will largely overlap with KiDS.

The random and faint extrapolation tests require an interpretation, which depends on the problem complexity and robustness of the inference at the faint end. When determining the appropriate value of a given model parameter, for example the number of epochs or trees, one might obtain ambiguous results, such as a range of acceptable values rather than one best value. This adds to the complexity of model optimization. The results on faint end extrapolation are reported to have a high impact on the estimation reliability (e.g., Shu et al., 2019; Clarke et al., 2020; Logan and Fotopoulou, 2020). We achieved satisfactory extrapolation results in $r < 23.5$, which is 1.5 magnitude larger than the SDSS limit. Our results are robust, because we not only find a limit at which the results diverge from expectations, but also make sure that the results are adequate for data brighter than this limit, $r < 23.5$ in our case.

The biggest source of incompleteness in our catalog comes from removing objects with at least one band missing out of the nine available. This decreases the size of the KiDS inference data by 55%, from 100 million to 45 million. The requirement of u-band detections may lower the completeness of QSOs at $z \gtrsim 2$. When looking for such high-z QSOs, one would have to perform a classification and redshift estimation using only the redder bands. The possible addition of red QSOs to the training may result in a higher QSO density at high redshifts, at the cost of limiting the feature space.

Another source of incompleteness is the removal of 13 million of the faintest objects for which the SExtractor morphological classifier CLASS_STAR fails. At $r > 23.5$, the unsafe subset constitutes a large fraction of all KiDS objects (65%) and dominates at $r > 24$ (81%) (Fig. 2.6). As the stellarity index is in fact one of the most important features for the classification (Fig. 4.2), its inaccuracy at the faint data end may account for the limit of reliable extrapolation, which is $r < 23.5$.

We plan several steps in order to further increase the catalog's completeness and interpretability. The missing data problem can be solved with either straightforward methods, such as assigning some specific values to the missing features, for example, zeros or mean values, or more sophisticated approaches such as predicting the missing values or using models designed to work with missing features (e.g., Śmieja et al., 2018). It might also prove necessary to skip the shape classifiers for the faint end estimations. The redshift uncertainties require epistemic uncertainty modeling in order to be fully useful in the extrapolation range of $r > 22$. This can be implemented in ANN with, for example, variational layers of Tensorflow, which represent each weight as a probability distribution.

It is possible to validate the faint-end predictions by fitting an SED to the QSO candidates in the catalog, using the estimated photo-zs as input to SED fitting. This will allow us to physically interpret the predictions and find the physical reasons for some of the model failures. Furthermore, this could be the best way of validating the estimations at the faint magnitude end by evaluating how physically acceptable the QSO SED fits are.

Dedicated spectroscopic observations might be yet another way of validating the estimations at extrapolation. They would allow us to determine more precisely the limit of reliability of our predictions at $r \approx 23.5$. It would be interesting to also probe the faintest objects to understand how the estimations cover the unsafe inference subset and find what is the actual portion of real QSOs in our selection in the faintest end. If the results are positive enough, this would show that the ML models optimized for the extrapolation can also serve as a method of candidate selection for follow-up spectroscopy in such faint data.

The correlation analysis might be extended by addressing systematics resulting from the observations and ML processing of the KiDS data, accounting for photo-z errors from the uncertainties obtained the with ANN, using tomographic approach based on the photo-z binning, measuring correlation between the catalog and stars, and, finally, constraining the cosmology itself. In this early study, however, we are mostly interested in providing the most reliable catalog of quasars, tested throughly with different methods.

6.4 Conclusion

In this work we have shown how artificial intelligence can be successfully used to process large amounts of astronomical data. The wide-angle KiDS DR4 catalog of 253k QSO candidates with reliable photometric redshifts can be used in both AGN and LSS studies, as shown here by estimating the bias function, and our work addresses important aspects for any other application of ML in astronomy. As we have demonstrated, well-designed inference models can be pushed to the limits and give reliable results even beyond the coverage of the training sets. The interested readers can test the approach of validation on the faint data proposed in this work in their own inference schemes, and compare what differences it brings to parameter optimization. This work, and ML processing in general, is important in a view of the upcoming large surveys such as the Rubin Observatory LSST or Euclid. Those new endeavors will provide unprecedented vast amounts of data much fainter than the current spectroscopic surveys, and also going deeper than most of the current wide-angle imaging datasets, which will require robust big data processing. Carefully designed, intepretable, and well-tested ML models can provide reliable and trustworthy results. We believe that the framework developed here is one step towards meeting the demands of these future missions.

Bibliography

Abadi, Martín et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Abdurro'uf et al. (Apr. 2022). The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. 259.2, 35, p. 35. DOI: 10.3847/1538-4365/ac4414. arXiv: 2112.02026 [astro-ph.GA].

Abolfathi, B. et al. (Apr. 2018). The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. 235, 42, p. 42. DOI: 10.3847/1538-4365/aa9e8a. arXiv: 1707.09322.

Alonso, David et al. (Apr. 2019). A unified pseudo-C_ℓ framework. 484.3, pp. 4127–4151. DOI: 10.1093/mnras/stz093. arXiv: 1809.09603 [astro-ph.CO].

Alonso, David et al. (Mar. 2021). Cross-correlating radio continuum surveys and CMB lensing: constraining redshift distributions, galaxy bias, and cosmology. 502.1, pp. 876–887. DOI: 10.1093/mnras/stab046. arXiv: 2009.01817 [astro-ph.CO].

Asgari, Marika et al. (July 2020). KiDS-1000 Cosmology: Cosmic shear constraints and comparison between two point statistics. *arXiv e-prints*, arXiv:2007.15633, arXiv:2007.15633. arXiv: 2007.15633 [astro-ph.CO].

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, 26, p. 26. DOI: 10.1088/0004-637X/772/1/26. arXiv: 1209.6055.

Assef, R. J. et al. (Feb. 2018). The WISE AGN Catalog. 234, 23, p. 23. DOI: 10.3847/1538-4365/aaa00a. arXiv: 1706.09901.

Benítez, Narciso (June 2000). Bayesian Photometric Redshift Estimation. 536.2, pp. 571–583. DOI: 10.1086/308947. arXiv: astro-ph/9811189 [astro-ph].

Bertin, E. and S. Arnouts (June 1996). SExtractor: Software for source extraction. 117, pp. 393–404. DOI: 10.1051/aas:1996164.

Bilicki, M. et al. (Aug. 2018). Photometric redshifts for the Kilo-Degree Survey. Machine-learning analysis with artificial neural networks. 616, A69, A69. DOI: 10.1051/0004-6361/201731942. arXiv: 1709.04205.

Bilicki, M. et al. (Sept. 2021). Bright galaxy sample in the Kilo-Degree Survey Data Release 4. Selection, photometric redshifts, and physical properties. 653, A82, A82. DOI: 10.1051/0004-6361/202140352. arXiv: 2101.06010 [astro-ph.GA].

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. Information science and statistics. Softcover published in 2016. New York, NY: Springer. URL: https://cds.cern.ch/record/998831.

Blanton, M. R. et al. (July 2017). Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. 154, 28, p. 28. DOI: 10.3847/1538-3881/aa7567. arXiv: 1703.00052.

Bovy, J. et al. (Mar. 2011). Think Outside the Color Box: Probabilistic Target Selection and the SDSS-XDQSO Quasar Targeting Catalog. 729, 141, p. 141. DOI: 10.1088/0004-637X/729/2/141. arXiv: 1011.6392 [astro-ph.CO].

Bovy, J. et al. (Apr. 2012). Photometric Redshifts and Quasar Probabilities from a Single, Data-driven Generative Model. 749, 41, p. 41. DOI: 10.1088/0004-637X/749/1/41. arXiv: 1105.3975.

Breiman, Leo (Oct. 2001). Random Forests. *Mach. Learn.* 45.1, pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Brescia, M., S. Cavuoti, and G. Longo (July 2015). Automated physical classification in the SDSS DR10. A catalogue of candidate quasars. 450, pp. 3893–3903. DOI: 10.1093/mnras/stv854. arXiv: 1504.03857.

Brescia, M. et al. (Aug. 2013). Photometric Redshifts for Quasars in Multi-band Surveys. 772.2, 140, p. 140. DOI: 10.1088/0004-637X/772/2/140. arXiv: 1305.5641 [astro-ph.IM].

Calistro Rivera, Gabriela et al. (Dec. 2016). AGNfitter: A Bayesian MCMC Approach to Fitting Spectral Energy Distributions of AGNs. 833.1, 98, p. 98. DOI: 10.3847/1538-4357/833/1/98. arXiv: 1606.05648 [astro-ph.GA].

Capaccioli, M. et al. (Oct. 2012). VST: the VLT Survey Telescope. VST: An Overview. *Science from the Next Generation Imaging and Spectroscopic Surveys*, 1, p. 1.

Carrasco, D. et al. (Dec. 2015). Photometric classification of quasars from RCS-2 using Random Forest. 584, A44, A44. DOI: 10.1051/0004-6361/201525752. arXiv: 1405.5298.

Carvajal, Rodrigo et al. (Oct. 2021). Exploring New Redshift Indicators for Radio-Powerful AGN. *Galaxies* 9.4, p. 86. DOI: 10.3390/galaxies9040086. arXiv: 2111.00778 [astro-ph.GA].

Chan, J. H. H. et al. (Mar. 2022). Discovery of strongly lensed quasars in the Ultraviolet Near Infrared Optical Northern Survey (UNIONS). 659, A140, A140. DOI: 10.1051/0004-6361/202142389. arXiv: 2110.09535 [astro-ph.GA].

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: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

Chisari, Nora Elisa et al. (May 2019). Core Cosmology Library: Precision Cosmological Predictions for LSST. 242.1, 2, p. 2. DOI: 10.3847/1538-4365/ab1658. arXiv: 1812.05995 [astro-ph.CO].

Chollet, François (2015). keras. https://github.com/fchollet/keras.

Ciesla, L. et al. (Apr. 2015). Constraining the properties of AGN host galaxies with spectral energy distribution modelling. 576, A10, A10. DOI: 10.1051/0004-6361/201425252. arXiv: 1501.03672 [astro-ph.GA].

Clarke, A. O. et al. (July 2020). Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra. 639, A84, A84. DOI: 10.1051/0004-6361/201936770. arXiv: 1909.10963 [astro-ph.GA].

Croom, S. M. et al. (Apr. 2004). The 2dF QSO Redshift Survey - XII. The spectroscopic catalogue and luminosity function. 349, pp. 1397–1418. DOI: 10.1111/j.1365-2966.2004.07619.x. eprint: astro-ph/0403040.

Croom, S. M. et al. (Jan. 2009). The 2dF-SDSS LRG and QSO Survey: the spectroscopic QSO catalogue. 392, pp. 19–44. DOI: 10.1111/j.1365-2966.2008.14052.x. arXiv: 0810.4955.

Cuoco, A. et al. (Sept. 2017). Tomographic Imaging of the Fermi-LAT γ-Ray Sky through Cross-correlations: A Wider and Deeper Look. 232, 10, p. 10. DOI: 10.3847/1538-4365/aa8553. arXiv: 1709.01940 [astro-ph.HE].

Curran, S. J. (Mar. 2020). QSO photometric redshifts from SDSS, WISE, and GALEX colours. 493.1, pp. L70–L75. DOI: 10.1093/mnrasl/slaa012. arXiv: 2001.06514 [astro-ph.IM].

Cutri, R. M. and et al. (Nov. 2013). VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013). *VizieR Online Data Catalog* 2328.

Dawson, K. S. et al. (Jan. 2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. 145, 10, p. 10. DOI: 10.1088/0004-6256/145/1/10. arXiv: 1208.0022.

de Jong, J. T. A. et al. (Dec. 2013). The Kilo-Degree Survey. *The Messenger* 154, pp. 44–46.

de Jong, J. T. A. et al. (Oct. 2015). The first and second data releases of the Kilo-Degree Survey. 582, A62, A62. DOI: 10.1051/0004-6361/201526601. arXiv: 1507.00742.

de Jong, Jelte T. A. et al. (Aug. 2017). The third data release of the Kilo-Degree Survey and associated data products. 604, A134, A134. DOI: 10.1051/0004-6361/201730747.

de Jong, R. S. et al. (Mar. 2019). 4MOST: Project overview and information for the First Call for Proposals. *The Messenger* 175, pp. 3–11. DOI: 10.18727/0722-6691/5117. arXiv: 1903.02464 [astro-ph.IM].

DESI Collaboration et al. (Oct. 2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. *ArXiv e-prints*. arXiv: 1611.00036 [astro-ph.IM].

DiPompeo, M. A., R. C. Hickox, and A. D. Myers (Feb. 2016). Updated measurements of the dark matter halo masses of obscured quasars with improved WISE and Planck data. 456, pp. 924–942. DOI: 10.1093/mnras/stv2681. arXiv: 1511.04469.

DiPompeo, M. A. et al. (Aug. 2014). The angular clustering of infrared-selected obscured and unobscured quasars. 442, pp. 3443–3453. DOI: 10.1093/mnras/stu1115. arXiv: 1406.0778.

DiPompeo, M. A. et al. (Sept. 2015). Quasar probabilities and redshifts from WISE mid-IR through GALEX UV photometry. 452, pp. 3124–3138. DOI: 10.1093/mnras/stv1562. arXiv: 1507.02884.

DiPompeo, M. A. et al. (Aug. 2017). The characteristic halo masses of half-a-million WISE-selected quasars. 469, pp. 4630–4643. DOI: 10.1093/mnras/stx1215. arXiv: 1705.05306.

D'Isanto, A. et al. (Aug. 2018). Return of the features. Efficient feature selection and interpretation for photometric redshifts. 616, A97, A97. DOI: 10.1051/0004-6361/201833103. arXiv: 1803.10032 [astro-ph.IM].

Edelson, R. and M. Malkan (May 2012). Reliable Identifications of Active Galactic Nuclei from the WISE, 2MASS, and ROSAT All-Sky Surveys. 751, 52, p. 52. DOI: 10.1088/0004-637X/751/1/52. arXiv: 1203.1942.

Edge, A. et al. (Dec. 2013). The VISTA Kilo-degree Infrared Galaxy (VIKING) Survey: Bridging the Gap between Low and High Redshift. *The Messenger* 154, pp. 32–34.

Eftekharzadeh, S. et al. (Nov. 2015). Clustering of intermediate redshift quasars using the final SDSS III-BOSS sample. 453, pp. 2779–2798. DOI: 10.1093/mnras/stv1763. arXiv: 1507.08380.

Falocco, S., F. J. Carrera, and J. Larsson (Feb. 2022). Automated algorithms to build active galactic nucleus classifiers. 510.1, pp. 161–176. DOI: 10.1093/mnras/stab3435. arXiv: 2111.12369 [astro-ph.GA].

Fan, Xiaohui (Nov. 2006). Evolution of high-redshift quasars. 50.9-10, pp. 665–671. DOI: 10.1016/j.newar.2006.06.077.

Foreman-Mackey, Daniel et al. (Mar. 2013). emcee: The MCMC Hammer. 125.925, p. 306. DOI: 10.1086/670067. arXiv: 1202.3665 [astro-ph.IM].

Fotopoulou, S. and S. Paltani (Oct. 2018). CPz: Classification-aided photometric-redshift estimation. 619, A14, A14. DOI: 10.1051/0004-6361/201730763. arXiv: 1808.04977.

Fotopoulou, S. et al. (June 2016). The XXL Survey. VI. The 1000 brightest X-ray point sources. 592, A5, A5. DOI: 10.1051/0004-6361/201527402. arXiv: 1603.03240 [astro-ph.GA].

Gaia Collaboration et al. (Nov. 2016). The Gaia mission. 595, A1, A1. DOI: 10.1051/0004-6361/201629272. arXiv: 1609.04153 [astro-ph.IM].

Gaia Collaboration et al. (Aug. 2018a). Gaia Data Release 2. Summary of the contents and survey properties. 616, A1, A1. DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365.

Gaia Collaboration et al. (Aug. 2018b). Gaia Data Release 2. The celestial reference frame (Gaia-CRF2). 616, A14, A14. DOI: 10.1051/0004-6361/201832916.

Górski, K. M. et al. (Apr. 2005). HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. 622.2, pp. 759–771. DOI: 10.1086/427976. arXiv: astro-ph/0409513 [astro-ph].

Harrell, Frank (Jan. 2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. *Chapter 5: Resampling, Validating, and Simplifying the Model* 3, pp. 88–103.

Hausen, Ryan and Brant E. Robertson (May 2020). Morpheus: A Deep Learning Framework for the Pixel-level Analysis of Astronomical Image Data. 248.1, 20, p. 20. DOI: 10.3847/1538-4365/ab8868. arXiv: 1906.11248 [astro-ph.GA].

Haykin, Simon (1998). *Neural Networks: A Comprehensive Foundation*. 2nd. Upper Saddle River, NJ, USA: Prentice Hall PTR. ISBN: 0132733501.

Heintz, K. E. et al. (July 2018). A quasar hiding behind two dusty absorbers. Quantifying the selection bias of metal-rich, damped Lyα absorption systems. 615, A43, A43. DOI: 10.1051/0004-6361/201731964. arXiv: 1803.09805.

Heymans, Catherine et al. (July 2020). KiDS-1000 Cosmology: Multi-probe weak gravitational lensing and spectroscopic galaxy clustering constraints. *arXiv e-prints*, arXiv:2007.15632, arXiv:2007.15632. arXiv: 2007.15632 [astro-ph.CO].

Hildebrandt, H. et al. (Jan. 2020). KiDS+VIKING-450: Cosmic shear tomography with optical and infrared data. 633, A69, A69. DOI: 10.1051/0004-6361/201834878. arXiv: 1812.06076 [astro-ph.CO].

Hinshaw, G. et al. (Oct. 2013). Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. 208.2, 19, p. 19. DOI: 10.1088/0067-0049/208/2/19. arXiv: 1212.5226 [astro-ph.CO].

Hivon, Eric et al. (Mar. 2002). MASTER of the Cosmic Microwave Background Anisotropy Power Spectrum: A Fast Method for Statistical Analysis of Large and Complex Cosmic Microwave Background Data Sets. 567.1, pp. 2–17. DOI: 10.1086/338126. arXiv: astro-ph/0105302 [astro-ph].

Ho, S. et al. (May 2015). Sloan Digital Sky Survey III photometric quasar clustering: probing the initial conditions of the Universe. 5, 040, p. 040. DOI: 10.1088/1475-7516/2015/05/040. arXiv: 1311.2597.

Ivezić, Željko et al. (Mar. 2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. 873.2, 111, p. 111. DOI: 10.3847/1538-4357/ab042c. arXiv: 0805.2366 [astro-ph].

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

Jarrett, T. H. et al. (Feb. 2017). Galaxy and Mass Assembly (GAMA): Exploring the WISE Web in G12. 836, 182, p. 182. DOI: 10.3847/1538-4357/836/2/182. arXiv: 1607.01190.

Joudaki, Shahab et al. (2017). KiDS-450: testing extensions to the standard cosmological model. 471.2, pp. 1259–1279. DOI: 10.1093/mnras/stx998. arXiv: 1610.04606 [astro-ph.CO].

Kauffmann, G. et al. (Dec. 2003). The host galaxies of active galactic nuclei. 346, pp. 1055–1077. DOI: 10.1111/j.1365-2966.2003.07154.x. eprint: astro-ph/0304239.

Kewley, L. J. et al. (Sept. 2013). The Cosmic BPT Diagram: Confronting Theory with Observations. 774, L10, p. L10. DOI: 10.1088/2041-8205/774/1/L10. arXiv: 1307.0514.

Khramtsov, Vladislav et al. (Dec. 2019). KiDS-SQuaD. II. Machine learning selection of bright extragalactic objects to search for new gravitationally lensed quasars. 632, A56, A56. DOI: 10.1051/0004-6361/201936006. arXiv: 1906.01638 [astro-ph.GA].

Kohonen, Teuvo, ed. (1997). *Self-organizing Maps*. Berlin, Heidelberg: Springer-Verlag. ISBN: 3-540-62017-6.

Kormendy, J. and L. C. Ho (Aug. 2013). Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies. 51, pp. 511–653. DOI: 10.1146/annurev-astro-082708-101811. arXiv: 1304.7762.

Kuijken, K. (May 2008). GaaP: PSF- and aperture-matched photometry using shapelets. 482, pp. 1053–1067. DOI: 10.1051/0004-6361:20066601. eprint: astro-ph/0610606.

— (Dec. 2011). OmegaCAM: ESO's Newest Imager. *The Messenger* 146, pp. 8–11.

Kuijken, K. et al. (Dec. 2015). Gravitational lensing analysis of the Kilo-Degree Survey. 454, pp. 3500–3532. DOI: 10.1093/mnras/stv2140. arXiv: 1507.00738.

Kuijken, K. et al. (2019). The fourth data release of the Kilo-Degree Survey: ugri imaging and nine-band optical-IR photometry over 1000 square degrees. 625, A2, A2. DOI: 10.1051/0004-6361/201834918. arXiv: 1902.11265 [astro-ph.GA].

Kurcz, A. et al. (July 2016). Towards automatic classification of all WISE sources. 592, A25, A25. DOI: 10.1051/0004-6361/201628142. arXiv: 1604.04229.

Laurent, P. et al. (July 2017). Clustering of quasars in SDSS-IV eBOSS: study of potential systematics and bias determination. 7, 017, p. 017. DOI: 10.1088/1475-7516/2017/07/017. arXiv: 1705.04718.

Leistedt, B., H. V. Peiris, and N. Roth (Nov. 2014). Constraints on Primordial Non-Gaussianity from 800 000 Photometric Quasars. *Physical Review Letters* 113.22, 221301, p. 221301. DOI: 10.1103/PhysRevLett.113.221301. arXiv: 1405.4315.

Lewis, Antony, Anthony Challinor, and Anthony Lasenby (Aug. 2000). Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models. 538.2, pp. 473–476. DOI: 10.1086/309179. arXiv: astro-ph/9911177 [astro-ph].

Li, Rui et al. (May 2022). Galaxy morphoto-Z with neural Networks (GaZNets). I. Optimized accuracy and outlier fraction from Imaging and Photometry. *arXiv e-prints*, arXiv:2205.10720, arXiv:2205.10720. arXiv: 2205.10720 [astro-ph.GA].

Limber, D. Nelson (May 1954). The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field. II. 119, p. 655. DOI: 10.1086/145870.

Lindegren, L. et al. (Aug. 2018). Gaia Data Release 2. The astrometric solution. 616, A2, A2. DOI: 10.1051/0004-6361/201832727. arXiv: 1804.09366 [astro-ph.IM].

Logan, C. H. A. and S. Fotopoulou (Jan. 2020). Unsupervised star, galaxy, QSO classification. Application of HDBSCAN. 633, A154, A154. DOI: 10.1051/0004-6361/201936648. arXiv: 1911.05107 [astro-ph.GA].

Lyke, Brad W. et al. (Sept. 2020). The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. 250.1, 8, p. 8. DOI: 10.3847/1538-4365/aba623. arXiv: 2007.09001 [astro-ph.GA].

Maaten, Laurens van der and Geoffrey Hinton (Nov. 2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, pp. 2579–2605. URL: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Maddox, N. et al. (May 2008). Luminous K-band selected quasars from UKIDSS. 386, pp. 1605–1624. DOI: 10.1111/j.1365-2966.2008.13138.x. arXiv: 0802.3650.

Małek, Katarzyna et al. (Jan. 2020). HELP project - a dreamed-of multiwavelength dataset for SED fitting: The influence of used models for the main physical properties of galaxies. *IAU Symposium*. Ed. by Médéric Boquien et al. Vol. 341. IAU Symposium, pp. 39–43. DOI: 10.1017/S174392131900471X. arXiv: 1904.12498 [astro-ph.GA].

Masters, D. et al. (Nov. 2015). Mapping the Galaxy Color-Redshift Relation: Optimal Photometric Redshift Calibration Strategies for Cosmology Surveys. 813, 53, p. 53. DOI: 10.1088/0004-637X/813/1/53. arXiv: 1509.03318.

McInnes, Leland, John Healy, and Steve Astels (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2.11. DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105%2Fjoss.00205.

Merloni, A. et al. (Mar. 2019). 4MOST Consortium Survey 6: Active Galactic Nuclei. *The Messenger* 175, pp. 42–45. DOI: 10.18727/0722-6691/5125. arXiv: 1903.02472 [astro-ph.GA].

Nakoneczny, S. et al. (Apr. 2019). Catalog of quasars from the Kilo-Degree Survey Data Release 3. 624, A13, A13. DOI: 10.1051/0004-6361/201834794. arXiv: 1812.03084 [astro-ph.IM].

Nakoneczny, S. J. et al. (May 2021). Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4. 649, A81, A81. DOI: 10.1051/0004-6361/202039684. arXiv: 2010.13857 [astro-ph.CO].

Oogi, Taira et al. (Feb. 2016). Quasar clustering in a galaxy and quasar formation model based on ultra high-resolution N-body simulations. 456.1, pp. L30–L34. DOI: 10.1093/mnrasl/slv169. arXiv: 1512.00458 [astro-ph.GA].

Palanque-Delabrouille, N. et al. (Mar. 2016). The extended Baryon Oscillation Spectroscopic Survey: Variability selection and quasar luminosity function. 587, A41, A41. DOI: 10.1051/0004-6361/201527392. arXiv: 1509.05607 [astro-ph.CO].

Pasquet-Itam, J. and J. Pasquet (Apr. 2018). Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey stripe 82. 611, A97, A97. DOI: 10.1051/0004-6361/201731106. arXiv: 1712.02777 [astro-ph.IM].

Pearson, Karl (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. DOI: 10.1080/14786440109462720.

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

Peebles, P. J. E. (Oct. 1973). Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory. 185, pp. 413–440. DOI: 10.1086/152431.

Piramuthu, Selwyn and Riyaz T. Sikora (Mar. 2009). Iterative Feature Construction for Improving Inductive Learning Algorithms. *Expert Syst. Appl.* 36.2, pp. 3401–3406. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2008.02.010. URL: http://dx.doi.org/10.1016/j.eswa.2008.02.010.

Planck Collaboration et al. (Sept. 2020a). Planck 2018 results. VI. Cosmological parameters. 641, A6, A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

Planck Collaboration et al. (Sept. 2020b). Planck 2018 results. VIII. Gravitational lensing. 641, A8, A8. DOI: 10.1051/0004-6361/201833886. arXiv: 1807.06210 [astro-ph.CO].

Prokhorenkova, Liudmila et al. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 6638–6648. URL: http://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf.

Richard, J. et al. (Mar. 2019). 4MOST Consortium Survey 8: Cosmology Redshift Survey (CRS). *The Messenger* 175, pp. 50–53. DOI: 10.18727/0722-6691/5127. arXiv: 1903.02474 [astro-ph.CO].

Richards, G. T. et al. (June 2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. 123, pp. 2945–2975. DOI: 10.1086/340187. eprint: astro-ph/0202251.

Richards, G. T. et al. (Dec. 2004). Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey: 100,000 z 3 Quasars from Data Release One. 155, pp. 257–269. DOI: 10.1086/425356. eprint: astro-ph/0408505.

Richards, G. T. et al. (Jan. 2009a). Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ˜1,000,000 Quasars from Data Release 6. 180, pp. 67–83. DOI: 10.1088/0067-0049/180/1/67. arXiv: 0809.3952.

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

Richards, G. T. et al. (Aug. 2015). Bayesian High-redshift Quasar Classification from Optical and Mid-IR Photometry. 219, 39, p. 39. DOI: 10.1088/0067-0049/219/2/39. arXiv: 1507.07788.

Salvato, M. et al. (Jan. 2009). Photometric Redshift and Classification for the XMM-COSMOS Sources. 690.2, pp. 1250–1263. DOI: 10.1088/0004-637X/690/2/1250. arXiv: 0809.2098 [astro-ph].

Salvato, M. et al. (Dec. 2011). Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples. 742.2, 61, p. 61. DOI: 10.1088/0004-637X/742/2/61. arXiv: 1108.6061 [astro-ph.CO].

Scranton, R. et al. (Nov. 2005). Detection of Cosmic Magnification with the Sloan Digital Sky Survey. 633, pp. 589–602. DOI: 10.1086/431358. eprint: astro-ph/0504510.

Secrest, N. J. et al. (Nov. 2015). Identification of 1.4 Million Active Galactic Nuclei in the Mid-Infrared using WISE Data. 221.1, 12, p. 12. DOI: 10.1088/0067-0049/221/1/12. arXiv: 1509.07289 [astro-ph.GA].

Shen, Yue et al. (June 2009). Quasar Clustering from SDSS DR5: Dependences on Physical Properties. 697.2, pp. 1656–1673. DOI: 10.1088/0004-637X/697/2/1656. arXiv: 0810.4144 [astro-ph].

Sherwin, B. D. et al. (Oct. 2012). The Atacama Cosmology Telescope: Cross-correlation of cosmic microwave background lensing and quasars. 86.8, 083006, p. 083006. DOI: 10.1103/PhysRevD.86.083006. arXiv: 1207.4543 [astro-ph.CO].

Shu, Yiping et al. (Nov. 2019). Catalogues of active galactic nuclei from Gaia and unWISE data. 489.4, pp. 4741–4759. DOI: 10.1093/mnras/stz2487. arXiv: 1909.02010 [astro-ph.GA].

Śmieja, Marek et al. (2018). Processing of missing data by neural networks. *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 2719–2729. URL: http://papers.nips.cc/paper/7537-processing-of-missing-data-by-neural-networks.pdf.

Spiniello, C. et al. (Oct. 2018). KiDS-SQuaD: The KiDS Strongly lensed Quasar Detection project. 480, pp. 1163–1173. DOI: 10.1093/mnras/sty1923. arXiv: 1805.12436.

Stalevski, Marko et al. (May 2016). The dust covering factor in active galactic nuclei. 458.3, pp. 2288–2302. DOI: 10.1093/mnras/stw444. arXiv: 1602.06954 [astro-ph.GA].

Stern, D. 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, 30, p. 30. DOI: 10.1088/0004-637X/753/1/30. arXiv: 1205.0811.

Stölzner, B. et al. (Mar. 2018). Updated tomographic analysis of the integrated Sachs-Wolfe effect and implications for dark energy. 97.6, 063506, p. 063506. DOI: 10.1103/PhysRevD.97.063506. arXiv: 1710.03238.

Strauss, M. A. et al. (Sept. 2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. 124, pp. 1810–1824. DOI: 10.1086/342343. eprint: astro-ph/0206225.

Stringer, K. M. et al. (Apr. 2021). Identifying RR Lyrae Variable Stars in Six Years of the Dark Energy Survey. 911.2, 109, p. 109. DOI: 10.3847/1538-4357/abe873. arXiv: 2011.13930 [astro-ph.GA].

Takahashi, Ryuichi et al. (Dec. 2012). Revising the Halofit Model for the Nonlinear Matter Power Spectrum. 761.2, 152, p. 152. DOI: 10.1088/0004-637X/761/2/152. arXiv: 1208.2701 [astro-ph.CO].

Taylor, M. B. (Dec. 2005). TOPCAT STIL: Starlink Table/VOTable Processing Software. *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell, M. Britton, and R. Ebert. Vol. 347. Astronomical Society of the Pacific Conference Series, p. 29.

van Uitert, Edo et al. (2018). KiDS+GAMA: cosmology constraints from a joint analysis of cosmic shear, galaxy-galaxy lensing, and angular clustering. 476.4, pp. 4662–4689. DOI: 10.1093/mnras/sty551. arXiv: 1706.05004 [astro-ph.CO].

Venemans, B. P. et al. (Nov. 2015). First discoveries of z
6 quasars with the Kilo-Degree Survey and VISTA Kilo-Degree Infrared Galaxy survey. 453, pp. 2259–2266. DOI: 10.1093/mnras/stv1774. arXiv: 1507.00726.

Warren, S. J., P. C. Hewett, and C. B. Foltz (Mar. 2000). The KX method for producing K-band flux-limited samples of quasars. 312, pp. 827–832. DOI: 10.1046/j.1365-8711.2000.03206.x. eprint: astro-ph/9911064.

Wright, Angus H. et al. (Aug. 2020). KiDS+VIKING-450: Improved cosmological parameter constraints from redshift calibration with self-organising maps. 640, L14, p. L14. DOI: 10.1051/0004-6361/202038389. arXiv: 2005.04207 [astro-ph.CO].

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].

Wu, X.-B. et al. (Aug. 2012). SDSS Quasars in the WISE Preliminary Data Release and Quasar Candidate Selection with Optical/Infrared Colors. 144, 49, p. 49. DOI: 10.1088/0004-6256/144/2/49. arXiv: 1204.6197.

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].

Yang, Qian et al. (Dec. 2017). Quasar Photometric Redshifts and Candidate Selection: A New Algorithm Based on Optical and Mid-infrared Photometric Data. 154.6, 269, p. 269. DOI: 10.3847/1538-3881/aa943c. arXiv: 1710.09155 [astro-ph.GA].

Yèche, C. et al. (Nov. 2010). Artificial neural networks for quasar selection and photometric redshift determination. 523, A14, A14. DOI: 10.1051/0004-6361/200913508.

York, D. G. et al. (Sept. 2000). The Sloan Digital Sky Survey: Technical Summary. 120, pp. 1579–1587. DOI: 10.1086/301513. eprint: astro-ph/0006396.