Astronomy & Astrophysics manuscript no. main
October 6, 2019
©ESO 2019

arXiv:1906.01638v1 [astro-ph.GA] 4 Jun 2019

KiDS-SQuaD II: Machine learning selection of bright extragalactic objects to search for new gravitationally lensed quasars

Vladislav Khramtsov[1], Alexey Sergeyev[1,2], Chiara Spiniello[3,4], Crescenzo Tortora[5], Nicola R. Napolitano[3,6], Adriano Agnello[7], Fedor Getman[3], Jelte T. A. de Jong[8], Konrad Kuijken[9], Mario M. Radovich[10], HuanYuan Shan[11], and Valery Shulga[12,2]

[1] Institute of Astronomy, V. N. Karazin Kharkiv National University, 35 Sumska Str., Kharkiv, Ukraine
[2] Institute of Radio Astronomy of the National Academy of Sciences of Ukraine
[3] INAF - Osservatorio Astronomico di Capodimonte, Salita Moiariello, 16, I-80131 Napoli, Italy
[4] European Southern Observatory, Karl-Schwarschild-Str. 2, 85748 Garching, Germany
[5] INAF – Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125, Firenze, Italy
[6] School of Physics and Astronomy, Sun Yat-sen University, 2 Daxue Road, Tangjia, Zhuhai, Guangdong 519082, P.R. China
[7] DARK, Niels Bohr Institute, Copenhagen University, Lyngbyvej 2, 2100 Copenhagen, Denmark
[8] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands
[9] Leiden Observatory, Leiden University, P.O.Box 9513, 2300RA Leiden, The Netherlands
[10] INAF - Osservatorio Astronomico di Padova, via dell'Osservatorio 5, I-35122 Padova, Italy
[11] Shanghai Astronomical Observatory (SHAO), Nandan Road 80, Shanghai 200030, China
[12] College of Physics of Jilin University, Qianjin Street 2699, Changchun, 130012, P.R.China
e-mail: [vld.khramtsov, alexey.v.sergeyev, chiara.spiniello]@gmail.com

Submitted on June 03, 2019

ABSTRACT

Context. The KiDS Strongly lensed QUAsar Detection project (KiDS-SQuaD) aims at finding as many previously undiscovered gravitational lensed quasars as possible in the Kilo Degree Survey. This is the second paper of this series where we present a new, automatic object classification method based on machine learning technique.
Aims. The main goal of this paper is to build a catalogue of bright extragalactic objects (galaxies and quasars), from the KiDS Data Release 4, with a minimum stellar contamination, preserving the completeness as much as possible, to then apply morphological methods to select reliable gravitationally lensed quasar candidates.
Methods. After testing some of the most used machine learning algorithms, decision trees based classifiers, we decided to use CatBoost, that was specifically trained with the aim of creating a sample of extragalactic sources as clean as possible from stars. We discuss the input data, define the training sample for the classifier, give quantitative estimates of its performances, and finally describe the validation results with *Gaia* DR2, AllWISE, and GAMA catalogues.
Results. We have built and make available to the scientific community the KiDS Bright EXtraGalactic Objects catalogue (KiDS-BEXGO), specifically created to find gravitational lenses. This is made of ≈ 6 millions of sources classified as quasars (≈ 200 000) and galaxies (≈ 5.7M), up to $r < 22^m$. From this catalog we selected 'Multiplets': close pairs of quasars or galaxies surrounded by at least one quasar, presenting the 12 most reliable gravitationally lensed quasar candidates, to demonstrate the potential of the catalogue, which will be further explored in a forthcoming paper. We compared our search to the previous one, presented in the first paper from this series, showing that employing a machine learning method decreases the stars-contaminators within the gravitationally lensed candidates.
Conclusions. Our work presents the first comprehensive identification of bright extragalactic objects in KiDS DR4 data, which is for us the first necessary step to find strong gravitational lenses in wide-sky photometric surveys, but has also many other more general astrophysical applications.

Key words. gravitational lensing: strong – methods: data analysis – surveys – catalogs – quasars: general – galaxies: general

1. Introduction

Strong gravitationally lensed quasars are very rare objects, especially in the case of quadruply lensed (Oguri & Marshall 2010). However, it was clear, since the first discovery (Walsh et al. 1979), that these systems are extremely useful tools for observational cosmology, cosmography and extragalactic astrophysics.

When the light coming from a distant quasar intercepts a massive galaxy, it gets blended and it forms multiple images of the same source, which are often also magnified, becoming brighter. The light-curves of these different images have different paths and thus are offset by a measurable time-delay that depends on the cosmological distances between the observer, the lens and the source, and on the gravitational potential of the lens (Refsdal 1964). This time-delays return a one-step measurement of the expansion history of the Universe (primarily H_0), and also allow to set constrains on the dark matter halo of the lens galaxy (Suyu et al. 2014).

Moreover, on top of the deflection caused by the lens, the light of the quasar can also be deflected by the gravitational field of other low-mass bodies moving along the line-of-sight ($10^{-6} < m/M_\odot < 10^3$, e.g., single stars, brown dwarfs, planets,

globular clusters, etc.). This phenomenon, known as microlensing, can be very effective to study the inner structure of the source (Anguita et al. 2008; Motta et al. 2012; Sluse et al. 2011; Guerras et al. 2013; Braibant et al. 2014), to estimate the masses of these compact bodies (Kochanek 2004) or to study the stellar content of the lens galaxies (Schechter & Wambsganss 2002; Bate et al. 2011; Oguri et al. 2014).

Unfortunately, in all these mentioned cases, and especially for cosmography, the biggest limitation lies in the relatively small number of confirmed lenses.

Thus, taking advantage from the high spatial resolution (0.2″/pixel, Capaccioli & Schipani 2011) and stringent seeing constraints (< 0.8″ in r-band) of the Kilo Degree Survey (KiDS, de Jong 2013; de Jong et al. 2017; Kuijken et al. 2019), we have recently started the KiDS Strongly lensed QUAsar Detection project, KiDS-SQuaD, presented in Spiniello et al. 2018, hereafter Paper I. We are carrying on a systematic census of lensed quasars with the final goal of building a statistically relevant sample of lenses, covering a wide range of parameters (geometrical configurations, deflector masses and morphologies, redshifts and nature of the sources) to study the the dark matter halos of lens galaxies up to z $\sim$ 1 (Schechter & Wambsganss 2002; Bate et al. 2011; Suyu et al. 2014) as well as the QSO-host galaxy co-evolution up to z $\sim$ 2 (e.g., Ding et al. 2017), to put constraints on the inner structure of the quasar accretion disk (size and thermal profile; e.g. Anguita et al. 2008; Motta et al. 2012) as well as the broad-line-region geometry (e.g., Sluse et al. 2011; Guerras et al. 2013; Braibant et al. 2014) and finally for precise cosmography (e.g Suyu et al. 2017).

The first step to find gravitationally lensed quasars is, obviously, to classify objects, selecting quasars and galaxies while minimizing as much as possible the stellar contamination.

More generally, the identification of extragalactic objects, quasars and galaxies at all redshifts, is a very important task, that can help to answer to a wide range of astrophysical and cosmological questions, such as the relationship between active galactic nuclei (AGN) and host galaxies or the cosmic evolution of Super Massive Black Holes (Kauffmann & Haehnelt 2000; Haehnelt & Kauffmann 2000; Wyithe & Loeb 2003; Hopkins et al. 2006; Shankar et al. 2009; Shen et al. 2009) or the formation and evolution of galaxies (Driver et al. 2009) across cosmic time.

Spectroscopy is without any doubt a powerful way to unambiguously identify and classify extragalactic objects. The most comprehensive dataset of spectroscopically confirmed quasars to date is the Sloan Digital Sky Survey (Pâris et al. 2018), and a few forthcoming spectroscopic surveys will exponentially increase the amount of confirmed quasars, e.g. Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration et al. 2016) and 4-meter Multi-Object Spectroscopic Telescope (4MOST, Richard, et al. 2019). However spectroscopy comes with a price: it is in fact time-expensive or effective only on small patches on the sky. Deep wide-field sky photometric surveys, on the other hand, offers nowadays an unprecedented opportunity to carry on this task on a much larger portion of the sky, modulo the development and the use of sophisticated automatic methods (e.g., Decision Trees, Quinlan 1986, Naives Bayes, Duda & Hart 1973, Neural Networks, Rumelhart et al. 1973, Support Vector Machines, Vapnik 1995; Cortes & Vapnik 1995) to process the very large amount of produced data.

KiDS, in particular, is the ideal platform to identify and classify objects and, more specifically to search for strong gravitational lenses, because of its excellent (for ground observation standards) seeing quality (mean r-band $\approx$ 0.70 $FWHM$), depth-ness (25^m in r-band) as well as its wide field of view (1350 deg^2 have been covered and will be released with DR5).

The power of KiDS in the objects classification has already been shown by Nakoneczny et al. 2019, hereafter N19, who built and released a catalogue of quasars from KiDS DR3 (440 deg^2), classified with a random forest supervised machine learning model, trained on Sloan Digital Sky Survey DR14 (SDSS DR14, Abolfathi et al. 2018) spectroscopic data. The approach we undertake in this paper is similar to the one presented in N19, as we also use KiDS data as input and SDSS as training sample, although we fine-tune and customize our pipelines to be more suitable for the search of lensed quasars. Moreover, the biggest difference between these works is that now we have available photometry in nine-bands. In fact, the optical data in KIDS are now (starting from DR4) complemented by infrared data from the VISTA Kilo-degree INfrared Galaxy (VIKING) survey, covering the same KiDS area in the Z, Y, J, H, K_s near-infrared bands (Edge et al. 2013). Thus, the KiDS×VIKING photometric dataset provides a unique deep, wide coverage in nine bands ($u, g, r, i, Z, Y, J, H, K_s$) which has been proved to be extremely effective to separate quasars from stars using photometric characteristics (e.g. Carrasco et al. 2015).

Indeed, one of the limitations of the first paper of this series was the manual optical colors selection we used to select quasars-like objects. In fact, in this way, the number of final lensed quasar candidates highly depends on the (somehow arbitrary and often calibrated on previous finding) selection criteria. Moreover, generally this number is of the order of 10 ÷ 30 per deg^{-2}, making the necessary second step of visual inspection very difficult and long. Thus, to make our research suitable to deal with the larger amount of data coming from the fourth (and in the future the fifth) KiDS Data Release (Kuijken et al. 2019) and also new deep wide-field surveys, e.g. Euclid (Laureijs et al. 2011) or LSST (Ivezić, Ž. & LSST Science Collaboration 2013), here, in this second paper, we developed a method based on machine learning (ML) and on the combination of VIKING and KIDS data that allow us to more efficiently pinpoint high redshift systems while eliminating as much as possible stellar contamination.

ML methods are, in fact, very effective in identifying quasars (and more generally, extragalactic sources,Eyer & Blake 2005; Ball et al. 2006; Elting et al. 2008; Kim et al. 2011; Gieseke et al. 2011; Kovács & Szapudi 2012; Brescia et al. 2015; Carrasco et al. 2015; Peters et al. 2015; Krakowski et al. 2016, 2018; Viquar et al. 2018; Khramtsov & Akhmetov 2018; Nolte et al. 2019; Bai et al. 2019) with respect to any manual colour cut. They allow to explore, with a little human intervention and affordable computing time, large datasets, thus selecting candidates with less stringent pre-selection criteria, maximizing the precision (recovery rate) and minimizing the stellar contamination.

Recently, a class of specific type of classifiers, the ensembles of decision trees, showed their advantage in the identification of extragalactic sources, and in particular quasars (Ball et al. 2006; Carrasco et al. 2015; Hernitschek et al. 2016; Schindler et al. 2017, 2018; Sergeyev et al. 2018; Jin et al. 2019), also, specifically, as already mentioned, within the KiDS collaboration N19. Moreover, ML methods to search for the strong gravitational lenses already exist, although we note that the large majority of them are build to find galaxy-galaxy lenses rather than lensed quasars using a deep learning approach (e.g., Cabanac et al. 2007; Paraficz et al. 2016; Lanusse et al. 2018; Metcalf et al. 2018; Petrillo et al. 2017, 2019a,b; but see also Agnello et al. 2015; Ostrovski et al. 2017; Krone-Martins et al. 2018; Jacobs et al. 2019 for lensed quasars). Moreover, many of these methods

are based on the analysis of imaging data directly, rather than on a catalog level like we do in this paper.

Nevertheless, we decided to build our own classifier in order to be able to fully customize the characteristics and parameters of the algorithm, also given the required completeness and purity we need for the resulting catalogue. It is of crucial importance for us to build a catalogue of extragalactic objects (not only quasars, since in some case the deflector can give a non-negligible contribution to the light of the whole system or the multiple images can be blended together and thus produce in the KiDS catalogue an 'extended' match rather than many 'point-like' ones), that is as clean as possible from stars, and, at the same time, as complete as possible. Thus, developing our own tool and releasing the resulting catalogue is the best possible choice.

As main result of the novel classification pipeline that, specifically developped for our specific task, we present here the catalogue of Bright EXtraGalactic Objects from KiDS DR4 – KiDS-BEXGO, which we then use to search for gravitationally lensed quasars, using some of the methods and idea already presented in Paper I.

This paper is organized as follows: in Section 2 we give a general overview on the catalogues and data we use. In Section 3 we discuss the method to classify objects and isolate extragalactic ones, using optical and infrared deep photometry, and we introduce and describe our own classification pipeline. In Section 4 we present the result of such a pipeline: the KiDS-BEXGO catalogue, and different validation techniques, based on external data, to test the performance of the classifier. Finally, in Section 5 we focus on 'Multiplets': close pairs of quasars, or galaxies surrounded by at least one quasar (within 5″), which represent the primary input catalogue for our search for strong gravitational lenses. We present our conclusions and future perspectives in Section 6. In addition, we present in the Appendix a direct comparison of three different machine learning methods, all based on decision trees.

2. Data overview

2.1. The input catalogue from KiDS DR4

The Kilo-Degree Survey (KiDS, de Jong 2013) is an European Southern Observatory (ESO) public survey, carried on with the VLT Survey Telescope (VST, Capaccioli & Schipani 2011; Capaccioli et al. 2012), that covers 1350 deg^2 on sky in four optical broad-band filters, namely u, g, r, i. Optical data from KiDS are complemented with data from the VISTA Kilo-Degree Infrared Galaxy Survey (VIKING, Edge et al. 2013), that has already completed the observations in five near-infrared bands (Z, Y, J, H, K_s) within the same region of the sky. The latest KiDS data release (KiDS DR4 Kuijken et al. 2019), encompasses all the survey tiles (1006 deg^2 in total) already released in the previous KiDS data releases (de Jong et al. 2017) with additional tiles covering ≈ 550 new deg^2, thus doubling the area coverage of DR3. In addition, infrared photometric data from VIKING is also included in the KiDS DR4 release for the aperture-matched sources (Wright et al. 2018). Typical magnitude limits for each band are 24.2, 25.1, 25.0, 23.6, 22.7, 22.0, 21.8, 21.1, 21.2 (AB magnitudes, 5σ in 2″ aperture), with seeing generally below 1.0″ in $u, g, r, i, Z, Y, J, H, K_s$ bands (Wright et al. 2018; Kuijken et al. 2019).

We started from the KiDS multi-band DR4 catalogue and selected ≈ 45M sources, that were detected in the r-band, which is the one with the best seeing (0.7″), and have a match in each of the other eight bands too. However, for the implementation of the classification method presented in this paper, we do not use the full catalogue but we limit to 9 583 913 sources with $r < 22^m$, covering ≈1000 deg^2 in all of the 9 filters with small errors on each magnitudes (we remove all the sources with `MAGERR_GAAP`$> 1^m$ in each of the band). In fact, as in N19, we also use spectroscopically confirmed objects from the Sloan Digital Sky Survey Data Release 14 (SDSS DR14, Abolfathi et al. 2018) as training sample and therefore we limit our inference to bright objects to avoid any extrapolation to unseen regions in the space of features.

Throughout this paper we always make use of the Gaussian Aperture and PSF (GAaP, Kuijken 2008; Kuijken et al. 2015) magnitudes, corrected for extinction. Finally, as we describe in more details below, we also use the magnitude-dependent parameter `CLASS_STAR` for the objects classification. This was already proven to be a very important feature in N19.

The histogram of the r-band magnitude distribution for the whole KIDS DR4 and for the spectroscopically confirmed objects that we use as training sample is shown in Figure 1. The training sample will be presented in details in the next Section.

2.2. The training sample from SDSS DR14

To provide accurate classification, we need to use a large sample of objects with known true classes. Such data can be obtained from spectroscopic surveys; for our purpose, following the approach of N19, we use the SDSS DR14[1] catalogue. The SDSS DR14 catalogue contains 4 311 571 spectroscopically confirmed objects, classified on the basis of their spectra in three main classes: galaxies (2 546 963 objects), quasars (824 548 objects) and stars (940 060 objects), which we will preserve in our classification setting up a 3-label classification system, as we will describe in details in Section 3.

We assume that a quasar (hereafter, `QSO`) is a point-like source[2] with `QSO` class and `QSO` or `BROADLINE` subclass; a normal galaxy (hereafter, `GALAXY`) is an extended source that has a `GALAXY` class label without `STARFORMING BROADLINE` and `STARBURST BROADLINE` subclasses. The stars labeling in SDSS does not have subclasses, so simply we assumed that the source is a star (hereafter, `STAR`) if it has the class `STAR` from the catalogue.

We cross-match this catalogue with the catalogue of bright sources from KiDS DR4 described above, using a 1.0 arcsec radius, and obtaining, as result, a training sample composed of 183 048 sources. However some of them have dubious spectroscopic classification.

A careful cleaning is very important for our scientific purpose, but an automatic masking procedure, eliminating all the dubious cases, that is often applied in classification pipelines to reach the highest possible pureness, is not appropriate here because it might cause the loss of interesting objects with complex morphology and photometric properties, that can be actually good lens candidates[3]. We therefore had to pay particular

[1] SDSS DR14 is the second release of the Sloan Digital Sky Survey IV phase (Blanton et al. 2017) and it includes data from all previous SDSS data releases.

[2] we note that this assumption has not been made in N19 that included in their `QSO` training sample the relatively near ($z < 0.2$) AGNs and visible host galaxies.

[3] as a matter of fact, inspecting the misclassified data from SDSS we found three interesting lensed quasar candidates that we selected for spectroscopic follow up. If confirmed, the system will be presented in a forthcoming publication.

attention to the cleaning procedure which we carried on in a rather manual and interactive way. In particular we use this first "unclean" training sample to train the classifiers (which we will describe in the next section). We then visually inspected all the misclassified objects (of the order of few hundreds)[4]. Interestingly, during the inspection, we discovered, that SDSS DR14, indeed, contains few objects with wrong labels, possibly due to a somehow imperfect procedure (among these we found few white dwarfs and few compact galaxies labeled as QSO, blended sources where one of the component is a star, or stars projected into a galaxy) and realized that classifiers trained on such dataset can inherit these mistakes. Thus we removed all the the sources, the true class of which did not fit with its imaging and/or spectral properties and we repeat the whole classification pipeline a few times (testing it also with different classifiers, see next section). We note that the total amount of removed sources does not exceed a few percents of the training sample, but that still, the classification results before and after this iterative cleaning procedure are not identical, with the classifier learned with the "clean" training sample producing better results in terms of pureness.

Finally, another test we did to get a better handle on the importance of our assumptions in building the input catalogue and the best training sample for it, was to change the chosen threshold in the photometric errors of each single band. In particular, we tested three different upper limit for the errors on the magnitudes of the training sample: 1^m, 0.5^m and 0.3^m. As for the cleaning, we trained the classifiers three times with three different training sets made of objects passing these three thresholds and then we compared the performances. We found negligible differences in purity and completeness (at the 0.1% level) in the classification of the training sample. Then we also compared the predictions for the whole input catalog obtained using the three different training samples, finding, again, no significant differences in the distribution of the sources among the classes. Thus, we decided to use the training sample with the largest number of object and the same error threshold as the input catalogue (1^m).

In conclusion, after removing the sources with 1) bad spectroscopic redshift estimation (for which zWarning > 0), 2) missing one or more of the 9 optical-infrared magnitudes, 3) high photometric errors (> 1^m in each filter) and the 4) accidental duplicates, retrieved after our cross-matching procedure, we ended up with 121 375 sources, of which 24 307 sources classified as STAR, 12 152 sources classified as QSO, and 84 917 sources labeled as GALAXY. This catalogue, which we will name for the rest of the paper SDSS×KiDS, will be used in Section 3.1 as training sample for the classifiers.

3. Classification

The main idea behind the classification problem that we have to solve is that it is possible to separate objects into stars, quasars and galaxies on the basics of their photometry, because each family of objects has specific photometric characteristics, which are different from objects belonging to a different family. Thus, our first task was to define the feature space where the quasars, galaxies and stars will be located in three different well separated regions. We identify optical-infrared colours as the most suitable features for the objects classification; since we have 9 magnitudes ($u, g, r, i, Z, Y, J, H, K_s$), there are 36 colours as pairwise

[4] We use the Navigate SDSS visual tool (http://skyserver.sdss.org/dr14/en/tools/chart/navi.aspx) to inspect misclassified sources

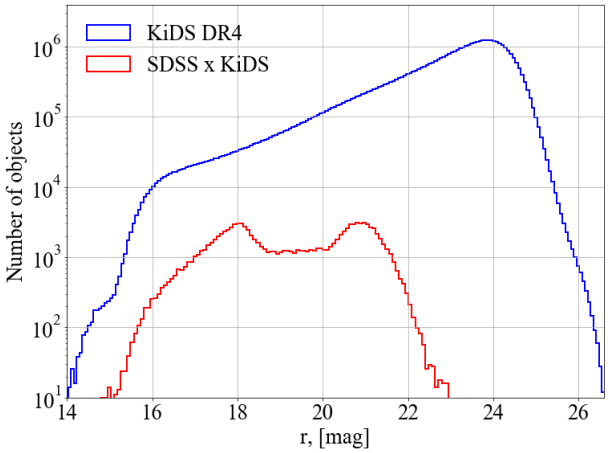

Fig. 1: Histogram of the r magnitudes for the full KiDS DR4 catalogue (blue) and the training sample from SDSS×KiDS (red).

differences of various magnitudes. We note that this approach is physically-motivated and model-driven. N19 showed that, although magnitudes contribute less to classification than colors and stellarity index, the output based on colors only was different from that using also magnitudes. However, we have many more colors at our disposal, thanks to the five additional infrared bands available to us, which make it easier to properly separate stars from quasars and galaxies. Moreover, considering the fact that it is somehow hard to find simple cut in magnitudes that allow to separate different classes of objects, we decided to do not use magnitudes and only consider colors which are more effective in separating the different classes.

Also, we add the CLASS_STAR flag to the feature set, corresponding to the 'stellarity' of a source and derived from the KiDS r-band images, the ones with the best seeing. This KiDS parameter is a continuous measure of whether the object is extended (CLASS_STAR=0) or point-like (CLASS_STAR=1) and has been proved to be a very powerful feature in the classification (e.g. N19). As showed in de Jong (2013) (Fig. 8), the CLASS_STAR parameter depends on the signal-to-noise ratio (SNR) and it is an effective way to separate stars from galaxies only for data with SNR> 50. Thus, an alternative for selecting the input data to classify, which would probably allow to investigate also fainter magnitudes, might be to put a cut in the SNR rather than on r-band magnitude. However, at the present stage, the more severe cut in magnitude is necessary given the limitation of the training sample that we use.

Colors and stellarity values of the sources correspond to the coordinates in the high-dimensional feature space, in which the classification has been performed.

As already specified in the introduction, we decided to define a 3-class problem, where the classes correspond to stars, quasars and galaxies. In fact, this 3-class labeling allows us to get the purest identification of quasars, unlike the 2-class (stars and quasars, assuming galaxies as extended sources) or the 4-class (stars, quasars, regular galaxies and galaxies with strong emission lines, e.g. starforming galaxies) schemes. Also, we stress that a 2-class problem, which only separates stars from extragalactic sources is not enough for our scientific purpose. It is true that to find gravitationally lensed quasars, we need a catalogue that contains both galaxies and quasars, but we need to be able to separate these two classes properly in order to find suitable lens candidates (see Section 5).

In the following sections we describe the final classification algorithm and calibration strategy that we use to build our catalogue, which was the end product of a large series of tests and experiments we carried out, also using different classification schemes, detailed in the Appendix A. In fact, we tested three classifiers based on decision trees (Random Forests and two different Gradient Boosting approaches). We decided, in the end, to use the CatBoost (Dorogush et al. 2018; Prokhorenkova et al. 2018), one of the two Gradient Boosting (GB, Friedman 2000) ensemble algorithms that we tested, because it was the one providing the best performance during the training process, as described in the following Section.

In general, Gradient Boosting (GB, Friedman 2000) is an ensemble algorithm that constructs a learner by fitting in an iterative way the gradients of the predictions' residuals of the previously constructed learners, typically decision trees (gradient boosting decision tree, GBDT). CatBoost in particular, is a novel, fast, scalable, high performance open-source GBDT library[5], developed by Yandex researchers and engineers[6]. CatBoost has the great advantage, with respect to other Gradient Boosting algorithms, that it uses Ordered Boosting (Prokhorenkova et al. 2018) to avoid the overfitting problem, as we highlight in the Appendix A.

To our knowledge, this is the first application of the CatBoost algorithm to an astronomical task.

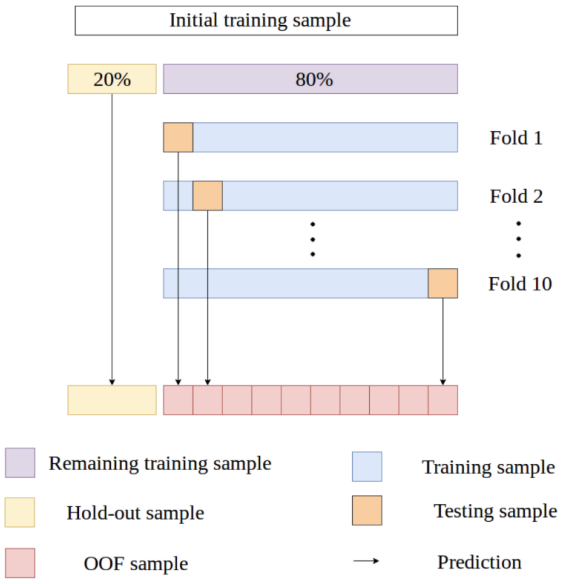

Fig. 2: A schematic view of the learning via 10-fold cross validation procedure and validation with the OOF and the hold-out samples, drown from the initial training sample

3.1. Fine-tuning and learning process

To be able to analyze the performance of classifier, one need to define a set of validation data and the type of learning with respect to the training-to-validation division. We therefore split the validation into two groups: out-of-fold (OOF) and hold-out. The hold-out sample consists of a random subsample of the initial training data which we keep to access the final performance of the classifiers. The remaining part of the initial training sample is used to learn the classifiers with a k-fold cross validation procedure. This method is one of the most commonly used way to train classifiers and directly compare classification algorithms. Briefly, one divides the training sample into k randomly partitioned disjoint equal parts. Then, the classification algorithm trains on $k-1$ parts and the remaining one is used as testing data. This process is repeated k times, each time using one of the k disjoint testing subsamples and obtaining a prediction from it. The combination of these k predictions is the so-called OOF sample. Finally, to obtain the prediction on the new data, one have to make k predictions, from each fold's model, and average them. A schematic view of the learning process is visualized in Figure 2.

Starting from the KiDS×SDSS sample of 121 376 sources, we randomly selected 20% of it as hold-out sample and use the remaining 80% as OOF training sample in the cross-validation process[7] We stress that, among the classifiers that we tested, CatBoost returned the best performance both on the hold-out and OOF samples.

Before training can take place, the classifier has a list of parameters that have to be tuned to reach the highest possible classification quality. This is true for each of the different classifiers that we tested (see the Appendix for more details on each of them). For this purpose, we performed optimal hyperparameter search on 60% of initial training sample via a 3-fold cross-validation with a a 'BayesSearch' for CatBoost (and XGBoost, while we use a 'GridSearch' method for RF).

While tuning the wide range of hyperparameters for CatBoost, we noticed that the most influential ones were the `max_depth` and the `early_stopping` parameters. We selected `max_depth` = 8 and `early_stopping` = 150 after BayesSearch, with a maximum number of trees equals 3500.

Moreover, we applied a weighting criterion to the loss function for the CatBoost model to further decrease the contamination by stars in the extragalactic objects catalogue (see Appendix A.1 for more details).

After the above described fine-tuning, we finally trained CatBoost with the same training and validation data with a 10-fold cross-validation (see Fig. 2). The result of the performance for the final CatBoost model (after the fine-tuning) is presented, as confusion matrices, in Figure 4 for the OOF sample (top) and the hold-out sample (bottom). Using the weighting for stars and galaxies, we received a significant improvement in the purity of the quasar sample; in fact, comparing the confusion matrices before weighting loss function (see Appendix) and after that, one can see, that the rate of stars, classified as quasars, decreased from $\approx 0.60\%$ to $\approx 0.30\%$. CatBoost lost only $< 1.50\%$ of the quasars, thus only marginally decreasing the resulting completeness of this class.

Another notable result, that we can get with CatBoost, is the relative importance that each feature has for the classification procedure. Feature importance, calculated with decision tree, shows the frequency, with which a certain feature occurs in the tree. In such a way, the higher frequency is directly related to the higher feature contribution to separate the sources, i.e. to the importance of a given feature. An excellent example of this kind of analysis, together with a full description of the feature importance technique is presented in D'Isanto et al. (2018). Figure 3 shows the 10 most informative features for the all of the CatBoost models, trained one by one via 10-fold cross-validation. Among these, the `CLASS_STAR` is certainly the most important

[5] https://catboost.ai/
[6] https://yandex.com/company/
[7] The hold-out and OOF samples are kept fixed for all the various algorithms that we tested (see Appendix).

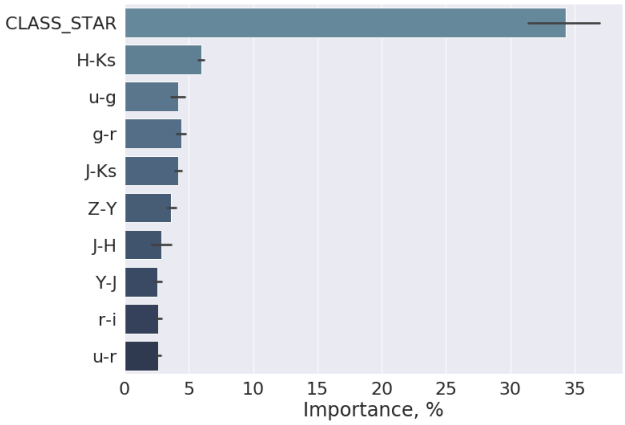

Fig. 3: Importance of the 10 most significant features, calculated with CatBoost in each of 10 folds. The dispersion of importance for each feature is represented by horizontal ticks at each bar.

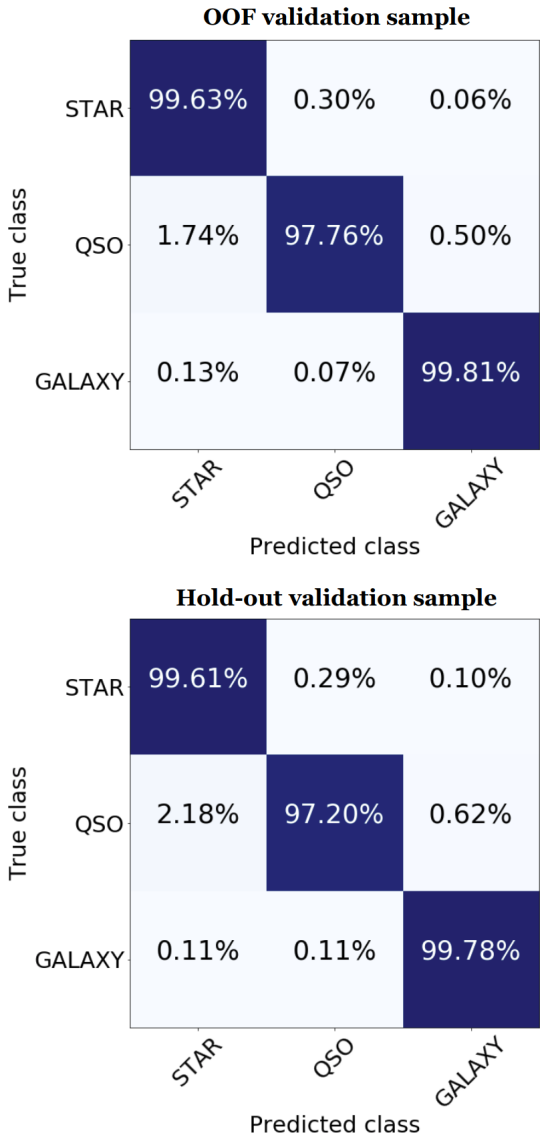

Fig. 4: Confusion matrices for the final version of CatBoost, after weighting the loss function, performed on the OOF sample (*top* panel) and the hold-out sample (*bottom* panel).

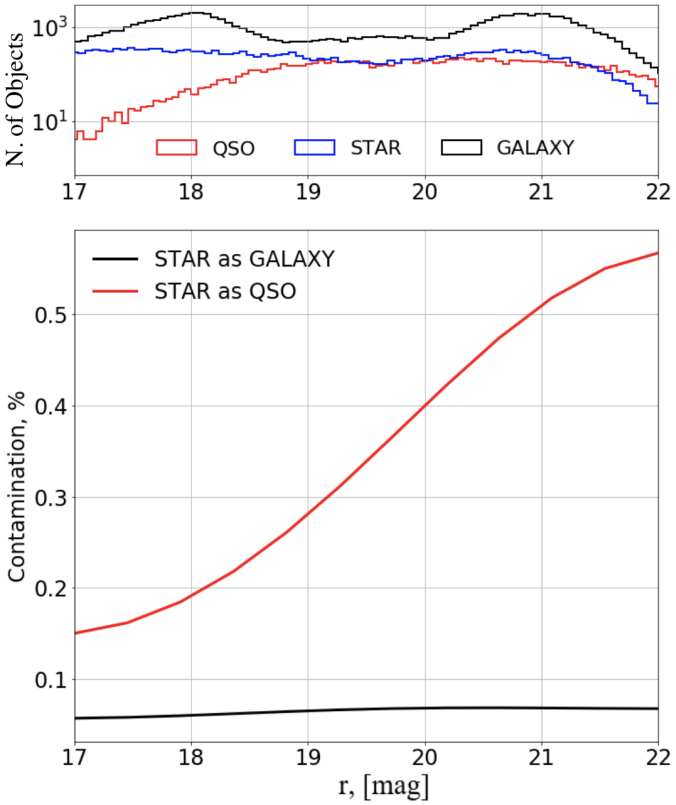

Fig. 5: Top panel: histogram of the r magnitudes for the three classes of training sources. Bottom panel: the rate of stars, misclassified as quasars (red curve) and galaxies (black curve), as a function of their r magnitude. The plots are produced for the full initial training sample.

one, followed by $H - K_s$, $u - g$, $g - r$, and $J - K_s$. This is in perfect agreement with a number of results in the literature, e.g., using $u-g$ and $g-r$ colour diagram it is possible to separate low-redshift quasars from stars (Abraham et al. 2012; Carrasco et al. 2015); these features, together with the stellarity, are in fact also the most important ones in other ML based classificators (e.g. N19). Also, quasars at $z \approx 2.5$ and $z \approx 5.6$ may be recovered by employing K-band information in the colour space (Chiu et al. 2007). And finally, it is well known, and also intuitively easy to understand, that morphological information (described here by the `CLASS_STAR` feature) allows us to clearly select galaxies, dividing them from stars and quasars in the relatively bright magnitude range that we consider ($r < 22^m$).

The maximum rate of stars-contaminators per bin of r-band magnitude in the quasars catalogue equals $\leq 0.6\%$ and is expected at the faint end of the sample ($r \approx 22$ mag). Instead, the stars misclassified as galaxies span over the full optical r magnitude range and does not exceed the 0.1%. This is clearly shown in the bottom panel of Figure 5).

Finally, we checked the contamination rate against the signal-to-noise ratio (SNR) in u-band, which is the noisiest one, finding that the relative contamination of stars decreases at each magnitude bin by $\sim 2\%$ when we only consider objects with SNR > 100.

For the input data, whose distribution in the feature space should be similar to that of the training set, we expect a contamination of 0.3% from stars and 0.1% from galaxies in the sample of quasars, and 0.1% from stars and 0.6% from quasars in the sample of galaxies. Thus, we conclude that the algorithm is able

to correctly classify up to 97.5% of all the bright quasars from the KiDS DR4 data, and up to 99.8% of the galaxies.

Surely, these estimations are ideal and do not really reflect the real situation, because they are only based on the training sample, that is a much smaller and simpler case than the full KiDS DR4 catalogue. A more realistic estimate of the quality of our extragalactic catalogue, in terms of purity and completeness, can be obtained using the external data to validate the resulting sample of classified sources, as we will do in Section 2).

We stress that for our final scientific purpose of finding gravitational lensed quasars, the most crucial point is to be able to get rid of the stellar contaminants. It is, in fact, of fundamental importance to separate as well as possible stars from quasars, being both point-like sources.

Since the KiDS DR4 input catalogue consists of galaxies mostly, there will be a non negligible number of galaxies contaminating the quasars sample. However, as we will explain in more details in Section 5, strong lenses can be classified as `GALAXY`, if the deflector gives a non-negligible contribution to the light and the multiple images of quasar are not deblended (thus, the whole system will result in the one extended object with colors which are a mix of galaxies and quasars typical colors), or they can be identified as multiple quasars. This is the main reason why we build and inspect a catalogue containing all the extragalactic sources (i.e. `QSO+GALAXY`), looking for 'multiplets'(i.e. sources classified as `QSO` and with at least one near-by `QSO` companion) to find lenses belonging to the latter group and by looking for galaxies with at least two quasars near-by (within 5″) to find lenses belonging to the former group.

4. The Bright EXtraGalactic Objects Catalogue in KiDS DR4 (KiDS-BEXGO)

The outputs of the CatBoost classifier for each object are three numbers which represent the probability of belonging to the different classes of objects: $p_{\texttt{STAR}}$, $p_{\texttt{GALAXY}}$, $p_{\texttt{QSO}}$.

In general, we assume that a source belongs to a given class when the probability of being in that class is the highest. With this simple assumption, starting from the input 9.5 million sources in the KiDS DR4 catalogue, we retrieved: 181 336 quasars, 3 711 692 stars and 5 690 885 galaxies. Using instead a more severe threshold, i.e., considering that an object belongs to a class when the corresponding probability is > 0.8, we obtain: 5 665 586 (59%) "sure" galaxies, 3 660 368 (38%) "sure" stars and 145 653 (1.5%) "sure" quasars, plus 122 306 objects (1.3%) with "unsecure" classification.

We note that for the classification of objects in the final catalog we stick to the original assumption that a source belongs to the class with the largest probability, without applying any further threshold, since "unsecure" extragalctic sources (with $p_{\texttt{GALAXY}} \sim p_{\texttt{QSO}}$) could very well be good lens candidates where the deflector and the quasar images are blended and all give a contribution to the light of the system.

Table 1: Number of resulting objects in the classified KiDS DR4 catalogue, for each class with different probability threshold to define class belonging.

cut-off	$p_{\texttt{QSO}}$	$p_{\texttt{GALAXY}}$	$p_{\texttt{STAR}}$
> 0.99	62 425	5 538 193	3 001 287
> 0.95	112 222	5 605 735	3 533 787
> 0.90	128 393	5 629 623	3 611 762
> 0.80	145 653	5 655 586	3 660 368
> 0.67	161 818	5 673 902	3 688 514
> 0.50	181 336	5 690 885	3 711 692

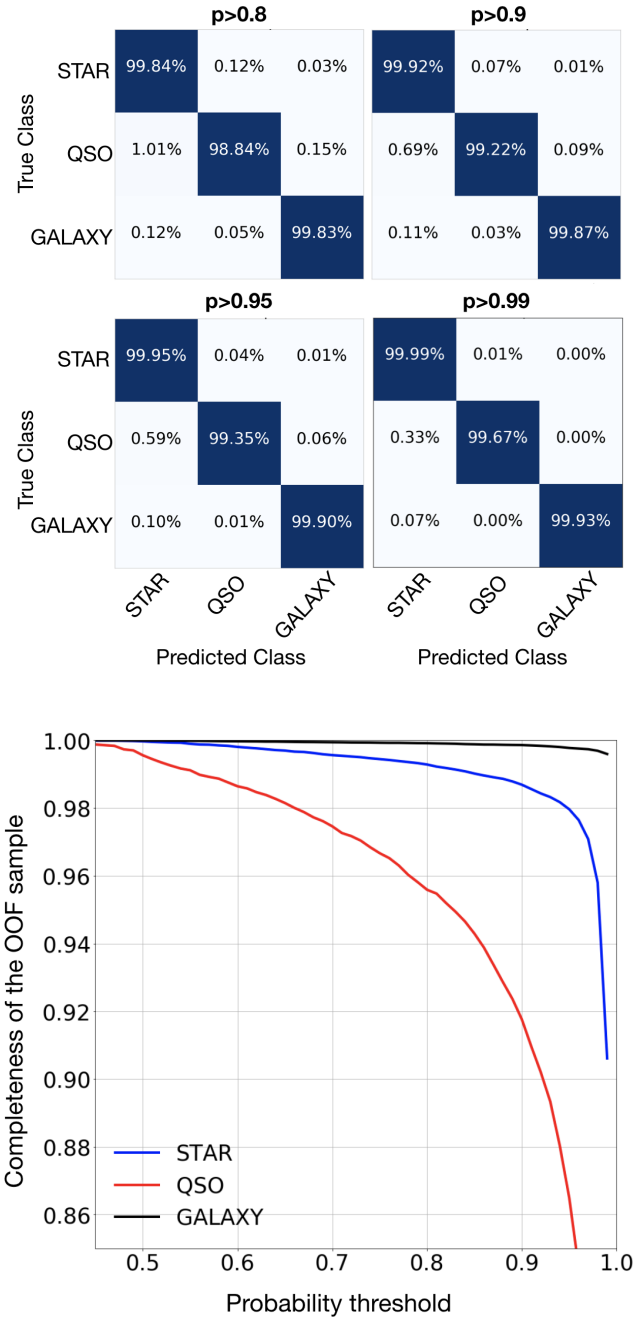

Fig. 6: The upper panel shows the confusion matrices for the final version of CatBoost performed on the OOF sample using different threshold of probability in the objects classification (see text for more details). The lower panel shows instead the completeness rate of the OOF sample as a function of the adopted probability threshold for each class.

However, since the levels of completeness and purity depend on the chosen probability, and different scientific cases might require different levels, we provide the number of objects classified in each subsample for 5 different thresholds in Table 1.

We also show the confusion matrices obtained for the OOF training sample for the four highest probability levels (0.8, 0.9, 0.95, 0.99) and the completeness rate as a function of the proba-

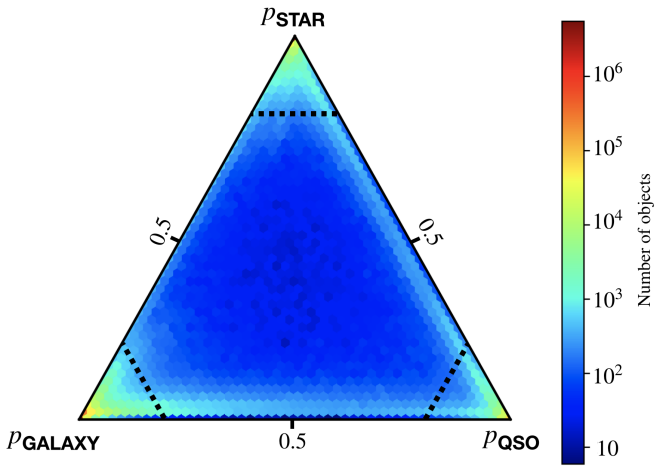

Fig. 7: Density plot of the final distribution of sources among the classes in the KiDS DR4 catalogue. The triangle corners show the maximum probability to belong to a given family (left QSO, right GALAXY, up STAR), and colors indicate number of objects. Dashed lines correspond to the $p = 0.8$ threshold.

bility threshold for the three classes, in Fig. 6, where the thresholds were applied to each of the classes.

Finally, Fig. 7 provides a visualization of the class distribution of the classified KiDS DR4 catalogue with a density plot, where each corner of the triangle represents the maximum probability to belong to a given class. Objects within the region delimited by dotted lines are "sure", according to the threshold given above ($p > 0.8$).

In the next section, we will only focus on all the objects with $p_{\mathrm{QSO}} > p_{\mathrm{STAR}}$ or $p_{\mathrm{GALAXY}} > p_{\mathrm{STAR}}$, that form the Bright EXtra-Galactic Objects Catalogue in KiDS DR4 (KiDS-BEXGO) that we will then use in Section 5 for the gravitational lens search. We discuss here instead three of the many possible validation procedures, for one or more classes of objects, performed using external data (from the *Gaia* astrometric survey, from the AllWISE infrared catalogue, from the GAMA survey). Using external dataset to validate catalogues obtained with ML techniques is a rather standard procedure, as e.g. already shown in N19 and Khramtsov et al. (2018), although in latter case the PMA (Akhmetov et al. 2017) catalogue of proper motions was used to validate purity of galaxies.

Given the results presented in the tests below, together with predictions on the hold-out sample, we are very confident that our ML classifier is able to minimize the stellar contamination in the BEXGO catalogue, which is the first, most crucial step if aiming at digging for gravitationally lensed quasars within very large catalogues.

4.1. Astrometric validation

Recently, the *Gaia* (Gaia Collaboration 2016) astrometric survey has provided optical realization of the International Celestial Reference System, materialized with ≈ 500 000 quasars, and named *Gaia* DR2 Celestial Reference Frame (*Gaia*-CRF2, Gaia Collaboration 2018b; Lindegren et al. 2018).

The latest data release, *Gaia* DR2 (Gaia Collaboration 2018a), introduced 5 astrometric parameters (positions α, δ, proper motions μ_α, μ_δ, and parallaxes $\bar{\omega}$) for 1.3 billion sources, covering the whole celestial sphere up to $G < 21^m$[8]. The systematical errors in *Gaia* DR2, estimated with a large sample of quasars, do not exceed 0.03 mas. Thus, the *Gaia* DR2 provides an excellent mean for testing the purity of our catalogue, especially for quasars. In fact, one of the main observational properties of quasars, that can be used to validate the sample of candidates classified as such, is positional stationary in the optical wavelength range. Being quasars very distant sources, they have proper motions of only few microarcseconds, due to different cosmological effects (Bachchan et al. 2016).

We cross-matched the KiDS DR4 sample of 9.5 million classified sources with the *Gaia* DR2 catalogue using a 0″.5 radius, and retrieved a sample of sources with defined astrometric parameters, of which 52 636 were classified as QSO, 2 369 414 were classified as STAR and 25 346 – as GALAXY.

We checked the proper motions and parallaxes of all the objects classified as quasars and with a match in *Gaia*, to test this assumption that quasars are indeed zero-proper motion and zero-parallax sources within the systematic errors. The results of this test are shown in Figure 9.

The behaviour of the proper motion components is consistent with the estimated contamination of stars within the quasar subsample of the KiDS DR4 catalogue (Fig. 5). In fact, at the faintest magnitudes ($20^m.5 < G < 21^m.0$), the proper motion components deviates strongly, due to a larger contamination from stars. At the bright end of magnitudes, the standard deviation of the mean of proper motions and parallaxes is also large, but this is rather due to relatively small amount of sources and, possibly, to contamination from stars. Also, it is important to note, that the parallax (right plot) is biased for the sample of extragalactic sources towards the value of −0.029 mas (Lindegren et al. 2018).

According to the statistical measures of astrometric parameters, that is reported in Table 2, we thus can conclude, that the sample of KiDS DR4 quasars mainly consists of motionless sources. There is some disagreement between median and mean values of the parameters, that can be explained by the existence of stars within the sample of quasars with high proper motions (up to ± 40 mas yr^{-1} at least in one of the components) and parallaxes (up to 35 mas). A more detailed astrometric analysis

[8] This limit corresponds to $r \approx 21^m$ for quasars at $z \leq 3$, (Proft & Wambsganss 2015)

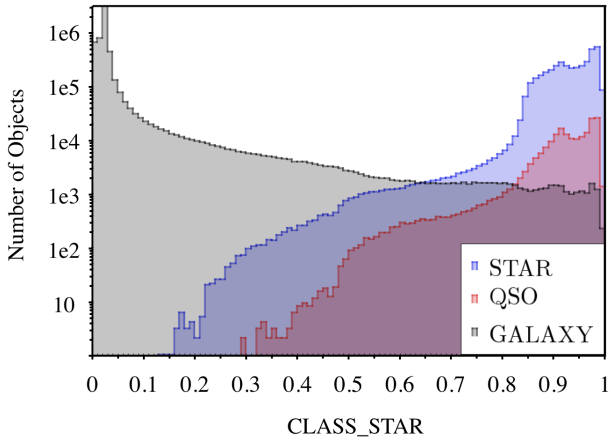

Fig. 8: Distribution of CLASS_STAR parameter for the sources from KiDS DR4, classified as galaxies (black), quasars (red) and stars (blue).

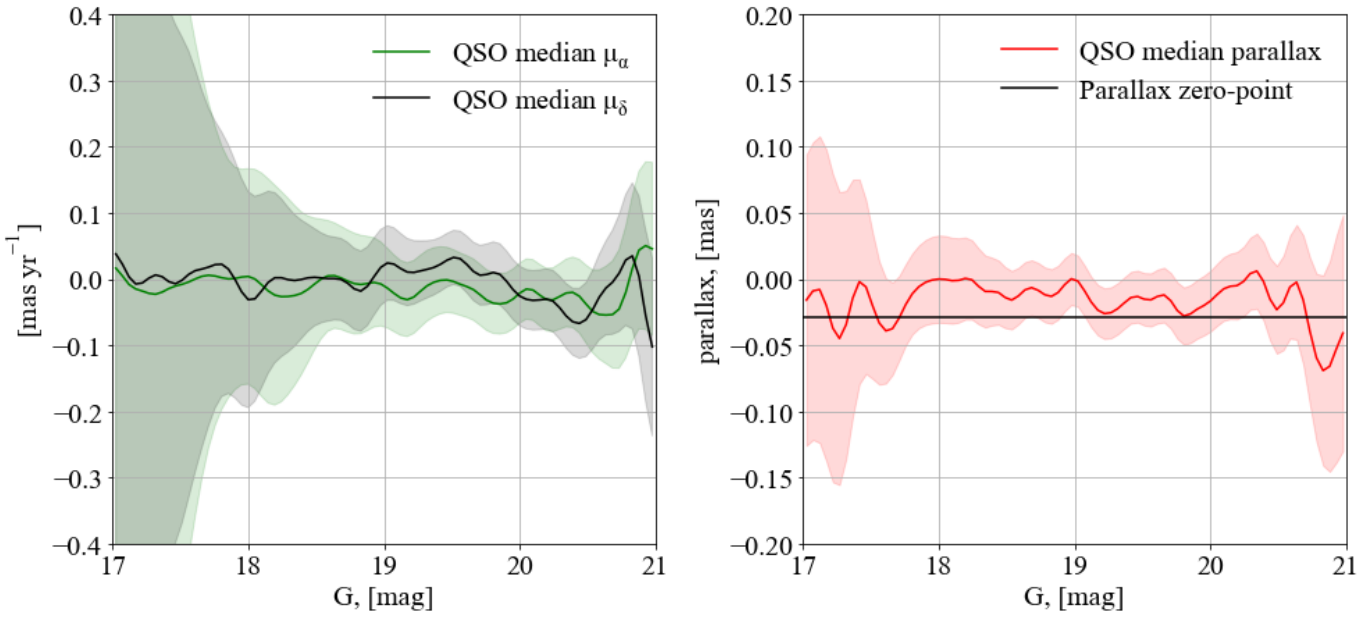

Fig. 9: Median right ascension (green curve, *left*) and declination (black curve, *left*) proper motion components, and parallaxes (*right*) for the KiDS DR4 QSO sample as a functions of G magnitude. The colored areas represent the standard deviation of the mean σ/N, where σ is the standard deviation and N is the number of quasars in each bin. The black line in the *right* plot represents the parallax zero-point (equals to −0.029 mas, Lindegren et al. 2018).

Table 2: Basic statistics of astrometric parameters for KiDS DR4 QSO, cross-matched with *Gaia* DR2

Parameter	Mean	Median	Standard deviation
$\bar{\omega}$, [mas]	-0.010	-0.014	1.125
μ_α, [mas yr^{-1}]	-0.028	-0.018	2.104
μ_δ, [mas yr^{-1}]	-0.104	-0.005	2.005

providing a more quantitative estimation of the rate of contaminating stars, cannot be produced without accurate modeling and involving another external datasets, which goes beyond the purposes of this paper.

In order to add something for the galaxies, we use the very simple argument that, by construction, *Gaia* should contains no galaxies at all (Robin et al. 2012). Thus all of the objects with high $p_{\texttt{GALAXY}}$ should not have a match in *Gaia* DR2. This is, of course, only a rough approximation since there might be a number of galaxies that *Gaia* still measures, as for instance, objects with bright cores.

Among the ≈25 000 GALAXY with a match in *Gaia*, we note that only 1 784 have CLASS_STAR> 0.5, thus they can be point-like sources in KiDS, that our algorithm misclassified, or very compact galaxies below the KiDS resolution.

In Figure 8 we show the distribution of the CLASS_STAR parameter for each class of full KiDS DR4 catalogue. Assuming that galaxies are all extended objects, we would expect to find in KiDS no objects classified as GALAXY with CLASS_STAR> 0.5.

However, there are objects that are point-like, according to their CLASS_STAR value, but have been classified as GALAXY by our algorithm. The number of point-like galaxies from Figure 8 is larger than a couple of thousands, as predicted by the cross-match with *Gaia*. This slight disagreement might be explained by the better resolution of *Gaia* (Krone-Martins et al. 2018): these sources might be seen as point-like in KiDS, but are extended and thus not identified in *Gaia*.

Despite this, the majority of GALAXY sources with a *Gaia* match are indeed extended objects in KiDS, or sources near by a bright star, as we directly verified on a random sample of ≈5000 objects, via the SDSS DR14 Navigate Tool[9], and then also checking the KiDS r-band images, finding, for most of the sources, bright features (e.g., cores, regions in arms, etc.), that could be resolved only for galaxies with significant angular size.

4.2. Validation of quasars with mid-infrared data from WISE

Using mid-infrared (MIR) colours is a very effective way to separate quasars from stars and passive galaxies. In fact, unlike stars and inactive galaxies, that show approximately zero MIR colors, the emission of AGNs conforms to the power-law emission in MIR wavelength range, that causes higher red MIR colours (Elvis et al. 1994; Stern et al. 2005; Assef et al. 2013).

As largely demonstrated by a number of published works, including Paper I, by using a combination of infrared color and magnitudes cuts, it is possible to separate quasars from stars and galaxies (e.g. the two-color criteria of Lacy et al. (2004), Stern et al. (2005), and Donley et al. (2012) with *Spitzer* (Werner et al. 2004) data; or the two-color criteria in Jarrett et al. (2011) and Mateos et al. (2012) or the one-color criteria of Stern et al. (2012) and Assef et al. (2013) using data from the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010).

Here, we decided to use the single infrared one-colour cut: [3.6]μm-[4.5]μm> 0.8 proposed by Stern et al. (2012), using data from WISE, the NASA space mission, aimed to map all sky in 4 MIR bands: $W1$, $W2$, $W3$, $W4$ (3.6,4.5,12 and 22 μm respectively). This criterion can separate quasars with resulting purity of ≈ 95%, but allows one to select quasars only up to $z \approx 3.5$ (Guo et al. 2018). We caution the reader that a given sample selected with this criterion can be contaminated by brown dwarfs,

[9] http://skyserver.sdss.org/dr14/en/tools/chart/navi.aspx

that have similar colours. The more elaborated two-colour criterion of Mateos et al. (2012) allows to reduce this contamination, but it requires reliable measurements in the [12]μm band, which would significantly decreases the total number of matched sources in our case.

We note that, in general, it is harder to validate the purity of galaxies in the same way since stars overlap with (non-active) galaxies in this dimension (see, e.g. Fig. 12 in Wright et al. 2010).

Finally, we clarify to the reader that in this paper, the WISE data is only used as validation for the quasars catalogue but not for the lenses search. In fact, in Paper I, we highlighted that the bottle-neck of our search was indeed the too severe colour WISE pre-selection. This could be caused by the fact that, in case the lens and the source are blended in WISE and the deflector gives a large contribution to the light, the colours of this *effective* source may be not quasar-like anymore and move indeed toward lower $W1-W2$ values. Here we rely on a much solid and trustable way to classify objects, our ML based classifier, and thus we do not need to apply any cut nor we need to require a match with WISE to build our candidate list.

We cross-matched the SDSS training sample as well as the catalogue of all the classified objects (Section 2) with the AllWISE (Cutri et al. 2013) data release using a 2″.0 radius. The resulting sample consists of 114 773 quasars, 3 289 858 galaxies, and 2 020 768 stars for the classified objects and of 8 879 quasars, 78 816 galaxies, and 13 249 stars for the training data.

Figure 10 shows the histograms of distribution of the $W1-W2$ color for the KIDS DR4 objects classified in the three classes (left panels, solid lines), and for the corresponding training sample (right panel, dotted lines), color coded by their classification: red for QSO, black for GALAXY and blue for STARS. In general, the distribution of the full catalogue shows a similar distribution to that of the training sample, with the peak of the QSO subsample shifted toward larger $W1-W2$ values, as expected. We note, however, that for the GALAXY and STAR classes, the distribution of the full catalogue is is much broader than the distribution of the corresponding training sample, especially towards larger values, both in negative and in positive. This might indicate a lower pureness for these families and consequently a larger contamination level in the QSO family or a lack of particular class of families (e.g. active galaxies) in the training sample.

As we show in the next subsection, the pureness of the objects classified as GALAXY seems to be quite high, according to the external validation of this class performed via a cross-match with the Galaxy And Mass Assembly Survey Data Release 3 (GAMA DR3, Baldry et al. 2018).

More and deeper investigations will be performed on pureness and completeness in the forthcoming paper of the KiDS-SQuaD series. Nevertheless, for the purposes of this paper, we are confident, and we will prove in Section 5, that our automatic classifier allowed us to obtain a starting catalogue of quasars and galaxy with a stellar contamination much smaller than the one obtained in Paper I, where we rely on simple and manual optical and infrared color-cuts.

4.3. Validation of galaxies with GAMA

To validate the pureness and completeness of the subsample of galaxies within the BEXGO catalogue, we cross-match the final catalogue of classified object with spectroscopically confirmed galaxies from GAMA.

In particular, following the suggestions given on the GAMA website, we retrieved all the objects[10] with redshifts $0.05 < z < 0.9$ and with a high "normalised" redshift quality ($nQ > 1$). We matched these $\approx$ 208k sources with our final catalogue of classified objects from KiDS, finding 105 334 systems in common. Among these 105 018 were indeed classified as GALAXY from CatBoost and 104 970 have a $p_{\texttt{GALAXY}} \geq 0.8$. Thus, only the 0.3% of the common objects have been misclassified (123 as STARS and 181 QSO).

Although we are aware that this test is not definitive and that it is not straightforward to directly translate the relative number of contaminant into a percentage of pureness of the final galaxy catalogue, it shows that, at least for this small but representative sample og galaxies, our CatBoost classifiers does a good job.

We speculate that one of the reasons for slight disagreement between the distribution of galaxies from the training sample and galaxies classified as such in the BEXGO catalogue in the $W1-W2$ space might arise from the fact that, although we limited the analysis and classification to only objects brighter than $r < 22^m$, the SDSS galaxies are generally more luminous than the KiDS ones.

We stress again that our final purpose is to create an authomatic and effective method to build a catalogue of extragalactic objects, with the smallest possible contamination from stars, that is the first necessary step to search for strong gravitational lensed quasars. We believe that these three validation steps with external data demonstrated that we succeeded in our goal and thus we can now use the newly created catalogue to search for lens candidates.

5. Searching for gravitationally lensed quasars

Strong gravitationally lensed quasars are valuable but very rare objects (according to Oguri & Marshall 2010, one quasars in $\sim 10^{3.5}$ is expected to be strongly lensed for i-band limiting magnitude deeper than $i \approx 21^m$, see e.g. their Fig. 3 and Sec. 3.1) that give direct, purely gravitational probes of cosmology and extragalactic astrophysics.

Generally speaking, we can separate lensed quasars in three families: systems where the quasar images dominate (mainly low-separation couples/quadruplets with a faint deflector in between), objects were the deflector is a bright, usually red, galaxy that dominates the light budget of the system, and finally systems where both lens and source give a non-negligible contribution to the light. In the last two cases, CatBoost will most probably return multiple matches of which at least one will be classified as extended (GALAXY), while in the former one it will classify them as QSO. However, we note that in cases where the separation between the quasar components is too low, the objects might be not resolved in the KiDS catalogue, and thus result in a single match/classification from our algorithm. Indeed, most of the known gravitationally lensed quasars with low separation between the multiple images, discovered in the SDSS, are identified as galaxies (only in few cases as single quasar), since the poor resolution does not allow deblending. Of course the better image resolution of KiDS helps in this case, however some lenses with very low separation are blended also in KiDS. This is why it is of crucial importance to have a catalogue of extragalactic objects as clean as possible from stellar contamination and as complete and efficient as possible in classifying galaxies and quasars. To demonstrate our statement that lensed

[10] also the ones observed by other surveys, i.e. we queried the table 'SpecAll'

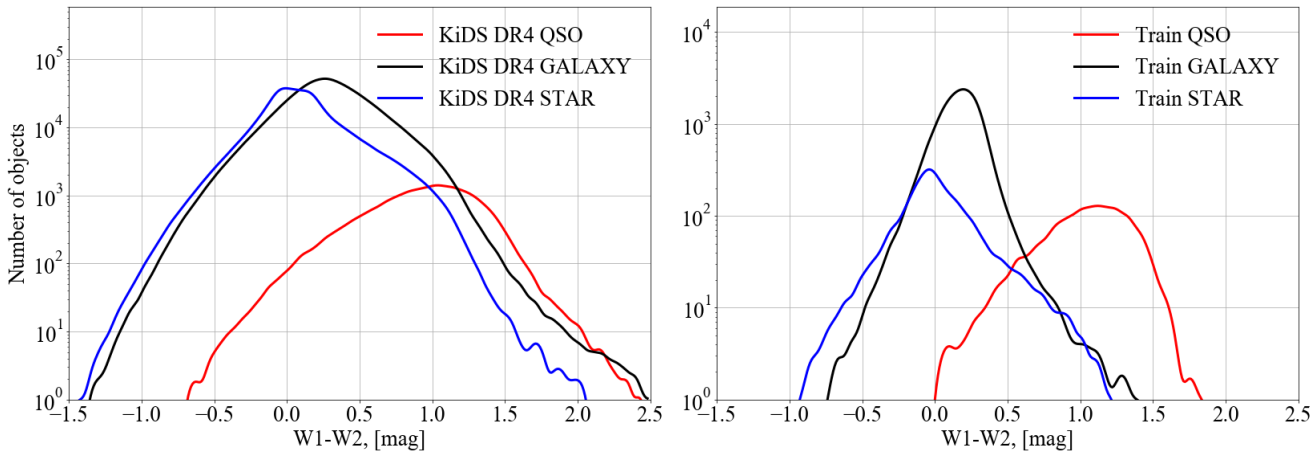

Fig. 10: Distribution of the $W1 - W2$ colour for the classified KiDS DR4 sources of different classes (*left* plot, *solid* lines, *red* for QSO, *black* for GALAXY, *blue* for STAR) versus their train samples (*right* plot, *dotted*, plotted with the same color-scheme). A fair agreement can be found between corresponding classes, although the distribution for the full catalogues is generally broader than the one of the training samples. The peak of the QSO distribution is shifted towards larger $W1 - W2$ with respect to the GALAXY and STARS, as expected.

quasars are not always classified as (multiples, near-by) QSO by ML based algorithms that work in a magnitude-color space, and at the same time to highlight the importance of having an extragalactic catalogue, we carried on a test on the recovery of known lenses, as already done in Paper I.

We started from the same list of $\approx$ 260 confirmed lensed quasars that we used in Paper I, collected from the CfA-Arizona Space Telescope LEns Survey (CASTLES, Muñoz et al. 1998) Project database, the SDSS Quasar Lens Search (SQLS, Inada et al. 2012) and updated it with systems recently discovered in wide-sky surveys (Agnello et al. 2018,b; Ostrovski et al. 2018; Anguita et al. 2018; Lemon et al. 2018; Spiniello et al. 2019b). The cross-match between the full KIDS DR4 catalogue of objects with $r < 22^m$ and the list of known lensed quasars (288 systems), gave us 17 known lenses. All of them have been retrieved in the KiDS-BEXGO catalogue, 10 classified as QSO and 7 as GALAXY (one of them with a $p_{\rm QSO} \approx 0.45$). These lenses are reported in Table 3, together with the probability to belong to each class. We do not explicitly report their right ascensions and declinations in separate columns because their ID already contain the J2000 coordinates in the usual 'hhmmss.ss±ddmmss.ss' format.

Based on this simple and qualitative test, it appears clear that selecting only quasars would allow one to find only lens systems were the contribution to the light from the quasars is much larger than the contribution of the deflector (selecting only QSO we would retrieve roughly 65% of the known lensed quasar population – 11 over 17 systems).

Finally, although this goes behind the scope of Paper II, we note that another important advantage of having an extragalactic source catalogue (rather than only quasars) is the possibility to search for galaxy-galaxy gravitational lenses

Such type of gravitationally lensed objects allow to investigate in great details the mass distribution in massive galaxies upt to $z \sim 1$, especially when combined with dynamics (Koopmans et al. 2006, 2009; Spiniello et al. 2011, 2015). Morphological and photometric criteria can be used to find this kind of lenses: one should look for red extended objects (GALAXY with red colors), with the presence of blue extended objects (GALAXY with blue colors) within small circular apertures. We will work in this direction in a forthcoming paper, possibly using authomatic, machine learning based routines to this scope (e.g. Petrillo et al. 2017, 2019a,b) and already available catalogues of Luminous red galaxies in the Kilo Degree Survey (e.g., Vakili et al. 2019.

Starting from the KiDS-BEXGO catalogue of 5 880 276 objects, we retrieve only systems belonging to the following distinct groups:

1. QSO-Multiplets: sources classified as QSO and with at least one near-by QSO companion (within a 5″ circular aperture radius) with similar colors,
2. GALAXY-Multiplets: sources classified as GALAXY and surrounded by at least one object classified as QSO within a 5″ circular aperture radius[11]

This simple procedure allowed us to obtain 347 unique objects for the first group and 611 unique objects belonging to the second one, which we then visually inspected. Among these, some where already known lenses, some are probably binary quasars and some are simply contaminants appearing asclose-by companions because of sky projection. Nevertheless, we found many very promising lens candidates, that two people of our team graded from 0 to 4, with 4 being a sure lens. We present the 12 candidates with grade $\geq$ 2.5 in Table 4 (divided into the two Multiplets kinds). We publicly release their coordinates to facilitate spectroscopic follow-up, which is the last necessary step for the unambiguous confirmation. Finally, the *gri*-combined KiDS cutouts of these 12 candidates are shown in Figure. 11. The first top rows show candidates belonging to the GALAXY-Multiplets family while the bottom row show QSO-Multiplets candidates. In the former group, the deflector give a much larger contribution to the light, as can be seen from the images. KiDSJ0008-3237 seem to be a very reliable galaxy-galaxy candidate, while KiDSJ0215-2909, definitively among the most promising objects, might be a fold-quadruplet, similar to the one recently found in the VST-ATLAS Survey, WISE 025942.9-163543 (Schechter et al. 2018). and very useful for cosmography studies (time-delay measurement of H_0, see e.g.'The H0 Lenses in COSMOGRAIL's Wellspring'[12] results).

[11] The choice of a 5″ circular aperture radius is motivated by the average separation of all the known lenses.

[12] https://shsuyu.github.io/H0LiCOW/site/

Fig. 11: The 12 best candidates, both of GAL-Multiplets type (first two rows) and of QSO-Multiplets type (bottom row). The cutouts show *gri* (for the moment ugr) combined KiDS images of $10'' \times 10''$ in size. The coordinates of the candidates are given in Table 4.

We note that, of the 17 known lenses, only 8 have been selected as multiplets (2 as QSO-Multiplets and 6 as GALAXY-Multiplets). The other 9 systems have not been deblended in the KiDS catalogue, and thus only have one single match in our classification scheme. These numbers are perfectly in line with the results obtained in Paper I where we found that the Multiplet method alone allowed the recovery of $\sim 40\%$ of the known lenses population. In a forthcoming paper of this series, fully dedicated to the lens search, we will perform a more careful candidate selection, also based on improved color and magnitude criteria to select objects with similar colors and applying to the full BEXGO catalogue the Blue and Red Offsets of Quasars and Extragalactic Sources (BaROQuES) scripts, already successful tested in Paper I.

We finally note that we re-discovered a very promising quadruplet: KIDS0239-3211 that was presented in a AAS research note (Sergeyev et al. 2018) and it was found by the first application of our ML based classifier[13]. The same system has also been detected Hartley et al. 2017 using image-based Support Vector Machine classifier and by Petrillo et al. 2019b using Convolutional Neural Networks; but since they do not released the coordinates in their paper, we re-discover it in a completely independent way.

Finally, we cross-match the list of all the lens candidates found in Paper I with the BEXGO catalogue. We find that among the 210 objects we have found in Spiniello et al. 2018, 148 are recovered in the extragalactic objects catalogue ($\approx 45\%$ classified as `QSO` and $\approx 55\%$ as `GALAXY`) and 66 are also selected as Multiplets. Of the 62 remaining objects, 33 have $r > 22^m$ and therefore were discarded at the input catalogue creation stage, and 29 were classified as `STAR` by CatBoost; these 29 stars indeed also have a match in *Gaia*, and all of them have non-negligible proper motions and parallaxes. Finally, among the DR3 candidates that have not been found in the DR4 KiDS-BEXGO, 4 have been spectroscopically followed-up and turned out to be stars[14]. These numbers nicely demonstrate that the employment of a ML based classifier further help in decreasing the risk of stellar contamination within gravitationally lensed quasar candidates.

Of the seven known lenses that we recovered in Paper I, six are still recovered. We only lose the nearly Identical Quasar (NIQ) couple QJ0240-343 (Tinney 1995; Tinney et al. 1997) be-

[13] We used in that case a Random Forests classifier, trained with spectroscopically confirmed objects from SDSS DR14.

[14] We have already started a spectroscopic follow-up campaign using different facilities (e.g. the NTT@La Silla, the SALT@Suthernland Observatory, the LBT@Mt. Graham). The detailed results will be presented in forthcoming dedicated papers.

ID	$p_{\texttt{STAR}}$	$p_{\texttt{QSO}}$	$p_{\texttt{GALAXY}}$
J004941.90-275225.87	2.0E-5	1.0E-5	**0.9999**
J033238.22-275653.32	2.0E-5	1.0E-5	**0.9999**
J115252.26+004733.11	2.0E-5	1.0E-5	**0.9999**
J220132.76-320144.73	0.0004	4.0E-5	**0.9999**
J234416.95-305625.98	0.0004	0.0012	**0.9983**
J105644.89-005933.34	7.0E-4	**0.9978**	0.0015
J112320.73+013747.53	0.0086	**0.9829**	0.0085
J142758.89-012130.31	0.0035	**0.9831**	0.0134
J025257.87-324908.65	0.0011	**0.9950**	0.0039
J145847.59-020205.87	0.0004	0.0011	**0.9985**
J145847.66-020204.86	2.0E-5	3.0E-5	**0.9999**
J032606.87-312254.21	0.0019	**0.9944**	0.0037
J032606.78-312253.52	0.0023	**0.9793**	0.0184
J143228.96-010613.51	0.0006	**0.9980**	0.0014
J143229.25-010615.98	0.0008	**0.9966**	0.0025
J104237.27+002301.42	0.0477	**0.8637**	0.0886
J104237.24+002302.76	0.0652	**0.7721**	0.1627
J092455.82+021923.69	0.0078	**0.8909**	0.1012
J092455.82+021925.30	0.0059	0.4543	**0.5397**
J122608.10-000602.31	0.0046	**0.9610**	0.0343
J122608.03-000602.25	0.0199	**0.5048**	0.4752
J122608.13-000559.09	0.0009	0.0016	**0.9997**
J133534.80+011805.61	0.0128	**0.9806**	0.0066
J133534.87+011804.45	0.0056	**0.9797**	0.0066
J133534.97+011809.32	0.0013	0.0050	**0.9937**
J152720.14+014139.66	0.0058	**0.9617**	0.0325
J152720.27+014140.96	0.0005	0.0006	**0.9999**

Table 3: Known lenses in the KIDS DR4 footprints. All of them are recovered in our extragalactic catalogue, 8 have multiple matches. For the single ones (upper 'block'), 5 are classified as `GALAXY` and 4 as `QSO` (we highlight in bold the highest probability). For the multiple matches, half of the time all the components belong to the same family (middle 'block') and the other half, they belong to different families. We report for each component of each system the J2000 coordinates in the ID column (in the usual 'hhmmss.ss±ddmmss.ss' format), and the probability to belong to each of the three classification families.

hind the Fornax dwarf spheroidal galaxy, once again because it has r mag of $r = 22.17^m$ and thus it does not satisfied our initial conditions.

6. Results and Conclusions

In this second paper of the KiDS Strongly lensed QUAsar Detection Project (KiDS-SQuaD) we have presented a new machine learning based classifier to identify extragalactic objects in order to find lensed quasars within the KiDS DR4 data. The technique adopted in this paper became quite standard in the extragalactic community to classify objects in multi-band photometric surveys (Gieseke et al. 2011; Kovács & Szapudi 2012; Brescia et al. 2015; Carrasco et al. 2015; Peters et al. 2015; Krakowski et al. 2016, 2018; Viquar et al. 2018; Khramtsov & Akhmetov 2018; Barrientos et al. 2018; Nolte et al. 2019; Bai et al. 2019), which provide a very large amount of data, and have been already tested on the KiDS DR3 (Nakoneczny et al. 2019).

In fact, Nakoneczny et al. (2019) presented a ML based pipeline that allowed them to classify objects into three classes (stars, galaxies and quasars) and successfully applied it to the KiDS DR3. Our work, although extending from their findings, has been developed within a different framework, i.e. the search of lensed quasars, and it therefore differs from N19 in many aspects, from the assumption that quasars are point-like source, to the cleaning procedure, optimization, and fine-tuning aimed at minimize as much as possible the stellar contamination in the catalogue of extragalactic objects. Finally, here we also add infrared data, using deep photometry in 9-bands (instead of four), which further helps in isolating stars.

6.1. Summary

We provide here a general summary of the archived results of this paper, highlighting with bullet points the main steps that we undertook from the presentation of a new pipeline to the search of gravitationally lensed quasars. In particular, we have:

- used the full potential of machine learning methods on broad optical-infrared photometry data, after having applied a careful cleaning on the training SDSS×KiDS sample, also visually inspecting the ambiguous cases when necessary;
- performed an ad-hoc customization and fine-tuning of the parameters of the CatBoost algorithm, which we identified as the best possible classifiers for our purposes, to reach the required levels for purity and completeness and to avoid overfitting poroblems. We also implemented a weighting procedure, that allowed us to reach the best possible purity of quasars (decreasing the rate of stars, classified as quasars, from 0.6% to 0.3%);
- splitted the training dataset into a hold-out and out-of-fold part to asses the performance of our classifier in terms of completeness and pureness;
- defined (and then solved) a 3-class problem (`STAR`, `GALAXY`, `QSO`), working with a simple basic assumption made for the classification, namely that quasars and stars are point-like sources while galaxies are extended. We therefore used the `CLASS_STAR` parameter – a 'stellarity' index from KiDS catalogue – which turned out to be the most important feature in our classification algorithm (as in N19), together with optical and infrared colors;
- applied CatBoost on all the data from KiDS DR4 with magnitude brighter than $r = 22^m$. For each source, the classifier calculated the probability of belonging to the three different classes of objects: $p_{\texttt{STAR}}$, $p_{\texttt{GALAXY}}$, $p_{\texttt{QSO}}$, and then we assumed that a source belongs to a given class when the probability of being in that class is the highest;
- studied the variation in completeness and pureness as a function of the probability threshold used to assign an object to a given class;

ID	RA (hh:mm:ss.ss)	DEC (±dd:mm:ss.ss)	Multiplet type	Num. of matches	Separation (arcsec)	Notes
KIDSJ0008-3237	00:08:16.01	-32:37:15.80	GAL	3	2.7	Gravitational arc
KIDSJ0106-2917	01:06:49.80	-29:17:12.20	GAL	2	3.4	Double with large shear
KIDSJ0206-2855	02:06:30.86	-28:55:42.22	GAL	2	1.2	Low-separation double
KiDSJ0208-3203	02:08:53.16	-32:02:03.51	GAL	3	2.0	Cross Quad candidate
KIDSJ0215-2909	02:15:14.4	-29:09:25.6	GAL	3	3.2	Fold Quad candidate
KIDSJ1204+0034	12:04:56.58	+00:34:06.02	GAL	3	4.6	Large-separation double
KiDSJ1346+0017	13:46:12.38	+00:17:20.18	GAL	4	2.8	Double with large shear
KIDSJ1359+0129	13:59:43.98	+01:28:13.90	GAL	2	0.9	Low-separation double
KIDSJ0000-3502	00:00:57.10	-35:02:54.15	QSO	2	1.0	Low-separation double
KIDSJ0139-3103	01:39:59.08	-31:03:35.06	QSO	2	1.6	Low-separation double
KIDSJ0201-3208	02:01:15.44	-32:08:34.57	QSO	2	1.6	Low-separation double
KIDSJ1334-0120	13:34:11.18	-01:20:52.22	QSO	2	1.1	Low-separation double

Table 4: The most reliable gravitationally lensed quasar candidates found with the multiplet method described in the text. Cutouts of the candidates are shown in Figure 11. We highlight in the table the number of matches found for each system in the BEXGO catalogue. In groups where the number of detection is smaller than the total number of objects visible from the cutout, most probably these objects are fainter than the magnitude threshold we set for the input catalog. We double checked that these missing matches were not objects classified as stars. The last column indicates the separation between the multiple QSO images.

- collected all the objects that were not classified as stars, building the KiDS DR4 Bright EXtraGalactic Objects catalogue (KiDS-BEXGO), which we then also validated using external data (*Gaia* DR2, AllWISE and GAMA);
- showed the potential of the KiDS-BEXGO catalogue in the gravitationally lensed quasar search, with a simple test on the recovery of known, confirmed lenses, and proved, in this way, that our method of selecting extragalactic sources (not only quasars) is a necessary condition to discover as many new systems as possible;
- used the KiDS-BEXGO catalogue to search for new, undiscovered gravitationally lensed quasars, looking for objects with a near-by companion. We have obtained a list of 958 'Multiplets' (347 `QSO` and 611 `GALAXY`) that we visually inspected, finding 12 very reliable lens candidates for which we release coordinates and KiDS images;
- showed the improvement, in terms of stellar contaminants in the final candidate list with respect to what obtained in in Paper I, but at the same time also demonstrated the need for different methods to search for lenses candidates within the catalogue (e.g. the BaROQuES) and directly analyzing images (DIA). These methods will be investigate in a forthcoming publication.

In addition, we present in the Appendix A a direct comparison of some of the most used classifiers based on decision trees. This test helped us to compare and quantify the performance of each of them on the same training sample in order to choose the most suitable one for our purposes, namely CatBoost.

6.2. Future perspectives and improvements

From the predictions of Oguri & Marshall (2010), we estimated that $\approx$ 50 lensed quasars are expected in the KIDS DR4 (1000 deg^2), when limiting to systems with $r < 22^m$; 17 lenses are already known, thus, in principle, more than half are still undiscovered (and even more going to fainter magnitudes).

In this Paper II, we focused on the first necessary step to find all the catchable gravitational lenses: an object classifier, that allowed us to get rid of the very numerous stellar contaminants and will allow us to analyze with a minimum human intervention very large datasets. We note that our classifier is build and trained for this specific purpose. A forthcoming paper within the KiDS consortium (Nakoneczny et al., in prep.) will present a machine learning based pipeline trained for general scientific purposes, providing photometric redshifts for galaxies and quasars in KiDS DR4 on top of the objects classification, and testing machine learning extrapolation to increase catalog completeness on fainter magnitudes.

Moreover, we also plan to further improve the classification model, working in a more complex and complete feature space and developping a more detailed classification scheme (e.g., spitting the classification of galaxies on late and early types, giving that massive early types are more likely acting as deflectors because on average more massive).

In Paper III, already in preparation (Sergeyev et al., in prep), we focus instead only on the gravitationally lens search, presenting a more systematic way, as automatic as possible, to select reliable candidates from the KiDS-BEXGO catalogue. We will apply photometric and morphological criteria, e.g. based on optical and infrared color, or on the simple fact that a centroid offsets of the same object among different surveys, covering different bands is expected since the deflector and quasar images contribute differently in different wavelength ranges (BaROQuES). We will also exploit the full potential of the Direct Image Analysis (DIA, see Paper I for more details) to get precise astrometry and fit the photometry of our most reliable candidates.

Finally, we already started the necessary spectroscopic follow-up, to get a final, unambiguous confirmation of the lensing nature for as many systems as possible, and to obtain secure redshift measurements that will allow us translate the lens model results (e.g., Einstein radii) into physical mass measurements.

Acknowledgements. The authors wish to thank Maciej Bilicki for the interesting discussion and the very constructive comments that helped in improving the final manuscript. CS has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie actions grant agreement No 664931. CT acknowledges funding from the INAF PRIN-SKA 2017 program 1.05.01.88.04. KK acknowledges support by the Alexander von Humboldt Foundation. JTAdJ is supported by the Netherlands Organisation for Scientific Research (NWO) through grant 621.016.402. HYS acknowledges

the support from the Shanghai Committee of Science and Technology grant No. 19ZR1466600. This work is based on data products from observations made with ESO Telescopes at the La Silla or Paranal Observatories under programme ID(s) 177.A-3016(A), 177.A-3016(B), 177.A-3016(C), 177.A-3016(D), 177.A-3016(E), 177.A-3016(F), 177.A-3016(G), 177.A-3016(H), 177.A-3016(I), 177.A-3016(J), 177.A-3016(K), 177.A-3016(L), 177.A-3016(M), 177.A-3016(N), 177.A-3016(O), 177.A-3016(P), 177.A-3016(Q), 177.A-3016(S), 177.A-3017(A), 177.A-3018(A), 060.A-9038(A), 094.B-0512(A), and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD and the KiDS production team, on behalf of the KiDS consortium. OmegaCEN and the KiDS production team acknowledge support by NOVA and NWO-M grants. Members of INAF-OAPD and INAF-OACN also acknowledge the support from the Department of Physics & Astronomy of the University of Padova, and of the Department of Physics of Univ. Federico II (Naples). This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration. This publication makes use of data products from the Sloan Digital Sky Survey IV. Funding for SDSS IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This work has made use of data from Galaxy and MAss Assemby Survey. GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is http://www.gama-survey.org/.

References

Abolfathi B., Aguado, D., Aguilar, G., et al. 2018, ApJS, 235, 42
Abraham, S., Philip, N., Kembhavi, A., Wadadekar, Y. G., & Sinha, R. 2012,MNRAS, 419, 80
Agnello A., Treu, T., Ostrovski, F., et al. 2015,MNRAS, 454, 1260
Agnello A., Kelly B. C., Treu T., Marshall P. J., 2015,MNRAS, 448, 1446
Agnello, A., Schechter, P. L., Morgan, N. D., et al. 2018, MNRAS, 475, 2086
Agnello A., Lin, H., Kuropatkin, N., et al. 2018,MNRAS, 479, 4345
Agnello, A., & Spiniello, C. 2018, arXiv:1805.11103
Akhmetov, V., Fedorov, P., Velichko, A., & Shulga, V. 2017, MNRAS, 469, 763
Amaro, V., Cavuoti, S., Brescia, M., et al. 2019,MNRAS, 482, 3, 3116
Anguita, T., Schmidt, R. W., Turner, E. L., et al. 2008, A&A, 480, 327
Anguita T., Schechter, P.L., Kuropatkin, N. et al. 2018,MNRAS, 480, 5017
Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, ApJ, 772, 26
Bachchan, R. K., Hobbs, D. & Lindegren, L. 2016, A&A, 589, A71
Bai, Y., Liu, J., Wang, S., & Yang, F. 2019, AJ, 157, 9
Baldry I. K., et al., 2018,MNRAS, 474, 3875
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Barrientos, F., Pichara, K., Troncoso, P., et al. 2018, VST in the Era of the Large Sky Surveys, 9
Bate, N. F., Floyd, D. J. E., Webster, R. L., & Wyithe, J. S. B. 2011, ApJ, 731, 71
Blanton, M., Bershady, M., Abolfathi, B., et al. 2017, AJ, 154, 28
Bilicki, M., Hoekstra, H., Brown, M., et al. 2018, A&A, 616, A69, 22
Braibant, L., Hutsemékers, D., Sluse, D., Anguita, T., & García-Vergara, C. J. 2014, A&A, 565, L11
Breiman L. 2001, Machine Learning, 45, 5
Brescia, M., Cavuoti, S., & Longo, G. 2015, MNRAS, 450, 3893
Cabanac R. A., et al., 2007, A&A, 461, 813
Carrasco D., Barrientos, L., Pichara, K., et al. 2015, A&A, 584, A44
Capaccioli, M. & Schipani, P. 2011, The Messenger, 146, 2
Capaccioli, M., Schipani, P., de Paris, G., et al. 2012, in Science from the Next Generation Imaging and Spectroscopic Surveys, 1
Chen, T. & Guestrin, C. 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785-794
Chiu, K., Richards, G., Hewett, P., and Maddox, N. 2007,MNRAS, 375, 1180
Cortes, C. & Vapnik, V. 1995, Mach. Learn., 20, 273
Cutri, R., Wright, E., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products, Tech. rep.
D'Isanto, A., Cavuoti, S., Gieseke, F., Polsterer, K. 2018, A&A, 616, A97, 21
de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2013, Exp. Astron., 35, 25
de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Ding, X., Treu, T., Suyu, S. H., et al. 2017,MNRAS, 465, 4634
Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, ApJ, 748, 142
Dorogush, A.V., Ershov, V., Gulin, A., arXiv:1810.11363
Driver S. P., et al., 2009, A&G, 50, 5.12
Duda, R. O. & Hart, P. E. 1973, Pattern classification and scene analysis (J. Wiley & Sons)
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32
Elting, C., Bailer-Jones, C. A. L., & Smith, K. W. 2008, American Institute of Physics Conference Series, 1082, 9
Elvis, M., Wilkes, B. J., McDowell, J. C., et al. 1994, ApJS,95, 1
Eyer L., Blake C., 2005,MNRAS, 358, 30
Friedman, J.H. 2000, Annals of Statistics, 29, 1189
Gaia Collaboration, (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration, (Brown, A., et al.) 2018, A&A, 616, A1
Gaia Collaboration, (Mignard, F., et al.) 2016, A&A, 616, A14
Gieseke, F., Polsterer, K. L., Thom, A., et al. 2011, arXiv:1108.4696
Guerras, E., Mediavilla, E., Jimenez-Vicente, J., et al. 2013, ApJ, 778, 123
Guo, S., Qi, Z., Liao, S. et al. 2018, A&A, 618, A144
Hartley P., Flamary R., Jackson N., Tagore A. S., Metcalf R. B., 2017,MNRAS, 471, 3378
Haehnelt, M. G., & Kauffmann, G. 2000, MNRAS, 318, L35
Hernitschek N., Schlafly, E., Branimir, S., et al. 2016, ApJ, 817, 73
Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Springel V., 2006, ApJS, 163, 50
Inada N., Oguri, M., Shin, M.-S., et al., 2012, AJ, 143, 119
Ivezic, Ž., & LSST Science Collaboration 2013, LSST Science Requirements Document, http://ls.st/LPM-17
Jacobs C., et al., 2019,MNRAS, 484, 5330
Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, ApJ, 735, 112
Jin, X., Zhang, Y, Zhang, J. 2019,MNRAS, 485, 4539
Kauffmann, G., & Haehnelt, M. 2000, MNRAS, 311, 576
Khramtsov, V., & Akhmetov, V. 2018, Proceedings of a IEEE XIIIth International Scientific and Technical Conference "CSIT", 2018, 72
Khramtsov, V., Akhmetov, V., & Fedorov, P. 2018b, A&A, under review
Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68
Kovács, A., & Szapudi, I. 2015, MNRAS, 448, 1305
Krakowski, T., Małek, K., Bilicki, M., et al. 2016, A&A, 596, A39
Krakowski T., Małek K., Bilicki M., Siudek M., Pollo A., 2018, pas7.conf, 7, 252
Kochanek C. S., 2004, ApJ, 605, 58
Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006, ApJ, 649, 599
Koopmans L. V. E., Bolton, A., Treu, T., et al. 2009, ApJ, 703, L51
Krone-Martins, A., Delchambre, L., Wertz, O., et al. 2018, A&A, 616, L11
Kuijken, K. 2008, A&A, 482, 1053
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015,MNRAS, 454,3500
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, in press
Lacy, M., Storrie-Lombardi, L. J., Sajina, A., et al. 2004, ApJS, 154, 166
Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018,MNRAS, 473, 3895
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Lindegren, L., Hernandez, J., Bombrun, A., et al. 2018 A&A, 616, A2
Lemon, C. A., Auger, M. W., McMahon, R. G., & Koposov, S. E. 2017, MNRAS, 472, 5023

Lemon, C., Auger, M., McMahon, R., & Ostrovski, F. 2018, MNRAS, 479, 5060
Mansour, Y. 1997, Proceedings of the 14th International Conference on Machine Learning, p.195-201
Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. 2012, MNRAS, 426, 3271
Matthews, B. 1975 , Biochimica et Biophysica Acta (BBA) - Protein Structure, 405, 2, p. 442-451.
Meylan, G., Jetzer, P., North, P., et al. 2006, Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro
Metcalf R. B., et al., 2018, arXiv, arXiv:1802.03609
Motta, V., Mediavilla, E., Falco, E., & Muñoz, J. A. 2012, ApJ, 755, 82
Muñoz J. A., Falco E. E., Kochanek C. S., Lehár J., McLeod B. A., Impey C. D., Rix H.-W., Peng C. Y., 1998, Ap&SS, 263, 51
Nakoneczny, S., Bilicki, M., Solarz, A. et al. 2019, A&A, 624, A13
Nolte, A., Wang, L., Bilicki, M., Holwerda, B., & Biehl, M. 2019, arXiv:1903.07749
Oguri M., et al., 2006, AJ, 132, 999
Oguri M., & Marshall P. J. 2010, MNRAS, 405, 2579
Oguri, M., Rusu, C. E., & Falco, E. E. 2014, MNRAS, 439, 2494
Ostrovski F., et al., 2017, MNRAS, 465, 4325
Ostrovski F., et al., 2018, MNRAS, 473, L116
Paraficz D., et al., 2016, A&A, 592, A75
Peters C. M., et al., 2015, ApJ, 811, 95
Petrillo C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
Petrillo C. E., Tortora, C., Chatterjee, S., et al., 2019a, MNRAS, 482, 807
Petrillo C. E., Tortora, C., Vernardos, G., et al., 2019b, MNRAS, 484, 3879
Proft, S. & Wambsganss, J. 2015, A&A, 574, A46
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin., A. 2018, in Advances in Neural Information Processing Systems, 31, 6638
Quinlan, J. R. 1986, Mach. Learn., 1, 81
Refsdal S., 1964, MNRAS, 128, 307
Richard J., et al., 2019, Msngr, 175, 50
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, (Cambridge, MA, USA: MIT Press), 318
Suyu S. H., Bonvin, V., Courbin, F., et al. 2017,MNRAS, 468, 2590
Wyithe, J. S. B., & Loeb, A. 2003, ApJ, 595, 614
Pâris I., Petitjean, P., Aubourg, É, et al. 2018, A&A, 613, A51
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. 2009, ApJS, 180, 67
Robin, A., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Rusu, C. E., Berghea, C. T., Fassnacht, C. D., et al. 2018, arXiv:1803.07175
Shankar, F., Weinberg, D. H., & Miralda-Escudé, J. 2009, ApJ, 690, 20
Shen, Y., Strauss, M. A., Ross, N. P., et al. 2009, ApJ, 697, 1656
Schechter, P. L., & Wambsganss, J. 2002, ApJ, 580, 685
Schechter P. L., Anguita T., Morgan N. D., Read M., Shanks T., 2018, RNAAS, 2, 21
Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2018, ApJ, 863, 144
Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2017, ApJ, 851, 13
Sergeyev A., Spiniello C., Khramtsov V., et al. 2018, RNAAS, 2, 189
Sluse, D., Schmidt, R., Courbin, F., et al. 2011, A&A, 528, A100
Spiniello C., Koopmans L. V. E., Trager S. C., Czoske O., Treu T., 2011,MNRAS, 417, 3000
Spiniello C., Koopmans L. V. E., Trager S. C., Barnabè M., Treu T., Czoske O., Vegetti S., Bolton A., 2015,MNRAS, 452, 2434
Spiniello, C., Agnello, A., Napolitano, N. R., et al. 2018, MNRAS, 480, 1163
Spiniello C., Sergeyev, A., Marchetti, L., et al. 2019,MNRAS.tmp..750S
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Suyu, S. H., Treu, T., Hilbert, S., et al. 2014, ApJ, 788, L35
Sutherland, W. 2012, Science from the Next Generation Imaging and Spectroscopic Surveys, 40
Treu T., Agnello A., Strides Team, 2015, AAS, 225, 318.04
Tinney, C. G. 1995, MNRAS, 277, 609
Tinney, C. G., Da Costa, G. S., & Zinnecker, H. 1997, MNRAS, 285, 111
Vakili M., et al., 2019,MNRAS, in press.
Vapnik, V. 1995, The nature of statistical learning theory (New York, USA: Springer-Verlag New York, Inc.)
Viquar, M., Basak, S., Dasgupta, A., Agrawal, S., & Saha, S. 2018, arXiv:1804.05051
Walsh D., Carswell R. F., Weymann R. J. 1979, Nature, 279, 381
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Wright, E. L., Eisenhardt, P.R.M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Wright, A.H., Hildebrandt, H., Kuijken, K., et al. 2018, arXiv:1812.06077

Appendix A: Testing different classifiers

The main result of the main body of the paper, is a catalogue of bright objects from KiDS DR4 that we classified into three families, `STAR`, `QSO`,`GALAXY`, using a machine learning based classifier, that uses the Gradient Boosting algorithm.

In order to choose the best possible algorithm for our purposes, we tested two different approaches (and three diffrent methods) based on ensembles of decision trees, namely the Gradient Boosting (GB, Friedman 2000) and Random Forest (RF, Breiman 2010) algorithms, which was the choice of N19. Here in this Appendix, we provide a more detailed description of the classifiers, their main characteristics, strength and weakness points to give to the reader a better understanding of the differences and similarities between them and to finally justify our final choice.

As already stated in the main body, the general classification problem can be simply explained considering a training dataset (D) with n samples and m features for each sample with defined label y_i: $D = \{\mathbf{x}_i, y_i\}$ where $i \in \{0, ..., n-1\}, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{N}$. The goal is then to create an approximation function $F : \mathbf{x} \rightarrow y$.

The two GB algorithms mentioned above follow two different approaches of ensemble learning, that could be propagated not only on the decision trees. We describe them in details in the following sections.

Appendix A.1: Gradient Boosting Decision Trees

XGBoost, Chen & Guestrin 2016 and CatBoost, Dorogush et al. 2018; Prokhorenkova et al. 2018), are two Gradient Boosting (GB, Friedman 2000) algorithms that implement two different schemes for calculating gradients.

Let us consider the ensemble of K trees, where the predicted score for an input $\mathbf{x}$ is given by the sum of the values predicted by the individual K trees: $\hat{y}^K(\mathbf{x}) = \sum_{j=1}^{K} f_j(\mathbf{x})$, where f_j is the output of the j-th decision tree.

Building the $(K+1)$th decision tree minimizes an objective function $L = \sum_{i=1}^{n} \ell(y_i, \hat{y_i}^K(\mathbf{x}_i) + f_{K+1}(\mathbf{x}_i)) + \Lambda(f_{K+1})$, where $\ell(y_i, \hat{y_i}^K(\mathbf{x}_i) + f_{K+1}(\mathbf{x}_i))$ depends on the first (and, possibly, second) deviation of the loss function $\ell(y_i, \hat{y_i}^K(\mathbf{x}_i))$, and $\Lambda(f_{K+1})$ is a regularization function that penalizes the complexity of the $(K+1)$-th tree to prevent overfitting. To build a $(K+1)$-th decision tree, the algorithm starts with a single decision node and iteratively tries to add a best split for each node, until a stop criterion on tree growth is satisfied.

XGBoost estimates the gradient value for all of the objects in a leaf and calculates the average gradient to determine the best split for each node. In this way, the gradient is estimated via the same data points, on which the current decision tree was built on. In general, such splitting procedure leads to the gradient bias (due to the repeated usage of the same objects through all iterations) and, as result, to an overfitting problem (Prokhorenkova et al. 2018).

CatBoost, the chosen algorithm for this paper, in its turn, implements the splitting technique called Ordered Boosting (Prokhorenkova et al. 2018), that overcomes this problem. With Ordered Boosting, the gradients are calculated not for all of the objects, but for the shuffled training dataset (so-called, random permutations), wherein the gradients are calculated for the objects before a given one. In such a way, gradient for the $j+1$ object is calculated based on prediction of the model, learnt by previous samples in shuffled dataset.

One of the limitations of the GBDT algorithms is a wide range of parameters that have to be tuned to get the highest classification quality; CatBoost has advantages also on this, because it performs well also without hyperparameters tuning.

For our task, the most influential hyperparameters, that have to be tuned in GBDT algorithms, are:

1. `learning_rate` – the rate of gradient descent;
2. `min_split_loss` – the minimum loss reduction required to split a node of the tree;
3. `max_depth` – the maximum depth of the decision trees;
4. `min_child_weight` – the minimum number of samples in the node of the decision tree required to make a split;
5. `max_delta_step` – the maximum step controlling convergence during gradient descent;
6. `colsample_bytree` – the subsample ratio of the features during building each decision tree;
7. `subsample` – the subsample ratio of the training objects

In particular, we noticed that the parameters that mostly affected the learning quality for our training dataset were `learning_rate` (greater values correspond to a sharper gradient descent, that is good for learning acceleration, but can lead to missing the loss minimum), and `max_depth` (greater values correspond to a large complexity of the trees, and can lead to overfitting).

Moreover, GB algorithms usually allow to use a stop criterion, responsible for the termination of the learning when an overfitting occurs (the so-called, `early_stopping` parameter). It is expressed via the number of constructed trees, after which the quality of the metric does not increase anymore. Usually this parameter ranges between 10 and 1000 trees, depending on `learning_rate`. If the early stopping criterion is met, the GB algorithm accepts the number of trees, satisfied to the best score. Paying particular attention to the `early_stopping` parameter is the best way to avoid as much as possible overfitting. To quantify, how the quality of the classification changes over the iterations, for XGBoost and CatBoost, it is necessary to define a quality metric. Defining this metric allows to express the changing of the classification quality against the complexity of GBDT algorithm and can be easily used to control the overfitting. Widely used quality metrics are accuracy, precision, recall and F1-score; however, these metrics are sensitive to the imbalance in number of training sources among different classes. Therefore, we decided to use Matthews correlation coefficient (MCC, Matthews 1975) which is instead insensitive to it.

Finally, a good way to decrease the number of the stars and galaxies classified by the algorithm as quasars is to apply a weight to the loss function of these two classes. In fact, this trick, applied to CatBoost, helped us in the paper to improve the final purity of the quasar selection with a minimal decreasing of the completeness on the training set. In particular, we weighted the loss function for the `STAR` and `GALAXY` samples in the following way:

$$L = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i [\ell(y_i, \hat{y_i}^K(\mathbf{x}_i) + f_{K+1}(\mathbf{x}_i)) + \Lambda(f_{K+1})] \qquad (A.1)$$

where $w_i = 1$ if source is a `QSO` and $w_i = 4$ if source is `STAR` or `GALAXY`.

Appendix A.2: Random Forest

Another method of ensemble learning with decision trees is based on the use of a Random forest (RF, Breiman 2010) algorithm. This was the choice adopted in KiDS DR3. The basic

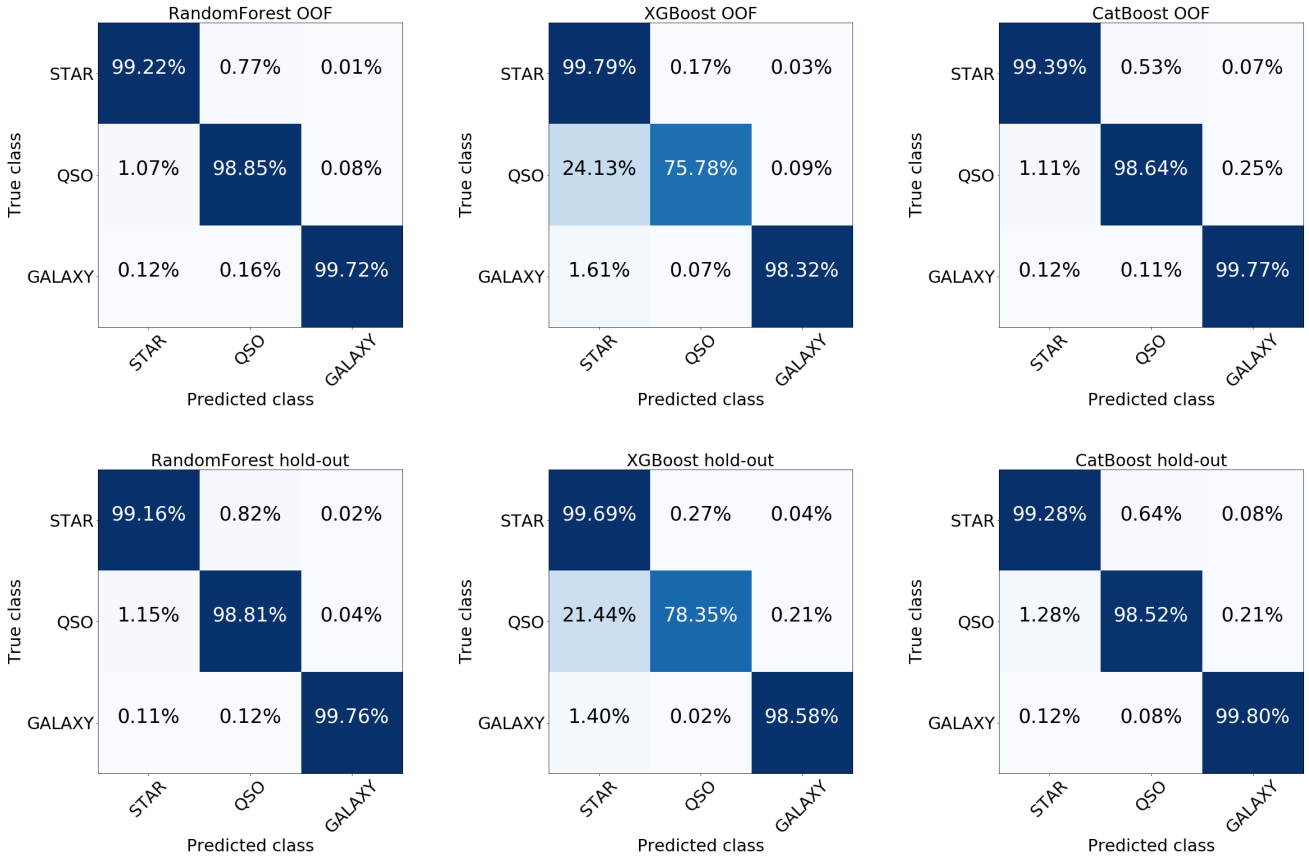

Fig. A.1: Confusion matrices for the RF (*left*), XGBoost (*central*) and CatBoost (*right*) predictions on OOF sample (*top* panels) and hold-out sample (*bottom* panels).

idea of RF is that a set of decision trees can fit a robust classifier by averaging their decisions. From the performance side, the RF constructs a large number of decision trees and then uses the majority vote among them.

The main pipeline of the single decision tree carries on the following steps:

1. generation of different samples from the training set with the same size, but using random subsets of all the objects. The repetition of some objects among different subsamples is required to make the sample complete (so-called, random subsample with replacement);
2. let the decision tree learning on the random subsample with replacement (1), but using randomly selected $\approx \sqrt{m}$ features.

The RF consists, then, in many learning processes, each performed on a single decision tree using different random subsamples with replacement and randomly selected features. Then, the single prediction of a class for given objects is a simple average on the predictions of all constructed decision trees (bootstrap bagging method).

The big advantage of RF is that it uses together the bootstrap bagging method (averaging prediction of the estimators learnt with random subsamples with replacement) and the learning of each estimator with random subset of features. These procedures prevent overfitting and, in most of the cases, help to improve the classification performance and to increase the generalization ability of the RF.

The principal hyperparameters, that are required to the RF fitting on the training dataset, are:

1. `n_estimators` – the number of decision trees;
2. `max_features` – the maximum number of features to be used in the node;
3. `max_depth` – the maximum depth of the decision trees;
4. `min_samples_split` – the minimum number of samples in the node of the decision tree required to make a split;
5. `min_samples_leaf` – the minimum number of samples required to be in the leaf node (the end node in which the splitting finishes) of each tree;
6. `class_weight` – weights associated with each class (is required in the case of imbalance training sample)

The number of estimators and the maximum depth (and/or minimum number of samples in the node) are mandatory hyperparameters. Moreover, a fine tuning of the parameters 3, 4 and 5 is crucial to avoid overfitting. For instance, setting the values of these parameters to their common values of $\{\infty, 2, 1\}$, respectively will most of the time lead to overfitting. In fact, if the depth of the decision tree is too high, and the minimum number of sample required to be in the leaf node is too small, then each single object in the training will have his own class characterized by its features. This will then make impossible to classify new, unknown objects, although the accuracy of classification of the training set will equals about 100% (Mansour 1997). Thus, to reduce overfitting, one has to limit the maximum depth of the decision trees (usually set to 3-20, depending on the amount and topology of the features), and/or the minimum number of samples in the nodes.

Appendix A.3: Performance of the classification algorithms

To directly compare the performance of the three algorithms that we tested, we use confusion matrices. These show the relative number of predicted objects in each of the three classes with respect to the number of true classes. The confusion matrices for RF, XGBoost and CatBoost are shown in Figure A.1. As we can see, RF provides the highest completeness for quasars ($\approx$ 98.8%), but with a contamination of $\approx$ 0.8% from stars and $\approx$ 0.1% from galaxies; XGBoost shows the largest purity of quasar selection, but with a very low completeness (only $\approx$ 75.8% quasars were classified as quasars).

For the `GALAXY` class, CatBoost and RF provide very similar completeness (higher than XGBoost), with a very low contamination from star ($<$ 0.1%) but CatBoost gives a much larger contamination from `QSO` ($\approx$ 0.25%).

As one can see, RF and CatBoost show very similar results. To better understand which was the best choice to make for our scientific purposes, we decided to compare the difference between the MCC value received for the training sample and the one received for the OOF sample, for each of the algorithms. Usually, a large difference between these two scores (training and validation) indicates an overfitting in the model, i.e. the classifier lose the generalization ability, and consequently is able to classify correctly only training data. For the RF, we received the following MCC values for training and OOF samples respectively: 0.9925 and 0.9892 (with a difference of 0.0033). For CatBoost the MCC equals 0.9901 for training data and 0.9894 for the OOF sample, thus, in this case, the difference (0.0007) is almost 5 times smaller.

Thus, the better ability to generalize good results on unseen dataset, combined with the fact that CatBoost keeps the purity and the completeness of the quasar selection at a very high level and it maximize the galaxies completeness, at the same time removing as much as possible stellar contaminants, convinced us that CatBoost is the optimal classifier for our purposes.