Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4

Nakoneczny, S.J.; Bilicki, M.; Pollo, A.; Asgari, M.; Dvornik, A.; Erben, T.; ... ; Valentijn, E.

Astronomy & Astrophysics manuscript no. kids_qso_dr4
October 29, 2021

Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4

S.J. Nakoneczny[1], M. Bilicki[2], A. Pollo[1,3], M. Asgari[4], A. Dvornik[5], T. Erben[6], B. Giblin[4], C. Heymans[4,5], H. Hildebrandt[5], A. Kannawadi[7], K. Kuijken[8], N.R. Napolitano[9], and E. Valentijn[10]

[1] National Centre for Nuclear Research, Astrophysics Division, ul. Pasteura 7, 02-093 Warsaw, Poland
[2] Center for Theoretical Physics, Polish Academy of Sciences, al. Lotników 32/46, 02-668, Warsaw, Poland
[3] Astronomical Observatory of the Jagiellonian University, 31-007 Kraków, Poland
[4] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, U.K.
[5] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, 44780 Bochum, Germany
[6] Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn / Germany
[7] Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
[8] Leiden Observatory, Leiden University, P.O.Box 9513, 2300RA Leiden, The Netherlands
[9] School of Physics and Astronomy, Sun Yat-sen University, Guangzhou 519082, Zhuhai Campus, P.R. China
[10] Kapteyn Institute, University of Groningen, PO Box 800, NL 9700 AV Groningen

October 29, 2021

ABSTRACT

We present a catalog of quasars with their corresponding redshifts derived from the photometric Kilo-Degree Survey (KiDS) Data Release 4. We achieved it by training machine learning (ML) models using optical *ugri* and near-infrared $ZYJHK_s$ bands, on objects known from 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, and employing its visualizations. 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 allows us to properly fit the bias vs. variance trade-off. We test 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 is the best for combined classification and redshift. The ANN inference results are tested using number counts, Gaia parallaxes and other quasar catalogs external to the training set. Based on these tests, we derive the minimum classification probability for quasar candidates which provides the best purity vs. 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 further 185,000 in the reliable extrapolation regime ($22 < r < 23.5$). Test-data purity equals 97%, completeness is 94%, the latter dropping by 3% in the extrapolation to data fainter by one magnitude than the training set. The photometric redshifts are derived with ANN and modeled with Gaussian uncertainties. 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 redshift range $0.14 < z < 3.63$ (1st 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 cosmological and Active Galactic Nucleus (AGN) analysis. We publicly release the catalog at kids.strw.leidenuniv.nl/DR4/quasarcatalog.php, and the code at: github.com/snakoneczny/kids-quasars.

Key words. quasars: general – large-scale structure of Universe – methods: data analysis – methods: observational – catalogues – surveys

1. Introduction

Object type and redshift or radial velocity are basic observables in astronomy. They can be precisely determined based on emission and absorption lines from spectroscopy, but 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 high number density and completeness, and samples of millions of objects. Upcoming large photometric surveys, e.g. 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.

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 quasars up to very high redshifts because of accretion of matter on supermassive black holes (Kormendy & Ho 2013), which leads to radiation of enormous amounts of energy. 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 et al. 2016), which makes them highly biased tracers of the LSS (DiPompeo et al. 2014; Laurent et al. 2017). Possible ap-

Send offprint requests to: S.J. Nakoneczny, e-mail: szymon.nakoneczny@ncbj.gov.pl.

plications of quasars in cosmology include tomographic angular clustering (Leistedt et al. 2014; Ho et al. 2015), analysis of cosmic magnification (Scranton et al. 2005), measurement of halo masses (DiPompeo et al. 2017), cross-correlations with various cosmological backgrounds (Sherwin et al. 2012; Cuoco et al. 2017; Stölzner et al. 2018), and even calibration of the reference frames for Galactic studies (Lindegren et al. 2018).

At any cosmic epoch, quasars are sparsely distributed in comparison to inactive galaxies. Therefore, wide-angle surveys are essential to obtain catalogs containing sufficiently many quasars 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 quasars. In spectroscopy, quasar detection and redshift measurement are based on broad emission lines like [OIII]λ5007/Hβ, [NII]λ6584/Hα (Kauffmann et al. 2003; Kewley et al. 2013). Many surveys exploit this approach: 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).

Spectral energy distribution (SED) fitting is a standard approach to analyse photometry of galaxies with active galactic nuclei (AGN), which include quasars in particular. It allows to derive the 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, 2011; Fotopoulou et al. 2016; Fotopoulou & Paltani 2018). The quasar selection in photometry is commonly based on color-color cuts (Warren et al. 2000; Maddox et al. 2008; Edelson & 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 quasar selection are the probabilistic methods (Richards et al. 2004, 2009b,a; Bovy et al. 2011, 2012; DiPompeo et al. 2015; Richards et al. 2015), while machine learning (ML) has been gaining on popularity in this respect as well (Brescia et al. 2015; Carrasco et al. 2015; Kurcz et al. 2016; Nakoneczny et al. 2019; Logan & Fotopoulou 2020). ML models have been also applied to derive quasar photometric redshifts (photo-zs, Brescia et al. 2013; Yang et al. 2017; Pasquet-Itam & Pasquet 2018; Curran 2020).

In the context of the Kilo-Degree Survey (KiDS, de Jong et al. 2013), which will be our focus in this paper, the quasar-related studies have so far dealt with high-redshift ($z \sim 6$) QSOs (Venemans et al. 2015), heavily reddened ones (Heintz et al. 2018), selecting them to search for strong-lensing systems (Spiniello et al. 2018; Khramtsov et al. 2019), while in Nakoneczny et al. (2019, hereafter N19) we presented an ML quasar detection analysis in KiDS Data Release 3 (DR3, de Jong et al. 2017). We note that in general, every quasar present in KiDS multi-band catalogs will have 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 usually will not be 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 quasars.

In our previous work (N19) we performed classification in KiDS DR3, using optical *ugri* broad-band data. The random forest (RF) achieved QSO purity of 91%, and completeness of 87%. The failures in quasar classification – mislabeling them as stars – occurred mostly at QSO redshift $2 < z < 3$. Due to the magnitude limit of training data available from SDSS, we restricted the catalog to $r < 22$. This resulted in 190,000 quasar candidates selected from 3.4 million objects taken as the inference data from KiDS DR3 based on four broad-band detections and data quality considerations.

In this paper, we perform classification and redshift estimation using optical and near-infrared broad-bands of KiDS DR4 (Kuijken et al. 2019), which incorporates the partner VISTA Kilo-degree Infrared Galaxy (VIKING, Edge et al. 2013) measurements. Our main goal is to create a catalog of quasars, optimized for the highest purity and completeness, with robust photometric redshift estimates. We test what near-IR imaging brings to classification in terms of separating quasars from stars. We aim at fitting ML models for the best bias vs. 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, 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 e.g. 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 quasar photometric redshift uncertainties with a Gaussian output layer in an Artificial Neural Network (ANN). Last but not least, we check whether projection of high-dimensional space on two dimensions (2D) can substitute the standard color-color plots as a tool to inspect the feature space coverage and differences between spectroscopic and photometric results, and to meaningfully interpret the data.

The paper is organised as follows: in Section 2 we describe the data and the methodology for quasar selection, redshift estimation, extrapolation in the feature space and bias-variance tuning; in Section 3 we provide results of experiments done on a cross-match with spectroscopic data, properties of the final catalog and purity-completeness calibration; in Section 4 we discuss the main findings, strengths and weaknesses of the approach, and outline possible extensions.

Where relevant, 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 $\Omega_m = 0.287$.

2. Data and methodology

2.1. Data

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 current 4th data release (Kuijken et al. 2019) is the penultimate one, covers a total of 1006 deg^2 and provides a list of $\sim$ 100 million (100M) objects based on the *r*-band detections. It includes also $ZYJHK_s$ photometry from the partner VIKING. The mean limiting AB magnitude (5 σ in a

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

2 arcsec. aperture) of KiDS is $\sim$ 25 in the r band. The optical depth, wide sky coverage, and multiwavelength imaging make this survey an ideal resource for quasar science.

Reliably applied supervised machine learning requires data from which it can learn a solution to the problem, and assurance that the inference data is well represented by its training subset. We solve the problem of QSO detection with classification models, and derive photo-zs with regression models. We create one feature set for both classification and regression to keep the models consistent in predictions. We limit 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 3.1). This feature set includes 9 magnitudes derived with the Gaussian Aperture and PSF (GAaP) photometry method (Kuijken 2008), 36 colors and 36 ratios of every magnitude pair, and two morphological classifiers: SExtractor-based CLASS_STAR (called *stellarity index* here; Bertin & Arnouts 1996), and the 3rd bit of SG2DPHOT – KiDS star/galaxy separation flag based on source r-band morphology[2] (de Jong et al. 2015, 2017). In Section 3.1 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 $\sim$ 45M, which creates the inference set.

The training set is derived from cross-matching the inference data with the Sloan Digital Sky Survey DR14 (SDSS, Abolfathi et al. 2018) spectroscopic observations[3]. The SDSS survey provides three basic classes: galaxies, quasars, and stars, which we use to define a three-class classification problem. After removing objects flagged with warnings by SDSS, we obtain a training subset of 152k objects (69% galaxies, 11% quasars, 20% stars). The training set is 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. The results of machine learning predictions for $r \gtrsim 22$ may be incorrect due to the resulting extrapolation in the feature space.

2.2. Inference subsets

In this Section we define the inference subsets 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 deterioration of results or a complete failure, due to new combinations of features or completely new feature values. For continuous features, such as magnitude, 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 use 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 large values indicating point-like objects, small values corresponding to extended objects, and intermediate values pointing to classifier failure. Because the failures are almost not present in the bright training data, it is not possible for ML models to understand their meaning (Fig. 1). 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. The second morphological classifier we use, SG2DPHOT, is a discrete one, whose 1st and 3rd bits indicate stars, and its failure is indicated by the zero value, the same as for galaxies. We find empirically that using only its 3rd bit provides the best improvement in our results. Cleaning the uncertain stellarity index values removes most of the SG2DPHOT failures.

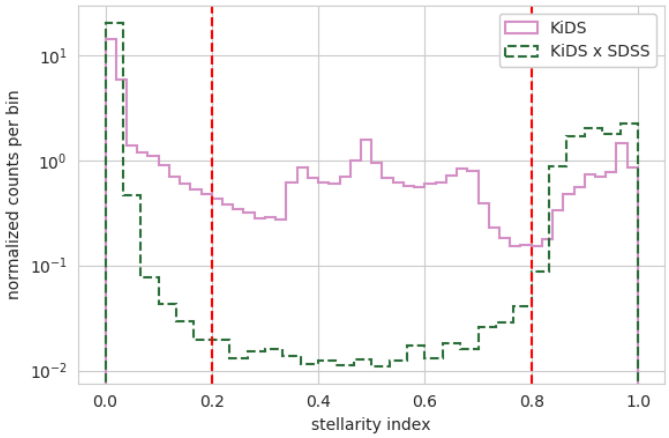

Fig. 1. Normalized histograms of the CLASS_STAR stellarity index in the training (KiDS x SDSS) and inference (KiDS) datasets. The intermediate values represent failures of the morphological classifier. Those values are not commonly present in the training data, thus we cannot expect the ML models to work correctly for objects with such index values. We consider the sources in between the red dashed lines as unsafe for the inference.

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:

1. safe: $r < 22$ **and** stellarity index $\notin (0.2, 0.8)$,
2. extrapolation: $r \in (22, 25)$ **and** stellarity index $\notin (0.2, 0.8)$
3. unsafe: $r > 25$ **or** stellarity index $\in (0.2, 0.8)$

We visualize the KiDS feature space and the inference subsets with t-distributed Stochastic Neighbor Embedding (t-SNE, van der Maaten & Hinton 2008) in Fig. 2. t-SNE belongs to a family of manifold learning algorithms, and allows us to visualize high dimensional and non-linear data structures with much simpler two dimensional embeddings. We create the visualization with the same set of 83 features that are used in classification and redshift estimation. Due to the computational complexity of t-SNE, we take 8k random objects from KiDS data and merge them with 4k 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; we observed the same result previously in N19, where we matched only the *ugri* magnitudes, colors and ratios. The main star and quasar groups are not separated in the KiDS photometric data.

[2] 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.

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

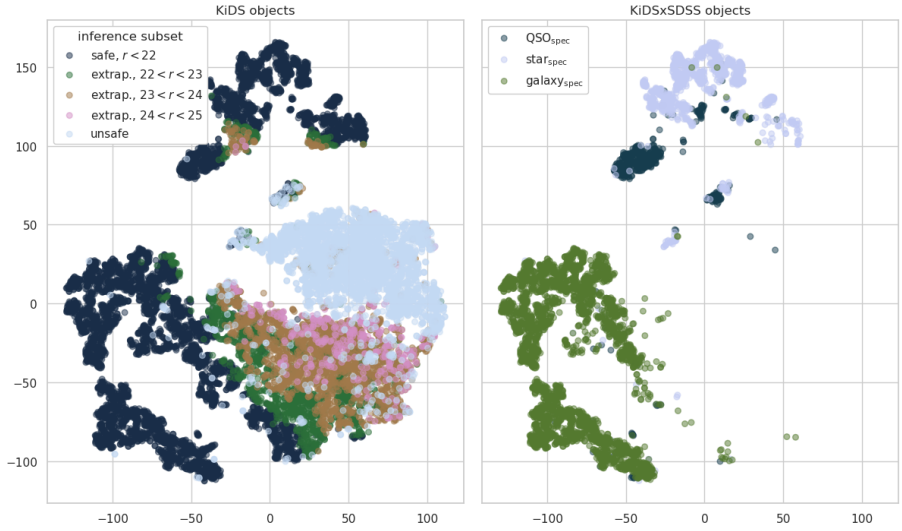

Fig. 2. t-SNE projections. *Left:* inference subsets, *right:* SDSS spectroscopic classification. The visualizations were made on subsets of 12k objects. The real density of objects at any part of the feature space is 3.8k times higher than visualized. We can see three main groups. The point-like objects 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 covers only the bright part of the photometric data; this visualizes the extrapolation problem to address with machine learning. The results of the inference are later investigated on similar plots (Section 3.3), which we consider a more robust approach than investigating color-color diagrams.

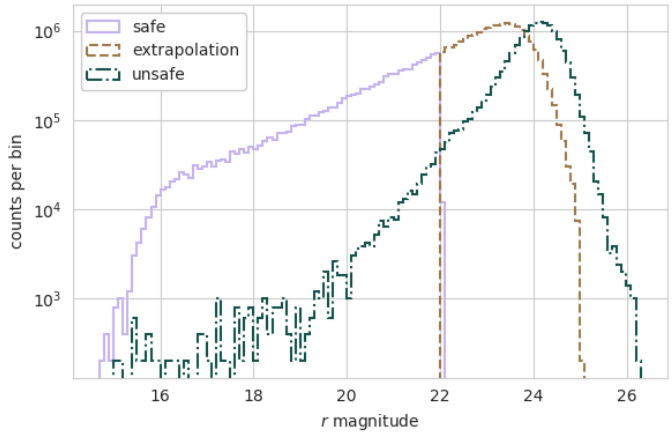

Fig. 3. Distribution of inference subsets over the r magnitude. The limit of the SDSS training data, $r = 22$, defines the lower limit of the extrapolation subset. The morphological classifier failure and sources beyond survey depth ($r > 25$) provide the unsafe subset (fig. 1). The extrapolation subset is complete up to $r < 23.5$. The safe subset covers 21% of data, extrapolation 45%, and unsafe 34%.

The first part of the extrapolation subset at $r \in (22, 23)$ is located close to the training data, and 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 quasar and star groups at $23 < r < 24$, thus such objects have a lower chance for their classification predictions to be correct.

We use 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 do 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 non-linear data structures and allows us to understand the models as well as, or better than, it would be possible with the color-color plots. Additionally, we use the embeddings to have insight into the extrapolation part of the feature space, which cannot be tested with methods based on ground-truth data.

Figure 3 shows r magnitude distributions for the inference subsets. The safe subset is 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 at our catalog. We expect that the number counts of quasars identified using the currently available training sets would become incomplete at $r > 23.5$.

2.3. Validation procedure

Validation data has to differ from the training set, to ensure proper model generalization and avoid overfitting. A randomly chosen sample of data which densely covers the feature space might not fully show the overfitting effects, and this might have very negative influence on the inference at the faint data end, both for classification and photo-zs. We use additional spectroscopic surveys to introduce some differences from the training data, and test the final predictions (Section 3.5). During the experiments, we use 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 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, velocities of stars do not correlate with photometry and we would not observe any variation of star colors between the training and validation data. As magnitude correlates with redshift in case of quasars and galaxies, we expect the faint test to evaluate the extrapolation accuracy of ML models with respect to the estimated redshift values. Figure 4 explains the whole methodology, in blocks illustrating experiments, inference and catalog testing.

Table 1 summarizes the training and validation sets. We select 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 allow us to correctly tune the models for 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 quasars, 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 $r > 22$.

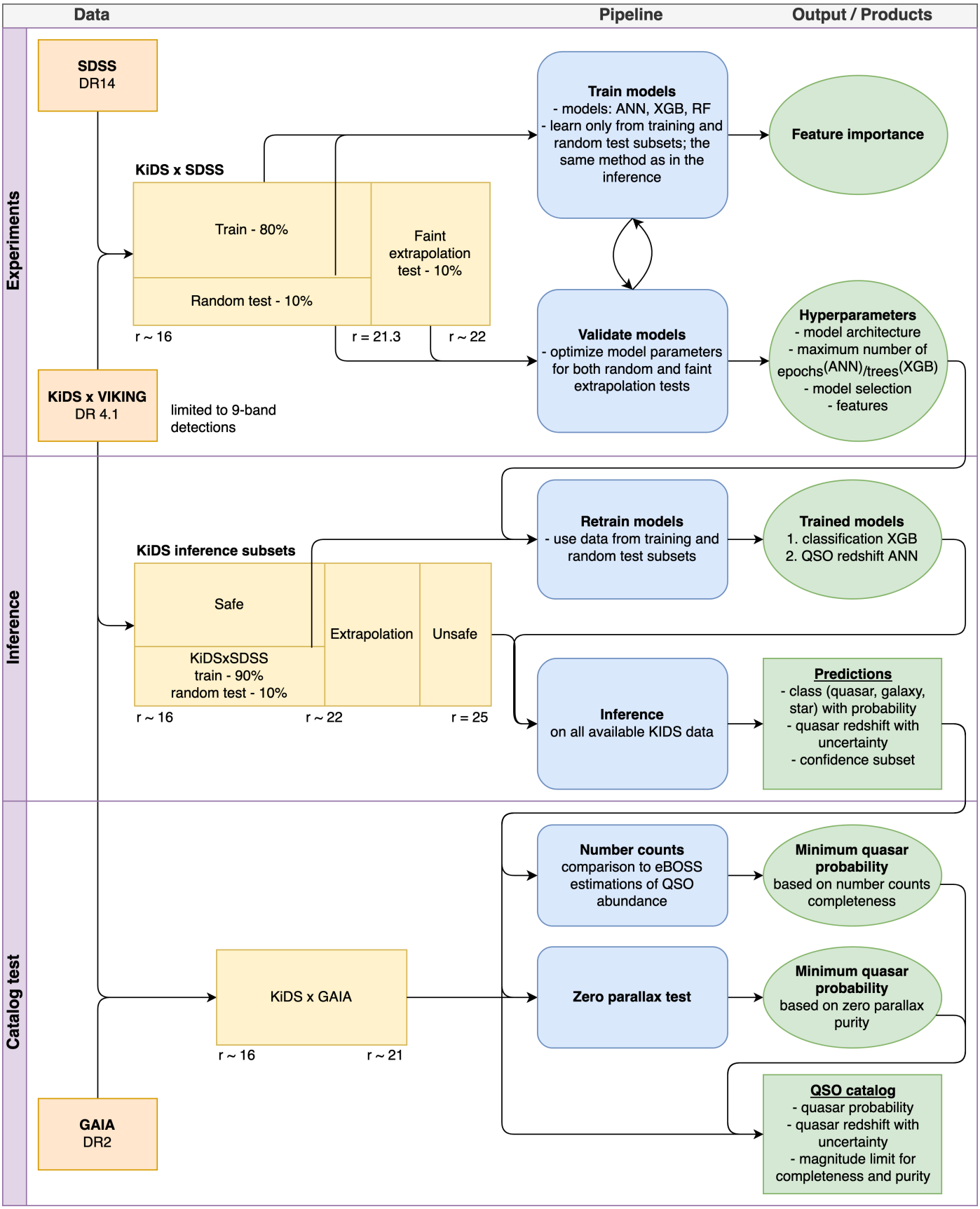

Fig. 4. Methodology diagram. The procedure consists of three main parts: experiments, inference and catalog tests. The experiments are based on the cross-match between KiDS and SDSS data, and include the repeatable process of training and evaluating ML models. The training is based only on the train 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 are then tested with number counts and Gaia parallaxes to calibrate the final catalog with probability cuts for the optimal purity-completeness trade-off.

Table 1. Train and test subsets of the KiDSxSDSS data. We take the faintest 10% of the data as the faint extrapolation test. This splits the training data at $r = 21.3$. We take 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%)

We test quasar redshifts on two subsets: the true spectroscopic quasars from SDSS, and quasar candidates from the output of an ML model. The quasar candidates may contain true stars and galaxies due to misclassification. As we are solving two distinct tasks: classification to identify quasars and regression to estimate their redshifts, a test of quasar candidates evaluates the consistency between classification and redshift models, and 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 quasar candidates as the most important metric for model selection.

We use the following classification metrics[4] (scikit-learn, Pedregosa et al. 2011): accuracy for three-class classification problem (quasar / galaxy / star), purity and completeness for quasar detection. For redshifts, we use:

– mean squared error

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

– R-squared

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

– redshift error

$$\delta z = \frac{z_{\mathrm{photo}} - z_{\mathrm{spec}}}{1 + z_{\mathrm{spec}}} \quad (3)$$

where 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.

2.4. Model selection

We test three of the most popular ML models: random forest (RF, Breiman 2001), XGBoost (XGB, Chen & Guestrin 2016) and artificial neural networks (ANN, Haykin 1998). We use 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 non-linear activation function in each neuron.

We test 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 assign zero redshift to stars. We also test a neural network model with multiple outputs for classification and redshifts, which allows us to solve both problems with only one model.

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

3. Results

3.1. Features

The final set of features consists of 83 values: optical *ugri* and near infrared $ZYJHK_s$ magnitudes, differences (colors) and ratios of every pair of magnitudes, and two point source classifiers: the stellarity index from SExtractor and the 3rd bit of SG2DPHOT from KiDS. We tested other bits of the SG2DPHOT without observable improvement in the results. Ellipticity and other apertures were tested in the previous work (N19) and no significant increase in performance was seen..

Figure 5 shows the most important features for the classification and redshift estimation. The importance for each feature is calculated as a sum of gain that a given features provides to a model in all the splits which are made based on that feature. 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 they are more challenging to estimate. The most important magnitudes for quasar 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 morphological parameters are also used for QSO redshift, allowing models to distinguish extended low-redshift AGNs.

We experimented with reducing the feature set used for classification to minimize possible overfitting and increase model interpretability. It provides stable results for classification, but worsens results of redshift in the subset of quasar 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 redshift assigned incorrectly, while the redshifts of quasar candidates require both the class and the redshift to be assigned correctly.

3.2. Experiment results

Figure 6 compares XGB training histories (number of trees used) for the classification and redshifts. 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 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 quasar 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

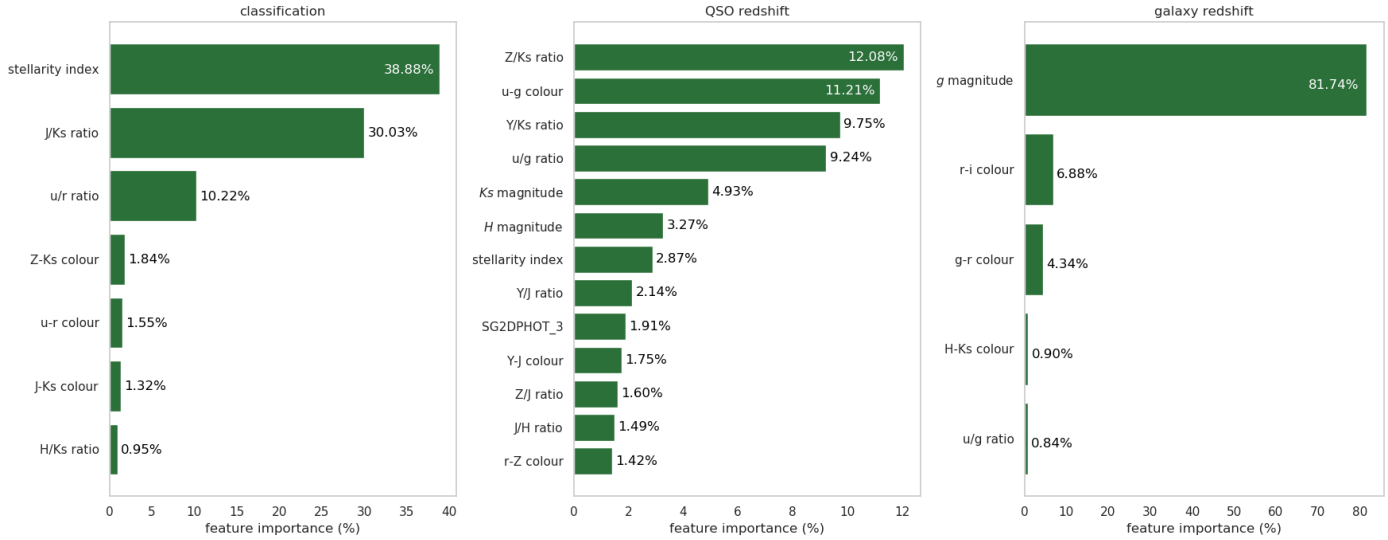

Fig. 5. Feature rankings from the XGB models. *Left:* classification, *centre:* QSO redshift, *right:* galaxy redshift. We use the total gain across all splits the feature is used in. The classification is mostly based on the stellarity index, near-IR JK_s and optical *ur* bands. The quasar redshifts 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 different importance.

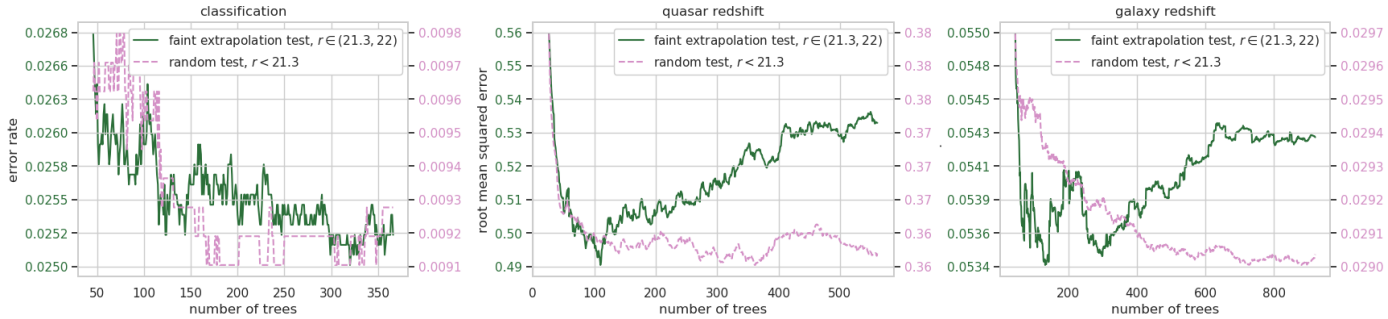

Fig. 6. Learning histories for the XGB models. *Left:* classification, *center:* QSO redshift, *right:* galaxy redshift. The x-axis shows the number of trees created iteratively during the model training, the y-axis shows the classification error rate and redshift root mean square error on two different scales for the random and faint extrapolation tests. The errors in the faint test are higher than in the random tests due to extrapolation and higher noise. The models are stopped if the results on the faint test do 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.

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 make sure that the final inference models, whose training is stopped based only on the random test, will achieve good results also at the faint data end.

ML models can be modified in many ways which control the bias vs. variance trade-off, in addition to the number of trees investigated in Fig. 6. In the case of ANNs we tune the number and size of layers, regularization, dropout and learning rate. Some attempts at model optimization showed improvement in results on both the tests, while 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 extrapolation subset is also incorporated, and uncontrolled failure of estimation for the faint end. In case of incorrectly regularized models, such 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 the both tests, we have the full picture of the bias vs. variance trade-off and we can tune the models to perform well on both bright and faint data, and extrapolate to magnitudes fainter than available from spectroscopy. We consider this an important success of our approach.

We test several ML strategies, and we conclude that the best approach for the final inference is two ANNs, one for classification and one for quasar redshifts (Table 2)[5]. 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

[5] 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.

Table 2. Model comparison on the random ($r < 21.3$) and faint extrapolation ($r \in (21.3, 22)$) tests. The redshifts are tested in two subsets of quasars: 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. Recall is the same as completeness, and MSE, R2, δz are given by equations 1, 2, 3 respectively.

		classification			redshift for true QSOs			redshift for QSO candidates		
test	model	accuracy	purity	recall	MSE	R^2	δz	MSE	R^2	δz
random	RF	99.00%	97.44%	94.31%	0.12	85%	0.018 ± 0.14	0.12	84%	0.032 ± 0.21
	XGB	**99.09%**	**97.85%**	**94.75%**	0.13	84%	0.017 ± 0.15	0.13	83%	0.030 ± 0.21
	ANN	98.98%	96.93%	94.67%	**0.10**	**88%**	**0.009 ± 0.12**	0.11	85%	0.023 ± 0.22
	class XGB, z ANN	**99.09%**	**97.85%**	**94.75%**	**0.10**	**88%**	**0.009 ± 0.12**	**0.10**	**87%**	**0.020 ± 0.19**
faint	RF	**97.44%**	96.12%	**92.37%**	0.31	31%	0.019 ± 0.25	0.33	31%	0.046 ± 0.38
extrap.	XGB	**97.44%**	96.48%	92.12%	0.27	39%	0.036 ± 0.23	0.34	29%	0.077 ± 0.41
	ANN	97.27%	**96.52%**	90.89%	**0.22**	**51%**	**-0.0004 ± 0.19**	**0.28**	**39%**	**0.042 ± 0.37**
	class XGB, z ANN	**97.44%**	96.48%	92.12%	**0.22**	**51%**	**-0.0004 ± 0.19**	0.31	35%	0.050 ± 0.40

on galaxies or quasars, are necessary for the best results, due to the differences in required model parameters between the two classes.

Table 2 shows the results of the specialized redshift models. The redshift metrics (Section 2.3) are calculated on two subsets of quasars: true spectroscopic ones, and our QSO candidates from photometric classification, as explained in Section 2.3. In our previous work (N19), which dealt with classification only, we did not observe 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 in redshifts. A combined approach, where classification is performed with XGB and redshifts with ANN gives the best results overall. 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. Mixing XGB classification with ANN redshifts gives worse results for quasar candidates in the faint test, due to different characteristics of both models resulting in fewer objects with both class and redshift assigned correctly. We find that XGBoost is the most robust and straightforward model for classification, while ANN is the best for combined classification and redshift.

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

Quasar misclassification occurs mostly at low redshift (Fig. 7), 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+NIR photometry those AGNs are more similar to galaxies than to quasars. It is due to the spectra taken through fibres in the SDSS, and in case of galaxies with AGN, the fibre is centred 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 to see the emission lines, therefore such AGNs are classified as galaxies from imaging. The quasars candidates consists of 96.9% quasars, 2.6% galaxies, and 0.4% stars. The bottom plots show results obtained using only the optical *ugri* broadbands. We observe misclassification with stars at QSO redshift $2 < z < 3$ (bottom left), and worse redshift estimates (bottom right), when only KiDS optical imaging is used, as studied previously in N19.

Figure 8 compares spectroscopic and photometric redshifts on the random and faint tests. The random test shows a well-fitted distribution and thus modelled 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 optimised 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 quasars in the random test is $\delta z = -0.38 \pm 0.43$, while in the extrapolation[6], $\delta z = -0.45 \pm 0.32$. The BPZ redshifts for quasars 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.

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 random test on objects with such contamination. To verify this, we evaluated ANNs on the objects flagged with any MASK bit (Table 3). 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. 9). We consider the classification probability limits as the primary way to calibrate the

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

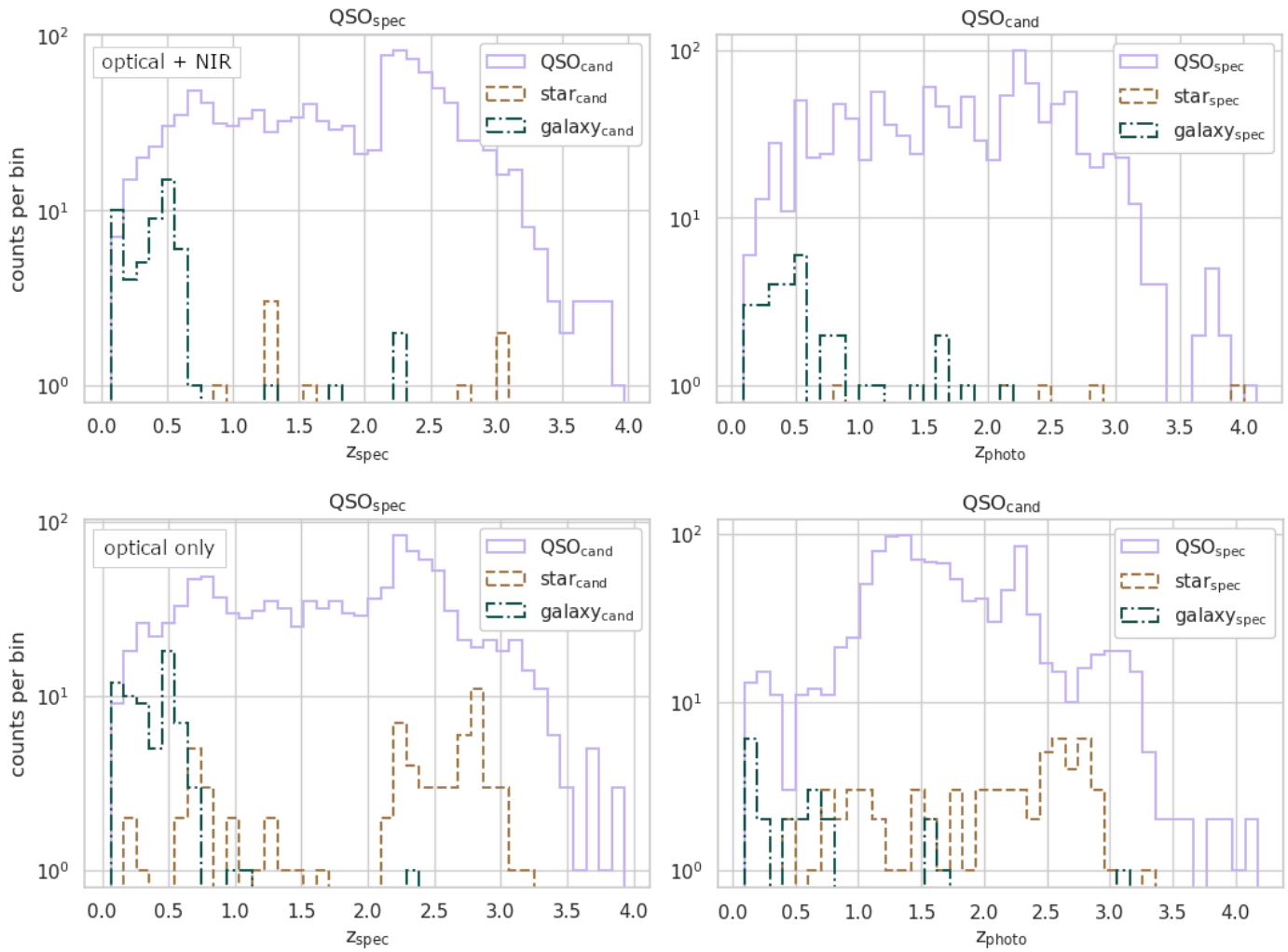

Fig. 7. Quasar misclassification as a function of redshift. *Top:* using optical KiDS and near-IR VIKING features, *bottom:* using only optical KiDS features, *left:* spectroscopic quasars and redshifts – a test for completeness; *right:* quasar candidates and redshifts – a test for purity.

Table 3. 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 2.

test	purity	recall	δz for true QSOs	δz for QSO candidates
random	96.26% (-0.67%)	93.92% (-0.75%)	0.004 (-0.005) ± 0.13 (+0.01)	0.017 (-0.006) ± 0.26 (+0.04)
faint extrap.	94.23% (-2.29%)	88.36% (-2.53%)	0.01 (+0.01) ± 0.22 (+0.03)	0.07 (+0.03) ± 0.42 (+0.05)

catalog's purity-completeness trade-off, while the uncertainties can be used to achieve the necessary redshift precision.

3.3. Final catalog properties

We apply the trained ML models to 45M objects of the KiDS DR4 inference data, and find a total of 3M quasar candidates, excluding the unsafe inference subset. In the final model training, we use the whole range of magnitudes of the training set, as well as a randomly selected validation sample. We employ the same set values of hyper-parameters as determined in the experiments which included the faint extrapolation test, and we only pick a new number of epochs based on new learning histories with a randomly selected test sample.

In Figure 10 we compare the number counts of quasars 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 two cuts on the photometric quasar probability to match the expected numbers: $p(QSO_{cand}) > 0.9$ for the safe magnitude range ($r < 22$), and $p(QSO_{cand}) > 0.98$ for the extrapolation. The fit of the quasar number counts to eBOSS predictions is reliable for $r < 23.5$, where the extrapolation subset is complete (Fig. 3). We do not observe the expected decrease of the quasar number counts at $r > 23.5$, which should result from reaching the completeness limit of the extrapolation subset. This suggests increased impurity of the quasars 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.

Figure 11 shows spatial number densities for KiDS quasar candidates based on the photometric redshifts, and for SDSS spectroscopic quasars based on the spectroscopic redshifts. We account for the V_{max} correction, taking the KiDS magnitude limit

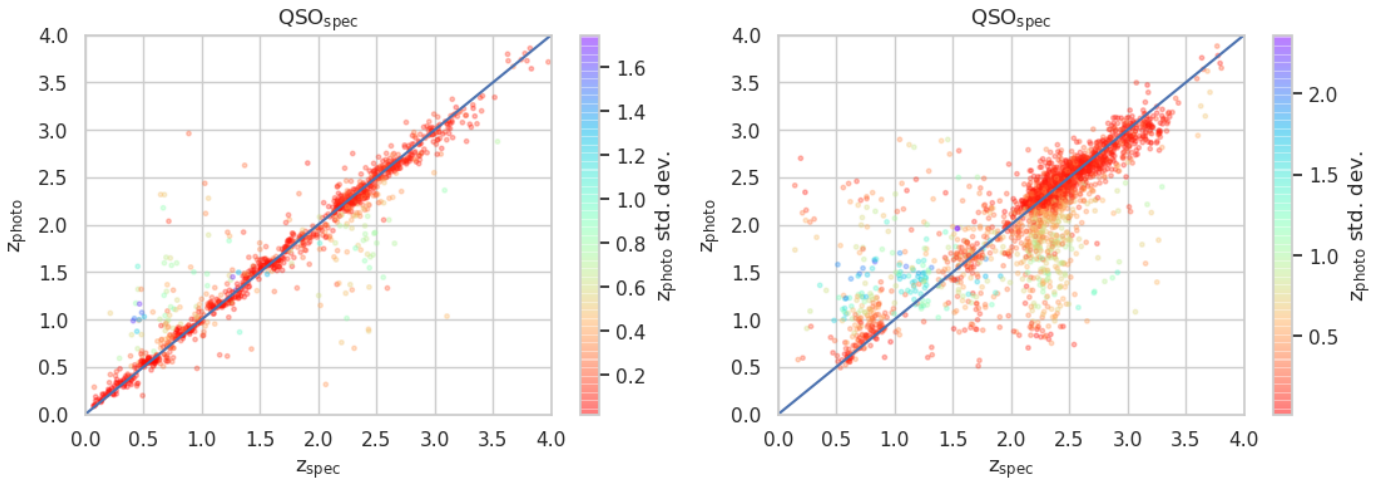

Fig. 8. Comparison of the spectroscopic and photometric redshifts for SDSS test-set quasars. *Left:* random test ($r < 21.3$), *right:* faint extrapolation ($21.3 < r < 22$). The mean photo-z error for the random and faint test equals 0.009 ± 0.12 and -0.0004 ± 0.19, respectively. Every redshift estimate is a Gaussian probability density function, the standard deviation of which represents the uncertainty (color coded).

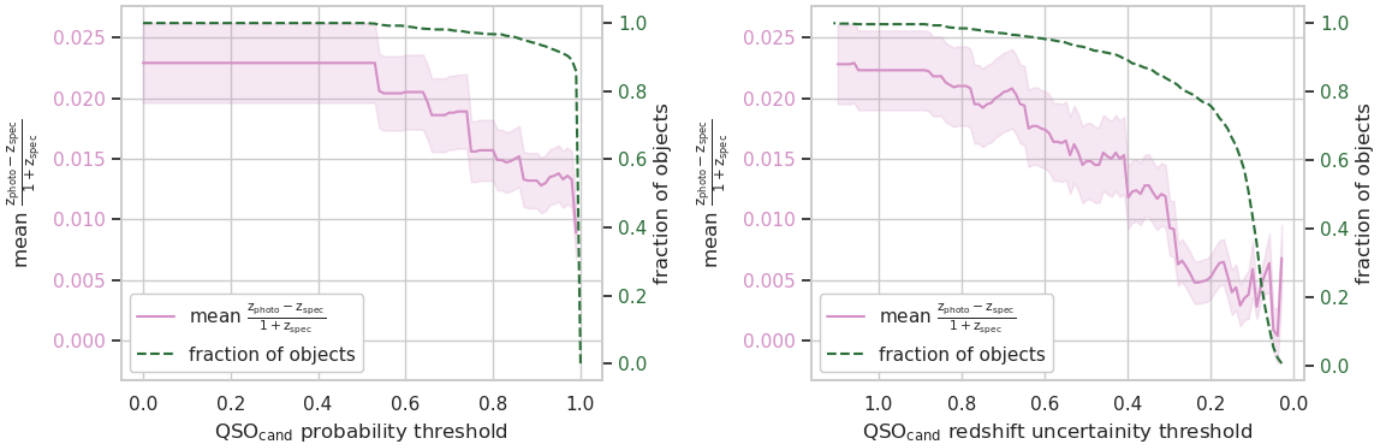

Fig. 9. Quasar photometric redshift errors as a function of thresholds in QSO probability (*left panel*) and model photo-z uncertainty (*right panel*). Increasing minimum classification probability yields better redshift estimations at a small cost in completeness. Low uncertainty estimations further increase redshift reliability at a cost of removing more objects.

$r = 25$, 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. 10) we estimate its completeness to be $r < 19$. We observe some differences between KiDS photometric and SDSS spectroscopic quasar densities at this limit. The quasars missing at low redshifts are due to previously discussed misclassification with galaxies (Fig. 7). 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 $p(\mathrm{QSO_{cand}}) > 0.98$. This is due to apparently faint galaxies classified by our model as quasars and assigned redshifts lower than one. This conclusion agrees with the number counts indicating quasar impurity at $r > 23.5$.

Table 4 summarizes the number of quasars 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 quasar 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 quasar surveys could help verify these finding, as they will include quasars fainter than SDSS, and will overlap with KiDS.

We visualize the outputs from the ML models, compare it to the spectroscopic information and show the final catalog properties for the inference subsets and suggested probability cuts using t-SNE in Fig. 12. The main spectroscopic quasar group is accurately covered with photometric classification and redshifts. In the part of feature space between spectroscopic stars and quasars, we observe an unphysical decision boundary, almost perpendicular to the gradient of magnitude. The predictions, however, appear regular and correct in the close extrapolation, which 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 quasars at the highest redshifts or to select candidates for follow-up spectroscopy.

Table 4. Number of photometrically selected quasars 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 quasar candidates and increase the purity.

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

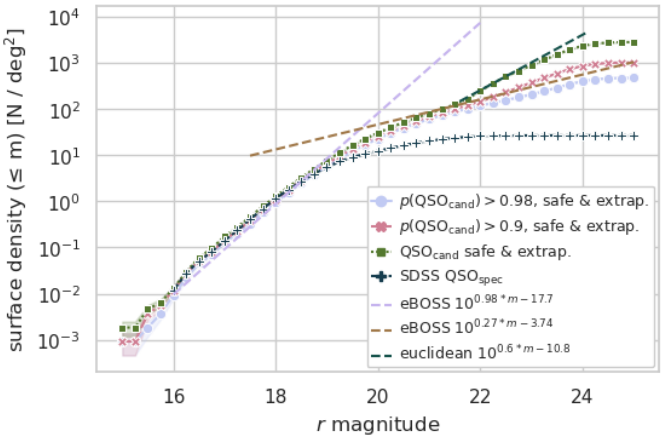

Fig. 10. Quasar number counts of SDSS spectroscopic quasars and KiDS quasars candidates ($\mathrm{QSO_{cand}}$) at progressing classification probability cuts, excluding the unsafe inference subset. The dashed lines show eBOSS predictions fitted with a broken power-law. The SDSS spectroscopic quasars are complete to $r < 19$. KiDS quasars candidates without a probability cut are too numerous at $r > 21.5$ due to misclassification, and 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 $r \lesssim 24$.

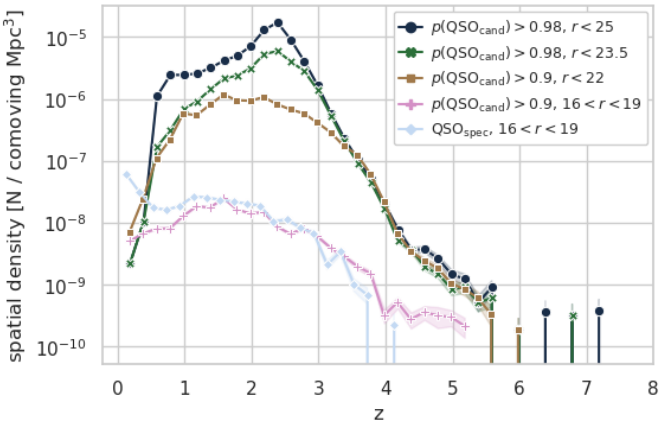

Fig. 11. Spatial number densities, excluding the unsafe inference subset, for KiDS quasars candidates. Two bottom lines compare the KiDS quasar candidates to the SDSS spectroscopic quasars, at the SDSS completeness range $16 < r < 19$. The three upper lines show the final quasar catalog at progressing magnitude limits with the suggested probability cuts. We chose magnitude limit for the middle line at $r < 23.5$, as above this limit the distribution of quasar candidates gains another peak at redshift $z < 1.5$.

3.4. Gaia parallaxes

We cross-match the quasar candidates identified here with Gaia DR2 (Gaia Collaboration et al. 2018) to estimate the star contamination. A clean set of quasars 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 quasars. Following the same procedure for KiDS quasar candidates would remove the star contamination, which we want to measure. Instead, we calculate a less precise mean offset for SDSS quasars 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 quasar candidates in the safe inference subset show a mean parallax offset of 0.003 mas, and this goes down at progressing minimum classification probability (Fig. 13). 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 quasars and 1k (0.1%) as galaxies. The test is limited to the Gaia magnitude $G < 21$, which corresponds to $r \lessapprox 20$. We then calculate an "acceptable offset" from a sample of the three spectroscopic classes, with the size of each class corresponding to the contamination of quasar candidates with stars and galaxies derived from the experiments: 96.9% quasars, 2.6% galaxies, 0.4% stars (Fig. 7). The minimum quasar photometric probability suggested by this test is $p(\mathrm{QSO_{cand}}) = 0.9$. This cut, 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.

3.5. Comparison with other quasar catalogs

We find good agreement with other quasar catalogs overlapping with the KiDS DR4 footprint (Fig. 14). Additional ground-truth samples, which were not used in the training, provide a good test of ML estimations. We use additional quasar catalogs built from different datasets and with different methodologies than ours. Those involve one spectroscopic catalog: 2QZ / 6QZ (Croom et al. 2004, hereafter 2QZ), and three photometric ones: Richards et al. (2009b, 2015); DiPompeo et al. (2015), hereafter R09, R15 and DP15 respectively. 2QZ includes quasars, stars and galaxies confirmed with spectroscopy, while the photometric catalogs are probabilistic, based on selection from SDSS (R09) and SDSS+WISE (R15 & DP15). DP15 publishes the whole range of QSO probabilities, which we limit to higher than 70%, according to the distribution with shows a minimum number of objects at this value. 2QZ, being spectroscopic, can be used as ground truth and confirms high quasar purity and completeness of our sample: 98.2% three class accuracy, 98.6% quasar purity and 99.4% quasar completeness. We note however that as 2QZ sources are on average brighter than those from the SDSS quasar catalog, these numbers should not be taken as measurements of the overall performance of our classification.

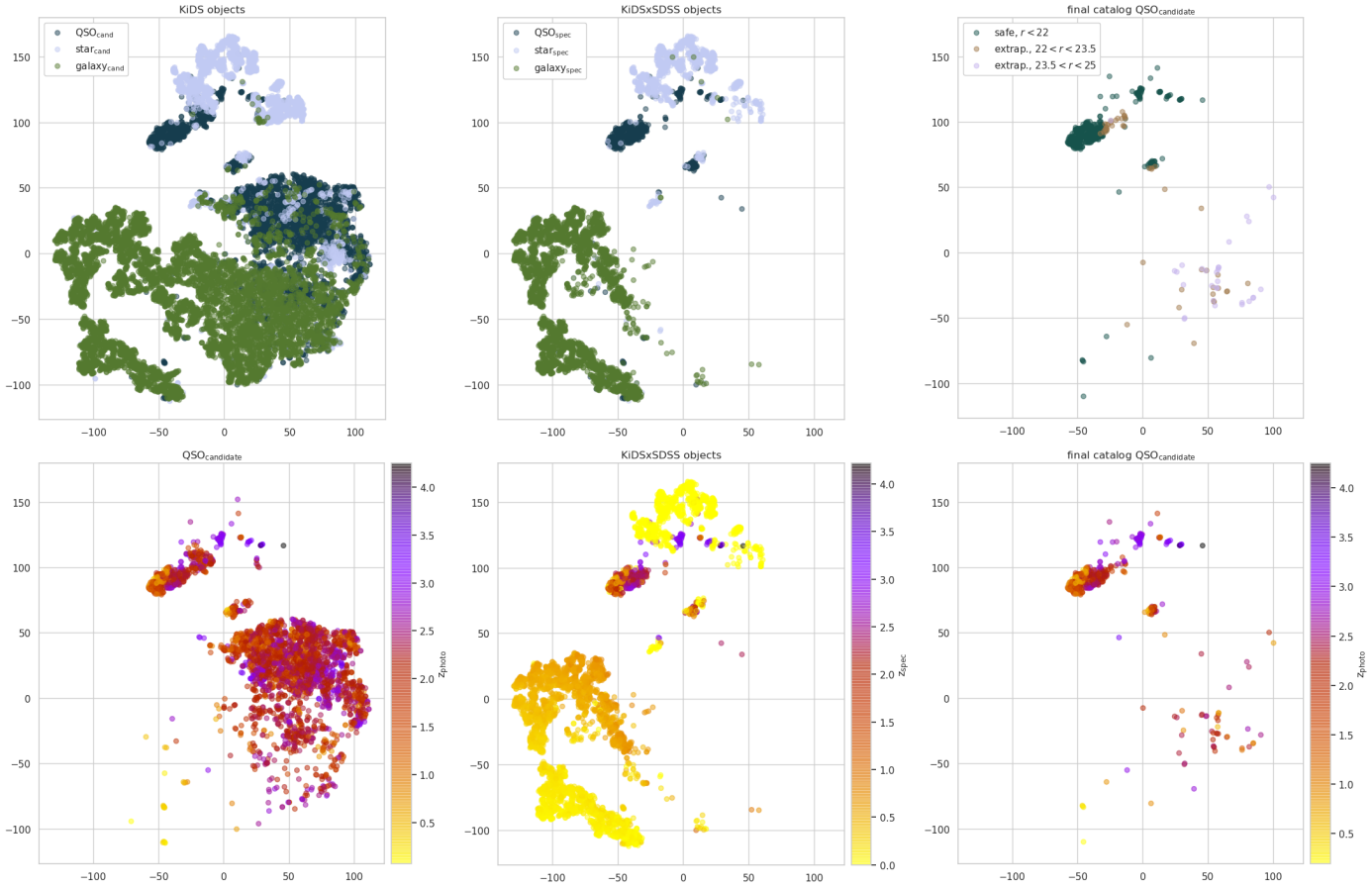

Fig. 12. t-SNE projections. Top: classification, bottom: redshifts, left: raw output from the ML models for all the inference subsets, center: spectroscopic SDSS distributions, right: the final quasar catalog at progressing magnitudes with the corresponding probability cuts, excluding the unsafe inference. 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.

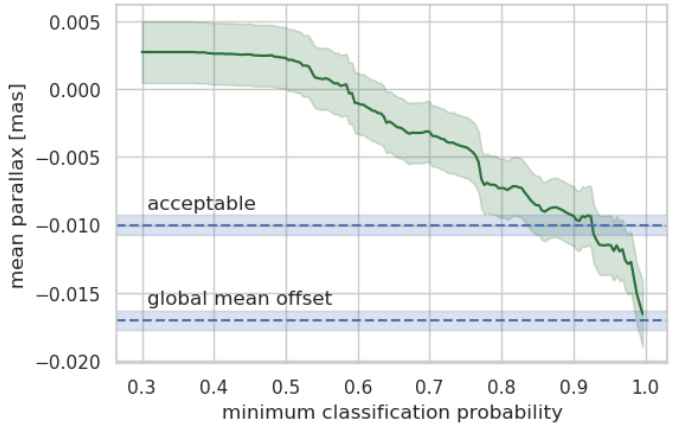

Fig. 13. Mean parallax for KiDS DR4 quasar candidates as a function of minimum classification probability. The Gaia observations have a global mean offset, which is imprinted in the quasar mean parallax distribution. The offset for SDSS spectroscopic quasars equals -0.017 ± 0.001 mas (standard error on the mean). We calculate the acceptable offset based on star and galaxy contamination estimated in the experiments. It equals -0.01 ± 0.0015 mas.

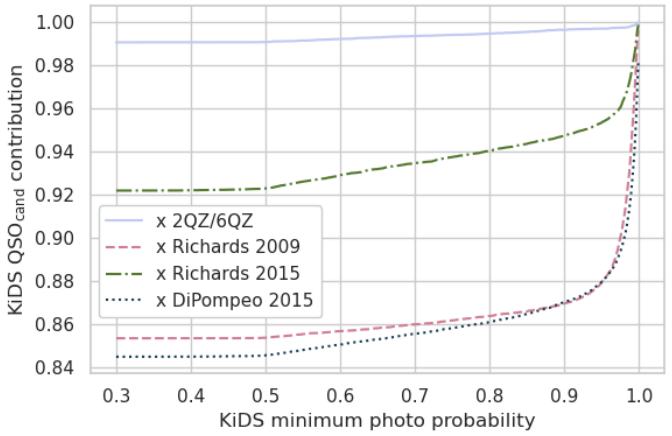

Fig. 14. Proportion of KiDS DR4 quasar candidates in cross-matches with other quasar catalogs as a function of KiDS minimum photometric classification probability.

4. Discussion

4.1. Main findings

In this paper we employed supervised ML models to identify quasars in KiDS DR4 and evaluate their redshifts. We found 158k quasar candidates with minimum classification probability $p(\mathrm{QSO_{cand}}) > 0.9$ at $r < 22$, and a total of 311k quasar candi-

dates with $p(\mathrm{QSO_{cand}}) > 0.98$ for $r < 23.5$, i.e. in the extension to the close extrapolation data. The far extrapolation at $r < 25$ provides a total of 507k quasar candidates at $p(\mathrm{QSO_{cand}}) > 0.98$. The catalog of quasars 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 94.7% for $r < 22$. The extrapolation by ∼0.7 magnitude lowers the purity by 0.4 percentage points and the completeness by 3.6 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 -0.0004 ± 0.19 in the extrapolation ($r < 23.5$).

We found that the traditionally adopted testing method, based on randomly selected samples of objects, was insufficient to tune the bias vs. variance trade-off. A faint-end test is necessary for 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 obtain useful redshift uncertainties also in the extrapolation data, even though we use 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 on which N19 was based, provided crucial information for quasar redshifts and helped us to distinguish stars from quasars at redshifts $2 < z < 3$. The most important bands for quasar 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 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 quasar redshifts. Possible further experiments may involve more custom feature engineering based on flux values, to find the most robust photometric features.

The comparison of ML models also shows clear trends: XGB performs better at classification, while ANN provides better redshift estimation, i.e. works better for regression. Many astronomical papers report no such differences, which was also the case in our previous work (N19). We uncovered these differences as more features are available from the VIKING imaging, which allows 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.

4.2. Relation to other work

Most of the quasar 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. Color-color cuts used for classification which ensure both high purity and completeness require modelling the data with many distributions or building 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, simultaneously optimizing both purity and completeness. The power of ML approaches comes with a 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 & 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, e.g. high and low mass end for galaxy cluster mass estimation, number of objects in n-body problems, cosmological parameters not available during training in cosmological problems, etc.

Our work is the first in which simultaneous selection of quasars from photometry and evaluation of their photometric redshifts is performed for samples selected from the KiDS+VIKING catalog. In a recent study, Logan & Fotopoulou (2020, L20) performed 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 et al. 2017), redshift estimation with random forest, 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/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 as high completeness 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.

As already mentioned in the Introduction, we addressed the quasar selection problem in KiDS in our previous work, N19, where we applied ML classification to the DR3 *ugri* photometry. In that study, we employed the Random Forest algorithm and reported 91% purity and 87% completeness for quasars. In the present work, most of the improvement in classification comes from adding the NIR bands, which allowed us to correctly classify quasars at $2.5 < z < 3$, where they are similar to stars in the *ugri* broad-bands. Additionally, two significant improvements were made: we now provide quasar photometric redshift estima-

tions, and publish estimations for fainter objects with achieved 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 quasars 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 is limited to $r < 22$.

The results of ML quasar identification are not directly comparable between our work and that of Khramtsov et al. (2019), as in the latter the quasars are defined as point-like objects, and any AGNs with visible galaxy host had been removed from the training data, unlike in our case. We keep quasars which appear extended in our training data, as such sources provide useful information on relation between quasars 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 at the specific purpose of quasar 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 best purity-completeness trade-off, which requires ML models which are the best in understanding the problem. 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 quasar.

We envisage that our catalog of quasars can have versatile applications in studies related to AGNs or LSS, as it is optimized solely for quasar identification without outside requirements. The availability of robust photometric redshifts with uncertainty estimates for the quasars 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.

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 quasars at higher redshifts, then based on the finding of our work, we suggest to train the ML classifier only on point-like objects – e.g. those with the stellarity index higher than 0.8 – and apply only quasar vs. star classification. Such a model is easier to train and interpret, and visualisations of the relevant data are simpler to understand than in the full three-class problem including both extended and point sources.

4.3. 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 quasars 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 (Merloni et al. 2019; Richard et al. 2019) quasar surveys, which will largely overlap with KiDS.

The random and faint extrapolation tests require 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, e.g. 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 & Fotopoulou 2020). We achieved satisfactory extrapolation results to $r < 23.5$, 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 of the 9 available. This decreases the size of the KiDS inference data by 55%, from 100 million to 45 million. Additionally, the inference set might not include some of the highest-redshift quasars due to the requirement of optical detections. When looking for such high-z quasars, one would have to perform classification and redshift estimation using only the near-IR bands.

Another source of incompleteness is the removal of the faintest objects for which the SExtractor morphological classifier CLASS_STAR fails; this is further 13 million objects that cannot be used for classification. At $r > 23.5$ the unsafe subset constitutes a large fraction of all KiDS objects (65%), and dominates at $r > 24$ (81%) (Fig. 3). As the stellarity index is in fact one the most important features for the classification (Fig. 5), 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, like imputing the missing values, or more sophisticated approaches such as predicting the missing values or using models designed to work with missing values (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 modelling, in order to be fully useful in the extrapolation range $r > 22$. This can be implemented in ANN with e.g. 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 quasar 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 quasar 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 quasars 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.

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 quasar candidates

with reliable photometric redshifts can be used in both AGN and LSS studies, 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.

Acknowledgements. We would like to express our gratitude to Sotiria Fotopoulou and Natasha Maddox for providing useful comments on the paper. This research was supported by the Polish Ministry of Science and Higher Education through grant DIR/WK/2018/12. SN is supported by the Polish National Science Center through grant UMO-2018/31/N/ST9/03975. MB is supported by the Polish National Science Center through grants UMO-2018/30/E/ST9/00698 and UMO-2018/31/G/ST9/03388. AP is supported by the Polish National Science Center through grant UMO-2018/30/M/ST9/00757. MA acknowledges support from the European Research Council under grant number 647112. AD acknowledges ERC Consolidator Grant (No. 770935). BG acknowledges support from the European Research Council under grant number 647112 and from the Royal Society through an Enhancement Award (RGF/EA/181006). CH acknowledges support from the European Research Council under grant number 647112, and support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. HH is supported by a Heisenberg grant of the Deutsche Forschungsgemeinschaft (Hi 1495/5-1) as well as an ERC Consolidator Grant (No. 770935). KK acknowledges support from the Royal Society and Imperial College. *Author Contributions:* All authors contributed to the development and writing of this paper. The authorship list is given in two groups: the lead authors (SJN, MB, AP), followed by an alphabetical group of those who have either made a significant contribution to the data products, or to the scientific analysis.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org
Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42
Asgari, M., Lin, C.-A., Joachimi, B., et al. 2020, arXiv e-prints, arXiv:2007.15633
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Benítez, N. 2000, ApJ, 536, 571
Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393
Bovy, J., Hennawi, J. F., Hogg, D. W., et al. 2011, ApJ, 729, 141
Bovy, J., Myers, A. D., Hennawi, J. F., et al. 2012, ApJ, 749, 41
Breiman, L. 2001, Mach. Learn., 45, 5
Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140
Brescia, M., Cavuoti, S., & Longo, G. 2015, MNRAS, 450, 3893
Calistro Rivera, G., Lusso, E., Hennawi, J. F., & Hogg, D. W. 2016, ApJ, 833, 98
Capaccioli, M., Schipani, P., de Paris, G., et al. 2012, in Science from the Next Generation Imaging and Spectroscopic Surveys, 1
Carrasco, D., Barrientos, L. F., Pichara, K., et al. 2015, A&A, 584, A44
Chen, T. & Guestrin, C. 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785–794
Chollet, F. 2015, keras, https://github.com/fchollet/keras
Ciesla, L., Charmandaris, V., Georgakakis, A., et al. 2015, A&A, 576, A10
Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, A&A, 639, A84
Croom, S. M., Richards, G. T., Shanks, T., et al. 2009, MNRAS, 392, 19
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, MNRAS, 349, 1397
Cuoco, A., Bilicki, M., Xia, J.-Q., & Branchini, E. 2017, ApJS, 232, 10
Curran, S. J. 2020, MNRAS, 493, L70
de Jong, J. T. A., Kuijken, K., Applegate, D., et al. 2013, The Messenger, 154, 44
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62
de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, ArXiv e-prints [arXiv:1611.00036]
DiPompeo, M. A., Bovy, J., Myers, A. D., & Lang, D. 2015, MNRAS, 452, 3124
DiPompeo, M. A., Hickox, R. C., Eftekharzadeh, S., & Myers, A. D. 2017, MNRAS, 469, 4630
DiPompeo, M. A., Hickox, R. C., & Myers, A. D. 2016, MNRAS, 456, 924
DiPompeo, M. A., Myers, A. D., Hickox, R. C., Geach, J. E., & Hainline, K. N. 2014, MNRAS, 442, 3443
Edelson, R. & Malkan, M. 2012, ApJ, 751, 52
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32
Eftekharzadeh, S., Myers, A. D., White, M., et al. 2015, MNRAS, 453, 2779
Fan, X. 2006, New A Rev., 50, 665
Fotopoulou, S., Pacaud, F., Paltani, S., et al. 2016, A&A, 592, A5
Fotopoulou, S. & Paltani, S. 2018, A&A, 619, A14
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Hausen, R. & Robertson, B. E. 2020, ApJS, 248, 20
Haykin, S. 1998, Neural Networks: A Comprehensive Foundation, 2nd edn. (Upper Saddle River, NJ, USA: Prentice Hall PTR)
Heintz, K. E., Fynbo, J. P. U., Ledoux, C., et al. 2018, A&A, 615, A43
Heymans, C., Tröster, T., Asgari, M., et al. 2020, arXiv e-prints, arXiv:2007.15632
Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, A&A, 633, A69
Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19
Ho, S., Agarwal, N., Myers, A. D., et al. 2015, J. Cosmology Astropart. Phys., 5, 040
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Joudaki, S., Mead, A., Blake, C., et al. 2017, MNRAS, 471, 1259
Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055
Kewley, L. J., Maier, C., Yabe, K., et al. 2013, ApJ, 774, L10
Khramtsov, V., Sergeyev, A., Spiniello, C., et al. 2019, A&A, 632, A56
Kormendy, J. & Ho, L. C. 2013, ARA&A, 51, 511
Kuijken, K. 2008, A&A, 482, 1053
Kuijken, K. 2011, The Messenger, 146, 8
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, 625, A2
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500
Kurcz, A., Bilicki, M., Solarz, A., et al. 2016, A&A, 592, A25
Laurent, P., Eftekharzadeh, S., Le Goff, J.-M., et al. 2017, J. Cosmology Astropart. Phys., 7, 017
Leistedt, B., Peiris, H. V., & Roth, N. 2014, Physical Review Letters, 113, 221301
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Logan, C. H. A. & Fotopoulou, S. 2020, A&A, 633, A154
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
Maddox, N., Hewett, P. C., Warren, S. J., & Croom, S. M. 2008, MNRAS, 386, 1605
Małek, K., Buat, V., Burgarella, D., et al. 2020, in IAU Symposium, Vol. 341, IAU Symposium, ed. M. Boquien, E. Lusso, C. Gruppioni, & P. Tissera, 39–43
McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2
Merloni, A., Alexander, D. A., Banerji, M., et al. 2019, The Messenger, 175, 42
Nakoneczny, S., Bilicki, M., Solarz, A., et al. 2019, A&A, 624, A13
Palanque-Delabrouille, N., Magneville, C., Yèche, C., et al. 2016, A&A, 587, A41
Pasquet-Itam, J. & Pasquet, J. 2018, A&A, 611, A97
Pearson, K. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. 2018, in Advances in Neural Information Processing Systems 31, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Curran Associates, Inc.), 6638–6648
Richard, J., Kneib, J. P., Blake, C., et al. 2019, The Messenger, 175, 50
Richards, G. T., Deo, R. P., Lacy, M., et al. 2009a, AJ, 137, 3884
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009b, ApJS, 180, 67
Richards, G. T., Myers, A. D., Peters, C. M., et al. 2015, ApJS, 219, 39
Richards, G. T., Nichol, R. C., Gray, A. G., et al. 2004, ApJS, 155, 257
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61
Scranton, R., Ménard, B., Richards, G. T., et al. 2005, ApJ, 633, 589

Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Sherwin, B. D., Das, S., Hajian, A., et al. 2012, Phys. Rev. D, 86, 083006
Shu, Y., Koposov, S. E., Evans, N. W., et al. 2019, MNRAS, 489, 4741
Śmieja, M., Struski, L. u., Tabor, J., Zieliński, B., & Spurek, P. a. 2018, in Advances in Neural Information Processing Systems 31, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Curran Associates, Inc.), 2719–2729
Spiniello, C., Agnello, A., Napolitano, N. R., et al. 2018, MNRAS, 480, 1163
Stalevski, M., Ricci, C., Ueda, Y., et al. 2016, MNRAS, 458, 2288
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Stölzner, B., Cuoco, A., Lesgourgues, J., & Bilicki, M. 2018, Phys. Rev. D, 97, 063506
van der Maaten, L. & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579
van Uitert, E., Joachimi, B., Joudaki, S., et al. 2018, MNRAS, 476, 4662
Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, MNRAS, 453, 2259
Warren, S. J., Hewett, P. C., & Foltz, C. B. 2000, MNRAS, 312, 827
Wright, A. H., Hildebrandt, H., van den Busch, J. L., et al. 2020, A&A, 640, L14
Wu, X.-B., Hao, G., Jia, Z., Zhang, Y., & Peng, N. 2012, AJ, 144, 49
Yang, G., Boquien, M., Buat, V., et al. 2020, MNRAS, 491, 740
Yang, Q., Wu, X.-B., Fan, X., et al. 2017, AJ, 154, 269
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Table A.1. Columns provided in data products.

Label	Description
ID	ESO ID
RAJ2000	Centroid sky position right ascension (J2000)
DECJ2000	Centroid sky position declination (J2000)
MAG_GAAP_r	r-band GAaP magnitude with optimal MIN_APER (extinction corrected)
CLASS_STAR	SExtractor star-galaxy classifier
MASK	9-band mask information
{CLASS}_PHOTO	Probability that the source is in one of the three classes: GALAXY, QSO, STAR
CLASS_PHOTO	Object class with the highest probability
Z_PHOTO_QSO	Photometric redshift for quasars
Z_PHOTO_STDDEV_QSO	Uncertainty of photometric redshift for quasars
SUBSET	ML inference subset (Section 2.2). Values: safe, extrapolation, unsafe.

Appendix A: Data products

Data are available at: http://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php. Table A.1 describes the data columns. Here we provide only a subset of the KiDS columns, the rest can be obtained by cross-matching with the full KiDS DR4 data by ID.

Appendix A.1: Catalog of quasar candidates

Filename: KiDS_DR4_QSO_candidates.fits
File size: 110 MB
Number of objects: 1,095,711
Data limited to:

- 9-band detections
- $r < 25$
- CLASS_STAR < 0.2 or CLASS_STAR > 0.8
- $p(\mathrm{QSO_{cand}}) > 0.9$

Possible values for the inference subset: safe, extrapolation. Suggested cut for the extrapolation subset: $r < 23.5$ and $p(\mathrm{QSO_{cand}}) > 0.98$ (table 4).

Appendix A.2: Catalog of all machine learning estimates

Filename: KiDS_DR4_all_ML_estimates.fits
File size: 5.5GB
Number of objects: 45,469,955
Data limited to 9-band detections.
Possible values for the inference subset: safe, extrapolation, unsafe.