





# Boost recall in quasi-stellar object selection from highly imbalanced photometric datasets

## The reverse selection method<sup>★</sup>

Giorgio Calderone<sup>1</sup> , Francesco Guarneri<sup>1,2</sup>, Matteo Porru<sup>1</sup>, Stefano Cristiani<sup>1,3,4</sup> , Andrea Grazian<sup>5</sup>, Luciano Nicastro<sup>6</sup> , Manuela Bischetti<sup>1</sup>, Konstantina Boutsia<sup>7,8</sup>, Guido Cupani<sup>1,3</sup>, Valentina D’Odorico<sup>1,3,9</sup> , Chiara Feruglio<sup>1</sup>, and Fabio Fontanot<sup>1,3</sup>

<sup>1</sup> INAF – Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, 34143 Trieste, Italy  
e-mail: [giorgio.calderone@inaf.it](mailto:giorgio.calderone@inaf.it)

<sup>2</sup> Dipartimento di Fisica, Sezione di Astronomia, Università di Trieste, via G.B. Tiepolo 11, 34143 Trieste, Italy

<sup>3</sup> IFPU – Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy

<sup>4</sup> INFN – National Institute for Nuclear Physics, via Valerio 2, 34127 Trieste, Italy

<sup>5</sup> INAF – Osservatorio Astronomico di Padova, Vicolo dell’Osservatorio 5, 35122 Padova, Italy

<sup>6</sup> INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via P. Gobetti 101, 40129 Bologna, Italy

<sup>7</sup> Cerro Tololo Inter-American Observatory/NSFs NOIRLab, Casilla 603, La Serena, Chile

<sup>8</sup> Las Campanas Observatory, Carnegie Observatories, Colina El Pino, Casilla 601, La Serena, Chile

<sup>9</sup> Scuola Normale Superiore, P.zza dei Cavalieri, 56126 Pisa, Italy

Received 8 April 2023 / Accepted 12 December 2023

### ABSTRACT

**Context.** The identification of bright quasi-stellar objects (QSOs) is of fundamental importance to probe the intergalactic medium and address open questions in cosmology. Several approaches have been adopted to find such sources in the currently available photometric surveys, including machine learning methods. However, the rarity of bright QSOs at high redshifts compared to other contaminating sources (such as stars and galaxies) makes the selection of reliable candidates a difficult task, especially when high completeness is required.

**Aims.** We present a novel technique to boost recall (i.e., completeness within the considered sample) in the selection of QSOs from photometric datasets dominated by stars, galaxies, and low- $z$  QSOs (imbalanced datasets).

**Methods.** Our heuristic method operates by iteratively removing sources whose probability of belonging to a noninteresting class exceeds a user-defined threshold, until the remaining dataset contains mainly high- $z$  QSOs. Any existing machine learning method can be used as the underlying classifier, provided it allows for a classification probability to be estimated. We applied the method to a dataset obtained by cross-matching PanSTARRS1 (DR2), *Gaia* (DR3), and WISE, and identified the high- $z$  QSO candidates using both our method and its direct multi-label counterpart.

**Results.** We ran several tests by randomly choosing the training and test datasets, and achieved significant improvements in recall which increased from ~50% to ~85% for QSOs with  $z > 2.5$ , and from ~70% to ~90% for QSOs with  $z > 3$ . Also, we identified a sample of 3098 new QSO candidates on a sample of  $2.6 \times 10^6$  sources with no known classification. We obtained follow-up spectroscopy for 121 candidates, confirming 107 new QSOs with  $z > 2.5$ . Finally, a comparison of our QSO candidates with those selected by an independent method based on *Gaia* spectroscopy shows that the two samples overlap by more than 90% and that both selection methods are potentially capable of achieving a high level of completeness.

**Key words.** methods: statistical – astronomical databases: miscellaneous – catalogs – surveys – quasars: general

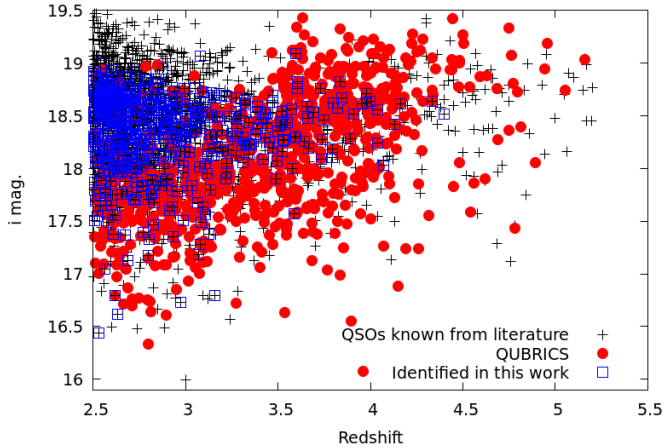
## 1. Introduction

Light from distant and powerful quasi-stellar objects (QSO) has proven to be a useful tool to probe the inter-galactic medium (IGM; Meiksin 2009; McQuinn 2016; Péroux & Howk 2020), investigate fundamental physics (Murphy et al. 2022), carry out cosmological studies (Grazian et al. 2022), probe the growth of supermassive black holes (Trakhtenbrot 2021), and even probe the dynamics of the Universe (Liske et al. 2008; Boutsia et al. 2020; Cristiani et al. 2023). The role of such cosmological beacons is expected to become even more important in the

next decade with the upcoming availability of high-resolution spectrographs on the 30 m-class telescopes. Therefore, comprehensive catalogs of high- $z$  bright QSOs are of the utmost importance to fulfill these goals, and a significant effort has been devoted from the astronomical community to identify new QSOs using machine learning techniques (e.g., Bailer-Jones et al. 2019; Jin et al. 2019; Wolf et al. 2020; Wenzl et al. 2021; Nakoneczny et al. 2021; Rodrigues et al. 2023). Similar techniques have been used to identify stars and extragalactic sources (e.g., Khrantsov et al. 2019; Nakazono et al. 2021; Barbisan et al. 2022; Hughes et al. 2022).

Most of the past research in the field has been, however, carried out using QSO samples in the northern hemisphere mainly due to the sky coverage of large-area surveys such as the Sloan

\* Table B.1 is available at the CDS via anonymous ftp to [cdsarc.cds.unistra.fr](https://cdsarc.cds.unistra.fr) (130.79.128.5) or via <https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/683/A34>



**Fig. 1.** Known QSO with  $\delta < 0$  and  $i \text{ mag} < 19.5$  from the considered catalogs in literature (black symbols, see Sect. 4), from the QUBRICS survey (red circles), and identified by the reverse selection method in this work (blue squares, see Sect. 7).

Digital Sky Survey (SDSS, Lyke et al. 2020). However, several advanced facilities such as the Ultraviolet and Visual Echelle Spectrograph (UVES), the Echelle SPectrograph for Rocky Exoplanets and Stable Spectroscopic Observations (ESPRESSO), and the ArmazoNes high Dispersion Echelle Spectrograph (ANDES) are or will be in operation in the southern hemisphere. In order to fill this gap, we started the QUBRICS<sup>1</sup> survey in 2018, aiming to identify the brightest high- $z$  ( $z > 2.5$ ) QSOs in the southern hemisphere, using data available in photometric databases such as the SkyMapper, the Panoramic Survey Telescope and Rapid Response System (PanSTARRS), the dark energy survey (DES), *Gaia*, and the Wide-field Infrared Survey Explorer (WISE) surveys, as well as machine learning selection algorithms. So far, we have obtained 1302 high-quality spectra<sup>2</sup> of QSO candidates using several facilities (*Magellan* Telescopes: Baade/IMACS and Clay/LDSS-3, du Pont/WFCCD, TNG/Dolores, and NTT/EFOSC2). Among these, 1219 (94%) were actual new Active Galaxy Nuclei (AGN) or QSO identifications, with 943 of them (72%) having  $z > 2.5$ , and 1079 (83%) having  $z > 2$ . We also identified 123 QSOs with  $z > 4$ , with the highest redshift being 5.16. The remaining sources were stars (56) and galaxies (27). Figure 1 shows the distribution in the  $i \text{ mag}$  versus  $z$  plane of the QSOs known from literature and of those identified by QUBRICS. The QUBRICS survey has already produced several papers discussing both of the selection methods (Calderone et al. 2019; Guarneri et al. 2021) and the exploitation of new QSO identifications (Boutsia et al. 2020, 2021; Cupani et al. 2022; Grazian et al. 2022).

The task of identifying bright, high- $z$  QSO candidates in photometric catalogs is the classical needle-in-a-haystack problem. For the catalogs considered in this work (Sect. 4), there are roughly  $10^4$  stars and 100 galaxies for each bright QSO with  $i \lesssim 19$  and  $z > 2.5$ , that is, the dataset is imbalanced toward stars. Hence the selection methods need to be carefully tuned, and their performance constantly monitored, in order to minimize the required telescope time and maximize the success rate. In the QUBRICS case the success rate (or precision, Sect. 2) has always been  $\sim 70\%$  (Calderone et al. 2019; Guarneri et al.

2021), and has steadily improved up to the most recent observing runs. The latest progress has been driven by the adoption of the probabilistic random forest (PRF; Guarneri et al. 2021) and XGBoost (this work) algorithms in place of the canonical correlation analysis used in the first works (Calderone et al. 2019; Boutsia et al. 2020). The initial goal of QUBRICS, namely to identify the brightest QSOs at redshift  $z > 2.5$  to probe the IGM and the dynamics of the Universe, has been partly achieved as shown in Fig. 1.

For other cosmological studies, however, it is mandatory to achieve a high and well-determined recall (i.e., the completeness within the considered photometric sample, Sect. 2). The recall is more difficult to estimate than precision since the former can only be assessed indirectly by estimating the (unknown) true number of high- $z$  QSOs in the subsample still lacking spectroscopic classification. Precision, on the other hand, can be extrapolated even with a limited set of observations. Until recently, the main obstacle has been the limited overlap between the southern hemisphere surveys used to search for new high- $z$  QSOs (namely, SkyMapper, DES, PanSTARRS) and other surveys with significantly higher QSO completeness in the North (such as SDSS). Depending on the adopted method and the underlying assumptions, we estimated a recall between 70% and 85% (Calderone et al. 2019; Guarneri et al. 2021, 2022). With the latest observations we now have more than 900 QSOs at  $z > 2.5$  and 600 at  $z > 3$ , that we can split into training and test datasets (with the latter containing a few tens of objects), in order to evaluate the recall in a self-consistent manner.

The QUBRICS survey aims to identify the remaining, not yet identified, QSOs in the redshift range 2.5–5.0 (with the upper limit due to the requirement of having a *Gaia* detection). Hence we need to maximize the recall, even though this could cause a reduced precision. We also need to take into account the overwhelming number of noninteresting sources (mainly stars) present in the photometric datasets compared to the number of bright and high- $z$  QSOs, which results in highly imbalanced datasets. The goal of this work is to present a method to boost the recall of a machine learning multi-label selection algorithm, under the only assumption that the latter provides an estimate of the probability for a source to belong to a given class. It operates by iteratively discarding objects with high probability of not being a QSO, thus automatically rebalancing the input datasets, and providing higher recall rates with respect to other classification methods (Sect. 3).

The paper is organized as follows: Sect. 2 describes the selection performance estimators used throughout the paper; Sect. 3 describes our method to boost recall in the multi-label selection case and in Sect. 5 we apply it to the specific problem of identifying new high- $z$  QSO candidates; Sect. 7 reports the observations of such candidates, and the comparison with the QSO candidates obtained with an independent method. Finally, in Sect. 8 we draw our conclusions.

## 2. Performance metrics

In order to measure the performance of a selection method, we need proper metrics as described in this section. We introduce the concepts of “positive” (P) and “negative” (N) classes in the context of a binary classification, where the former represents the class of objects of interest for a specific purpose, and the latter contains all the objects supposed to be rejected by a selection algorithm. Whenever an object is correctly classified as belonging to either the P or N class we call it a “true

<sup>1</sup> QUasars as BRIght beacons for Cosmology in the Southern hemisphere.

<sup>2</sup> Plus 149 spectra with uncertain classification due to low S/N or too few available features for a robust classification.

**Table 1.** Representation of a confusion matrix for a two-class (P vs. N) problem.

		Predicted class	
		pP	pN
True Class	P	TP (true positives)	FN (false negative)
	N	FP (false positive)	TN (true negative)

**Notes.** The rows represent the sources whose actual class is P or N. The columns represent the Predicted Positive (pP) and Predicted Negative (pN) sources respectively.

positive” (TP) or “true negative” (TN) prediction respectively. Whenever the prediction is wrong we call it a “false negative” (FN) or “false positive” (FP) respectively. The number of “true positives”, “false negatives”, etc. can be arranged in a tabular form known as “confusion matrix” (see Table 1 for an example). With these definitions, the relevant metrics are:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{pP}}, \\ \text{True Positive Rate (TPR or recall)} &= \frac{\text{TP}}{\text{P}}, \\ \text{False Positive Rate (FPR)} &= \frac{\text{FP}}{\text{N}} \end{aligned} \quad (1)$$

where “pP” represents the total number of sources predicted to belong to the positive class (i.e., the sum along the first column), while “P” and “N” represent the total number of sources actually belonging to each class (i.e., the sum along the rows).

In a classification process the “precision” metric is the expected success rate in identifying new sources belonging to the P class, while the “recall” is the expected P-class completeness within the considered sample. We note that the above metrics may be biased if the datasets are imbalanced (e.g., when  $N \gg P$ ) and tend toward values which depend on the class ratio P/N rather than measuring the actual capabilities of the method. As a consequence, there is not an absolutely “good” or “bad” value for precision and recall, since they depend on the specific case. On the other hand, it is always possible to compare the performance of two algorithms and decide which one provides the best precision or recall performance, regardless of the “goodness” of the absolute values of such metrics.

If the classifier algorithm also provides a probability estimate for a source to belong to a specific class, it is possible to adopt a discriminating threshold and accept a P-classification as reliable only if its probability exceeds the threshold (Provost & Fawcett 1997; Provost 2000). In a single run of a binary classifier this would alter the metrics in a correlated way. As an example, a conservative discriminating threshold would typically lead to higher precision but lower recall and FPR metrics (see Sect. 3.1 and Fig. 4).

The above mentioned concepts can be easily extended to the multi-label case by introducing the relevant classes such as stars, galaxies, low- $z$  QSOs, high- $z$  QSOs etc., in place of the P and N ones. For instance, the top panel of Table A.1 provides the example of a confusion matrix in the multi-label case. An alternative representation of the same confusion matrix is given by normalizing each row by the total number of sources predicted to belong to a class (middle panel of Table A.1), or by the total number of sources actually belonging to each class (lower panel of Table A.1). In the former case the diagonal numbers

represent the precision (i.e., the expected success rate in identifying members of a given class), while in the latter case they represent the recall of the method (i.e., the completeness for a given class within the considered sample).

As mentioned in Sect. 1, we are mainly interested in high redshift sources with  $z \gtrsim 2.5$  hence we consider separate classes for low- $z$  and high- $z$  QSOs, with  $z = 2.5$  as discriminating threshold. Within the high- $z$  QSO class we are much more interested in sources with  $z > 3$  or 4 rather than those at  $z \sim 2.5$ , but the latter are more abundant than the former due to the uneven redshift distribution of detectable QSOs. As a consequence, the recall metric for high- $z$  QSO might be biased to represent the population at  $z \sim 2.5$ , providing little information about the selection performance at  $z > 3$ . To overcome this issue, we introduced a new recall metric at  $z > 3$  by considering only the QSOs with  $z > 3$  in the TP and P calculations. As we subsequently show in the next sections, the recall at  $z > 3$  is typically higher than the standard recall metric at  $z = 2.5$ , the reason being that the QSOs with redshift slightly larger than  $z = 2.5$  may easily be misclassified as low- $z$  QSOs resulting in a lower high- $z$  recall.

For the sake of completeness, we introduce here the Normalized Median Absolute Deviation metric (NMAD) which we use to estimate the scatter when comparing the estimated and true redshift of the QSO candidates in Sect. 6:

$$\text{NMAD} = 1.4826 \text{ median}(|z_{\text{est}} - z_{\text{true}}|).$$

The NMAD is more robust than the standard deviation in presence of outliers, and the normalization factor 1.4826 makes the two quantities equal for a normal distribution (Rousseeuw & Croux 1993; Leys et al. 2013).

### 3. Boosting recall in selection methods

As discussed in Sect. 1, the purpose of this work is to present a method to significantly boost recall in QSO selection over highly imbalanced photometric datasets (at the possible expense of slightly reducing the precision), while keeping the number of additional hyper-parameters<sup>3</sup> at a minimum. The use of imbalanced datasets in machine learning algorithms is, however, known to be detrimental to performance (Prati et al. 2009), as well as being a source of bias for the performance estimators themselves (e.g., Batista et al. 2004). The issue is even more compelling when trying to identify the brightest QSOs, since they represent only a small subsample of the considered source catalogs, and their identification may be challenging. Several approaches have been suggested to address the issue of rebalancing an imbalanced dataset (e.g., Batista et al. 2004; Prati et al. 2009), such as “undersampling” (random elimination of sources in the majority class, in our case: the stars); “oversampling” (random duplication of sources in the minority class, in our case: high- $z$  QSOs); “synthetic data generation” (simulate the availability of further data for the minority class, approach discussed in Guarneri et al. 2022). These methods, although effective in rebalancing the dataset, come with drawbacks such as the possible elimination of relevant sources in the undersampling case, the possible overfit due to replication of nonrelevant features in the oversampling case, and the difficulties associated with conveying useful knowledge to the machine learning method by means of synthetic data, and generalizing the results to avoid being model-dependent. Moreover, each of the above methods

<sup>3</sup> A parameter affecting the learning process of an algorithm.



requires the addition of one or more hyper-parameters to the already long list characterizing each machine learning methods, making the exploration of the hyper-parameters space even more challenging. Besides, the worse performance of selection algorithms dealing with imbalanced training datasets may not be due to the imbalance itself, but possibly also to the “class overlapping” issue, that is, the difficulty in distinguishing members of two different classes with the available information (Prati et al. 2004). Smith et al. (2013), in particular, provides a definition of the instance hardness as the likelihood for a source to be misclassified, and proposes an undersampling method to rebalance the training dataset by removing problematic sources with high instance hardness (using a calibrated threshold). Such supervised undersampling method is called Instance Hardness Threshold (IHT). Other methods based on threshold tuning to address specific problems are discussed in Zou et al. (2016); Johnson & Khoshgofaar (2021). Several other methods had been proposed to deal with imbalanced datasets (Fernández et al. 2019).

Our heuristic method is similar to the IHT, but focuses on removing sources with high probability of being noninteresting ones, rather than on those being hard to classify. It builds upon existing classifier algorithms such as random forest (e.g., Parmar et al. 2019), probabilistic random forest (Reis et al. 2019), gradient boosting (Friedman 2001), or any other classifier able to provide an estimate of the classification probability. Note in particular that our method capability to handle missing values (which are very common in astronomical photometric data) is exactly the same as the underlying classifier algorithm. The following sections illustrates the method in the binary case (Sect. 3.1) and in the general multi-label case (Sect. 3.2).

### 3.1. The binary classifier case

By varying the classification probability threshold (hereafter,  $\tau$ ) to reliably accept a P-classification we can alter the algorithm performance metrics. More specifically, by increasing  $\tau$  the number of sources whose classification probability exceeds the threshold would be smaller, and both TP and FP decrease. As a consequence, pP decreases faster than TP, and the resulting precision is increased. On the other hand, the recall and FPR decrease since both P and N are fixed (although unknown). Figure 4 shows an example of such correlations.

In the binary case the association of the “interesting” sources with the P or N class is arbitrary and we can consider the case of exchanging their roles. The TP in the first case becomes the TN in the swapped case,  $FP \rightarrow FN$  and, most importantly,  $FPR \rightarrow 1 - TPR$ . The consequence is that applying a threshold  $\tau$  to accept a P-classification amounts to decrease its FPR, and boost the recall on the complementary class (N).

### 3.2. The multi-class case (reverse selection method)

The photometric sample discussed in Sect. 4 contains at least four different classes: stars, galaxies, low- $z$  and high- $z$  QSOs. A multi-label classifier, hereafter “direct selection method”, trained using four such classes can be built upon a binary one following either the one-vs-rest or one-vs-one heuristics (Allwein et al. 2000). In the former case we train a binary classifier to distinguish objects of one class (say high- $z$  QSOs) from all other sources, repeat for all available classes (four in our case), and consider the label with the highest score as the output classification. In the one-vs-one heuristic all possible combinations of class pairs are fed to a binary classifier, and the final prediction is

based on the majority of votes in each run. However, the performance of both heuristics is badly affected by the fact that there is no mitigation for the imbalanced datasets.

The method proposed here is a mixture of the one-vs-rest and undersampling techniques, the former being necessary to apply a threshold  $\tau$  on the classification probability for a source in the P class (thus improving the recall of the sources in the N class), and the latter being used to discard all sources belonging to the noninteresting P class (thus rebalancing the datasets). We note that in this case the discarded sources are not chosen randomly as in the standard undersampling approach (Sect. 3), but chosen among the sources with a high probability of belonging to the noninteresting P class. Our algorithm proceeds through the following steps (see also Fig. 2): (i) consider the class with the largest number of sources (stars in our case). Train a binary classifier to distinguish a star (P) from all other sources (N). Predict a classification on the training dataset and discard all the entries with correct<sup>4</sup> P-prediction whose probability is greater than  $\tau$ . Similarly, predict a classification on the test and unclassified (Sect. 7) datasets, remove entries having P-prediction with probability greater than  $\tau$  and associate them with the “Star” label. Ignore predictions with probabilities  $\leq \tau$ ; (ii) repeat for the next most abundant classes, namely galaxies and (iii) low- $z$  QSOs; (iv) train a multi-label classifier to associate a label to the remaining sources in the test and unclassified datasets. All classifications are accepted as reliable in this step, that is, no threshold on the classification probability is adopted. We note that, unlike other methods based on an a-posteriori “moving threshold” (e.g., Esposito et al. 2021; Baqui et al. 2021), our method requires the models to be re-trained whenever the threshold  $\tau$  is changed.

The method sketched above is supposed to provide a higher recall metric than its direct multi-class counterpart, although with a possibly slightly lower precision. Also, we note that the method is extremely simple to implement, and is agnostic with respect to the underlying classification framework, provided the latter allows a classification score or probability to be estimated. The interpretation of the classification probability threshold  $\tau$  is straightforward: the higher the threshold, the more conservative the method is in discarding sources. Finally, it deals naturally with highly imbalanced datasets by discarding noninteresting sources and simultaneously rebalancing all datasets (Table 4 shows the size the relevant datasets at each step of the method). Our approach is dubbed reverse selection method, since it focuses on the items to be removed from the photometric datasets, rather than on those to keep, in order to maximize the recall on high- $z$  QSOs.

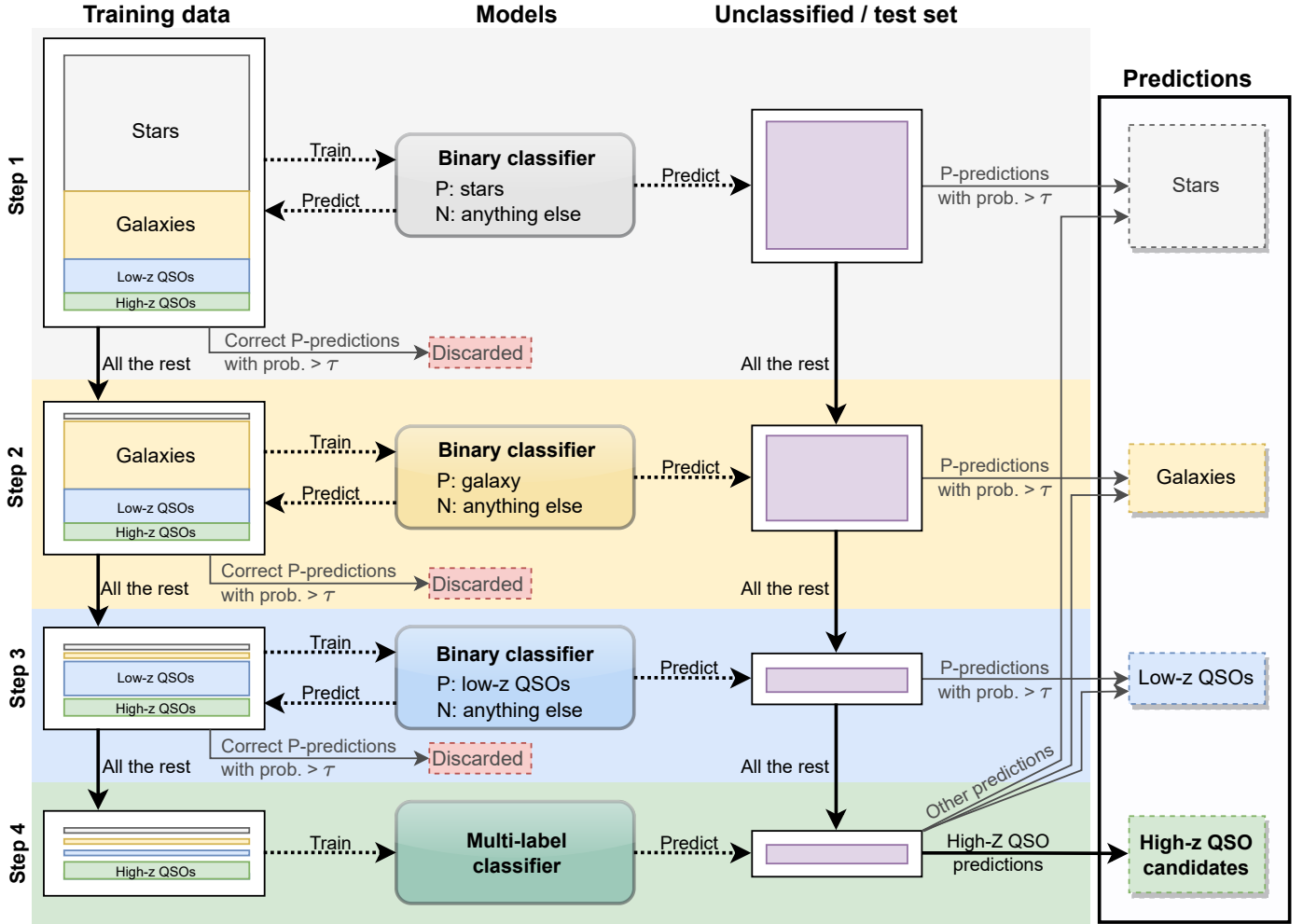
## 4. Datasets and features

The photometric dataset used throughout this paper has been prepared as follows: we selected all objects from the PanSTARRS1 (DR2) survey<sup>5</sup> (Chambers et al. 2016), with declination  $< 15^\circ$ , galactic latitude  $> 25^\circ$  (in absolute value) and  $Y$  band PSF magnitude  $14 < Y < 19$ .<sup>6</sup> We considered the PanSTARRS magnitudes in the  $g, r, i, z,$  and  $Y$  bands. Also, we cross-matched the

<sup>4</sup> We can check the predictions are correct because we are in the training dataset. We keep the wrong predictions since they may involve the very rare high- $z$  QSOs.

<sup>5</sup> <https://outerspace.stsci.edu/display/PANSTARRS/PS1+StackObjectView+table+fields>

<sup>6</sup> The lower limit on magnitudes is used since our training data does not cover such bright objects at high redshifts.



**Fig. 2.** Schema of the reverse selection method. In the first three steps a binary classifier is used to predict classification on all datasets, including the training one. If the probability of belonging to the noninteresting P-class is greater than a threshold  $\tau$  the source is discarded before proceeding to the next step. By doing so all datasets decrease in size and, most importantly, they are rebalanced toward the interesting sources, namely the high- $z$  QSOs. The last step is a simple multi-label classification, just like the direct selection method.

resulting table with *Gaia* DR3 (De Angeli et al. 2023) with a matching distance of  $0.5''$ . To avoid possibly spurious matches, we discarded all sources with multiple matching counterparts ( $\sim 0.004\%$  of the total). We considered the  $G$ ,  $RP$ , and  $BP$  magnitudes. Finally, we cross-matched with the AllWISE catalog (Wright et al. 2010) with a matching distance of  $0.5''$ , and discarded sources with multiple matching counterparts ( $\sim 0.06\%$ ). We considered the magnitudes in all the four WISE bands. The overall sample contains 30 796 027 sources (hereafter “main sample”). We identified stars in this sample as those sources having a proper motion or a parallax (as measured by *Gaia*) greater than zero with a  $3\sigma$  confidence level. Also, we cross-matched the main sample against several catalogs from the literature (Colless et al. 2001; Jones et al. 2009; Véron-Cetty & Véron 2010, Yang et al. 2016, Schindler et al. 2019a,b; Wolf et al. 2020; Lyke et al. 2020; Onken et al. 2022), as well as from our QUBRICS catalogs (Calderone et al. 2019; Boutsia et al. 2020; Guarneri et al. 2021, 2022) to assign a spectroscopic classification and, for AGN, QSOs, and galaxies, a redshift. For 2 639 184 sources we did not find any classification, hence this subset constitutes our “unclassified” sample where we search for new high- $z$  QSO candidates to be observed spectroscopically.

**Table 2.** Composition of the main sample used in this work.

Class	No. of sources	Fraction
Stars	27 985 913	90.875%
Galaxies	150 581	0.489%
Low- $z$ QSOs and AGNs	16 269	0.053%
High- $z$ QSOs	1908	0.006%
Other <sup>(*)</sup>	2172	0.007%
Unclassified	2 639 184	8.570%

**Notes.** The last line represents the sources for which we could not find a spectroscopic classification, nor significant proper motion or parallax measurements in the *Gaia* catalog, hence it is the subset we use to search for new high- $z$  ( $z > 2.5$ ) QSO candidates (Sect. 7). <sup>(\*)</sup>Any other spectral classification, such as Type 2 AGN, HII region, BL Lac, etc.

The composition of the main sample is shown in Table 2. As expected, the main sample turns out to be highly imbalanced toward stars and galaxies, and the high- $z$  QSOs are just “needles in a haystack” (0.006% of the whole main sample). A similar

order of magnitude imbalance likely affects the unclassified sample, and that a simplistic approach such as observing the whole unclassified sample would definitely recover a recall of 100%, but would also be extremely inefficient in terms of telescope time since the maximum achievable precision would be of the order  $10^{-4}$ .

We used the magnitude difference between neighboring bands (i.e., colors), rather than the magnitudes themselves, as features to train the classifier model since they provide better performance when searching for very bright and rare QSOs (Guarneri et al. 2022). We note that identifying the optimal feature selection is not a purpose of this work: we are just interested in comparing the reverse selection method with its direct counterpart using the color features.

## 5. Methodology

In the following sections, we apply three different selection methods to the dataset described above, and compare their precision and recall metrics in identifying high- $z$  QSOs in photometric catalogs. We note that our analysis does not require a specific classifier algorithm, the only requirement is that it provides an estimate of the probability for a source to belong to a given class. In this work we used XGBoost<sup>7</sup> (Chen & Guestrin 2016) as the underlying framework for all our analysis.

### 5.1. Training and test datasets

In the following sections we describe two subsets of the main sample for which we have a reliable spectroscopical classification and redshift estimate into two subsets. The first one, containing 22 525 474 sources (i.e., 80% of the whole sample), is used to train the classifier. The second one, containing 5 631 369 sources (20% of the sample) is used to estimate its performance and generate the confusion matrix. We chose the 80–20 splitting schema rather than, say, 70–30 since the former is closer to the case described in Sect. 7 where we used the 100% of the known dataset to train the models. The split is chosen randomly by shuffling the data before splitting, and following a “stratified” approach, that is, by keeping the class ratios approximately constant in both the training and test datasets. We did not use any validation dataset since the optimal value for the  $\tau$  parameter can be established using the procedure described in Sect. 5.7. Also, the extreme imbalance would not allow us to have a sufficient number of high- $z$  QSO in all datasets. We note that we used exactly the same training-test split for the specific analysis runs discussed in the following sections and to produce the results shown in Table 3 and Appendix A. When multiple runs are involved (Figs. 4 and 6, Sect. 5.6) we generated randomized training-test splits as described above. The unclassified dataset contains 2 639 184 sources ( $\sim 8.6\%$  of the main sample) and it is used in Sect. 7 to identify the list of high- $z$  QSOs candidates for the observations.

### 5.2. XGBoost hyperparameters

The values of the hyperparameters for the XGBoost classifier are as follows:<sup>8</sup> we set `num_round` (number of iterations for boosting) to 40 for the direct selection, 15 for the direct selection with

<sup>7</sup> <https://xgboost.ai/>

<sup>8</sup> The list of all XGBoost hyper-parameters, along with their default values, is available at <https://xgboost.readthedocs.io/en/stable/parameter.html>

**Table 3.** Comparison of the metrics for the selection of high- $z$  QSOs using the three methods discussed in this work.

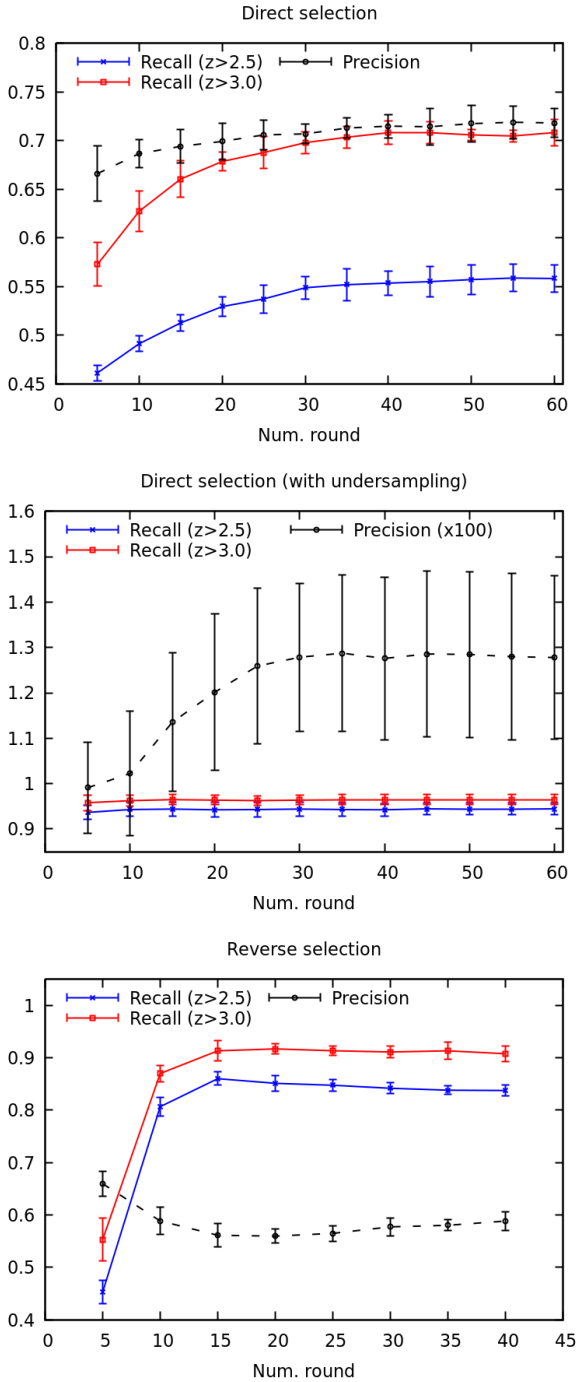
Method	Precision	Recall
Direct (Sect. 5.3)	70.3%	57.1%
Direct with undersampling (Sect. 5.4)	1.1%	94.8%
Reverse (Sect. 5.5)	55.1%	85.6%

undersampling, and 20 for the reverse selection method. The maximum depth of a decision tree (`max_depth`) is set to 20, and the tree construction algorithm (`tree_method`) to `hist` as this is the suggested setting for large datasets. Finally, (learning objective) (`objective`) is set to `multi:softprob` as this is the only setting producing probability estimates for a source to belong to a class, in a multi-class problem. The value for `num_round` has been chosen as the one where the recall curves for the considered selection method stop increasing, as determined from Fig. 3 which shows precision and recall metrics averaged over five runs (with randomly chosen training-test splits) for the direct (upper panel), direct with undersampling (middle panel) and reverse (lower panel) selection methods. We note that at higher values of `num_round` both precision and recall are approximately constant, hence our results would be scarcely affected by adopting larger values. The reverse selection method involves training several models (see Fig. 2), hence a `num_round` value may in principle be identified for each step. However, the performance at each step is affected by the outcomes of all other steps, and identifying the optimal `num_round` values for each step separately does not necessarily yield the maximum overall recall. A thorough exploration of the parameter space would thus be required to identify optimal `num_round` values for each step. This is beyond the scope of this paper, hence we simply adopt the same `num_round` value for each step in the reverse selection method.

The `max_depth` parameter controls the allowed complexity of the model, hence it is strictly related to `num_round`: we may achieve similar performance by increasing the former and decreasing the latter. The chosen value of `max_depth` = 20 enables us to achieve the performance shown in Fig. 3 in a reasonable time (training time is  $\sim 2$  min). All the remaining hyper-parameters are left at their default values. We note that the probabilities provided by XGBoost are not necessarily calibrated, that is, their reliability diagrams (Niculescu-Mizil & Caruana 2005) may deviate from the linear 1:1 relation. This is, however, not a problem for the reverse selection method since the probabilities are relevant only when compared to the threshold  $\tau$ , which in any case needs to be calibrated for the specific problem following the procedure described in Sect. 5.7. Hence there is no need to calibrate both the probabilities and the threshold. Finally, the method described in Sect. 3.2 aims to be agnostic with respect to the underlying classifier, hence the XGBoost framework used in the following sections may be replaced with any other as long as a classification probability can be estimated.

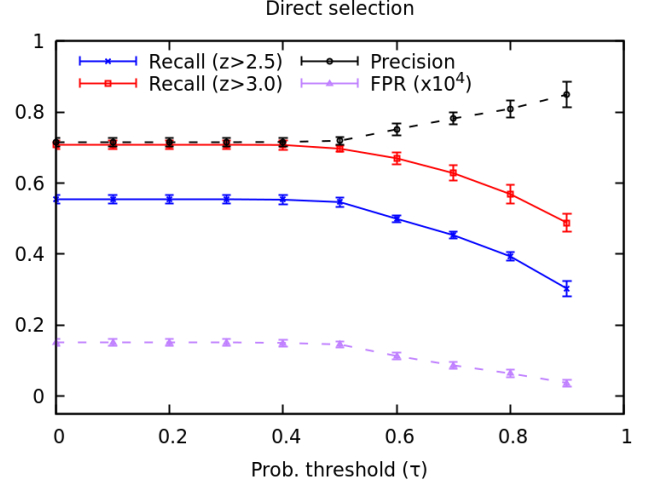
### 5.3. Results with the direct selection method

The XGBoost framework natively provides multi-label classification capabilities, hence we used it to train a model and predict the classes of the test dataset without applying any threshold on classification probability. The relevant metrics for the test dataset of a specific run are shown in Table 3, while the entire confusion



**Fig. 3.** Precision (dashed black line), recall at  $z > 2.5$  (blue line), and at  $z > 3$  (red line) as a function of the XGBoost `num_round` parameter for the direct selection method (upper panel), direct selection with undersampling (middle panel), and reverse selection (lower panel) methods. The precision curve in the middle panel is multiplied by a factor 100 to provide a clearer view. The point and error bars represent respectively the mean and the standard deviation over five runs with randomly chosen training-test splits.

matrix is available in Table A.1). On this specific run, the direct selection method features a 70% precision and 57% recall for the high- $z$  QSOs. The relatively low value for the recall is likely due to the aforementioned class imbalance problem, as shown by the significantly higher values obtained for the stars and galaxies.



**Fig. 4.** Precision (dashed black line), recall at  $z > 2.5$  (blue line), and at  $z > 3$  (red line) as a function of the classification probability threshold  $\tau$  for the direct selection method (Sect. 5.3) over five runs (with randomly chosen training-test splits). The point and error bars represent respectively the mean and the standard deviation over the five runs.

Although no probability threshold is involved in the direct selection method we can, a-posteriori, apply one such threshold to ignore those predictions whose confidence is too low (e.g., Esposito et al. 2021; Baqui et al. 2021). By considering such cases as noninteresting ones (i.e., not a high- $z$  QSO) we can estimate threshold-dependent precision, recall, and FPR metrics. Figure 4 shows such values calculated over five runs (with randomly chosen training-test splits). As expected, applying a threshold allows the algorithm to select only the most reliable high- $z$  QSOs candidates, increasing the precision but lowering the recall and FPR metrics, in contrast to our goal of raising the recall. In principle, we can tune the threshold to achieve arbitrarily high values for the precision. On the other hand, there is no way to tune the threshold to improve the recall, as the latter achieves the maximum value when no threshold is being applied (i.e., when  $\tau = 0$ <sup>9</sup>).

#### 5.4. Results with the undersampling heuristic

A common approach to deal with class imbalance is to use the undersampling heuristic (Fernández et al. 2019), namely to discard randomly chosen sources in the majority class(es) in order to rebalance the training dataset. We followed this approach by creating a balanced training dataset which consists of 1526 sources (i.e., 80% of the total number of high- $z$  QSOs, see Table 2) randomly chosen from each class, and used it to train a new classifier following the direct selection method to predict the classification on the test dataset. We note that we did not attempt to undersample the test dataset, which is now much larger than the training set. The justification for this choice is that we want to use a test dataset with the same, or similar, imbalance of the unclassified dataset, which we cannot undersample or rebalance.

The relevant metrics for the test dataset of a specific run are shown in Table 3, while the entire confusion matrix is available in Table A.2. On this specific run, the direct selection method with under sampled training dataset provides a significant boost

<sup>9</sup> For a generic multi-label classification problem with  $N$  classes all predictions have probabilities  $\geq 1/N$ , hence any threshold  $\tau$  smaller than  $1/N$  would be equivalent to  $\tau = 0$ .



**Table 4.** Evolution of dataset sizes and class imbalance at the beginning of each step of the reverse selection method.

	Step 1	Step 2	Step 3	Step 4
$N^{\text{train}}$	22 525 474	138 505	18 568	5749
$N^{\text{test}}$	5 631 369	32 141	5237	1891
$N^{\text{unclassified}}$	2 639 184	1 324 967	107 193	81 853
$f_{\text{high-}z}^{\text{train}}$ QSO	0.007%	1.102%	8.218%	26.544%
$f_{\text{high-}z}^{\text{test}}$ QSO	0.007%	1.126%	6.912%	17.610%

**Notes.** The first three rows report the size of the training, test, and unclassified datasets. The last two rows report the fraction of high- $z$  QSOs in the training and test datasets (this quantities are not available for the unclassified dataset).

in recall reaching 95%, but the precision falls down to  $\sim 1.1\%$  and the number of predicted high- $z$  QSOs increases to 33 207 sources (to be compared to the 310 candidates from Table A.1). In other words, the undersampling classifier predicts a high- $z$  QSO much more frequently than the simple direct method, achieving a very large recall rate. However, the resulting list of high- $z$  QSO candidates identified in the unclassified sample would be so large that it would be impossible to perform a spectroscopic follow-up. Since the precision is so low, we no longer consider the undersampling heuristic in the following discussion.

### 5.5. Results with the reverse selection method

We now consider applying the reverse selection method sketched in Sect. 3.2 using a probability threshold of  $\tau = 0.9$  (see Sect. 5.7 for a justification). In brief, the purpose of the method is to iteratively rebalance the datasets toward the high- $z$  QSOs by discarding sources with high probability (exceeding  $\tau$ ) of being noninteresting ones. Table 4 shows the size of the datasets and the high- $z$  QSO fractions at each step of a specific run of the reverse selection method. We note that the size of all datasets decreases, and that the high- $z$  fractions increases significantly from  $\sim 0.007\%$  at the beginning of first step, to  $\sim 20\%$  in the last step, allowing to approximately get rid of the imbalance dataset problem. Similarly, also the number of sources in the unclassified dataset decreases, possibly increasing the fraction of high- $z$  QSO contained therein.

The relevant metrics for the test dataset of a specific run are shown in Table 3, while the entire confusion matrix is available in Table A.3. On this specific run, we obtained a significantly higher recall (85.6%) at the cost of a relatively lower precision (55.1%). We note that the loss in precision is not dramatic, unlike in the undersampling heuristic. On the other hand, the high recall is not due to a specific choice of the training-test split in a specific run, as shown by the point in Fig. 6 corresponding to  $\tau = 0.9$ , whose symbol and error bar represent the average and standard deviation over five runs (with randomly chosen training-test splits).

### 5.6. Estimation of recall improvement

In order to quantify the recall improvements due to the reverse selection method when compared to the direct selection one we ran 100 analysis for both the direct and reverse selection method, randomly selecting the training and test datasets at each run (following the same procedure outlined in Sect. 5.1). The histograms

of the resulting precision and recall (for QSOs with  $z > 2.5$  and  $z > 3$ ) are shown in Fig. 5.

The legend also reports the mean and standard deviation for each metric, confirming that the reverse selection method allows us to improve the recall from  $\sim 50\%$  to  $\sim 85\%$ , and that the recall for the QSOs with  $z > 3$  is  $\geq 90\%$ , while the precision shows only a slight decrease from  $\sim 70\%$  to  $\sim 60\%$ , when compared to the direct selection method.

### 5.7. Optimal value for the probability threshold $\tau$

The reverse selection method relies on a probability threshold  $\tau$  to be calibrated in order to achieve the best possible results. Having a threshold which is too low implies we are rejecting sources for which the classification is questionable, and this may be harmful for recall. On the other hand, having a threshold which is too high prevents the algorithm from rebalancing the datasets, again limiting the recall. Hence, the proper value depends on the specific problem and possibly on the underlying classifier, and should be identified by exploration of the possible range until the recall is maximized. Note in particular that the search for the optimal values for  $\tau$  implies that a calibration of the probability estimates (Niculescu-Mizil & Caruana 2005) is, in general, not needed.

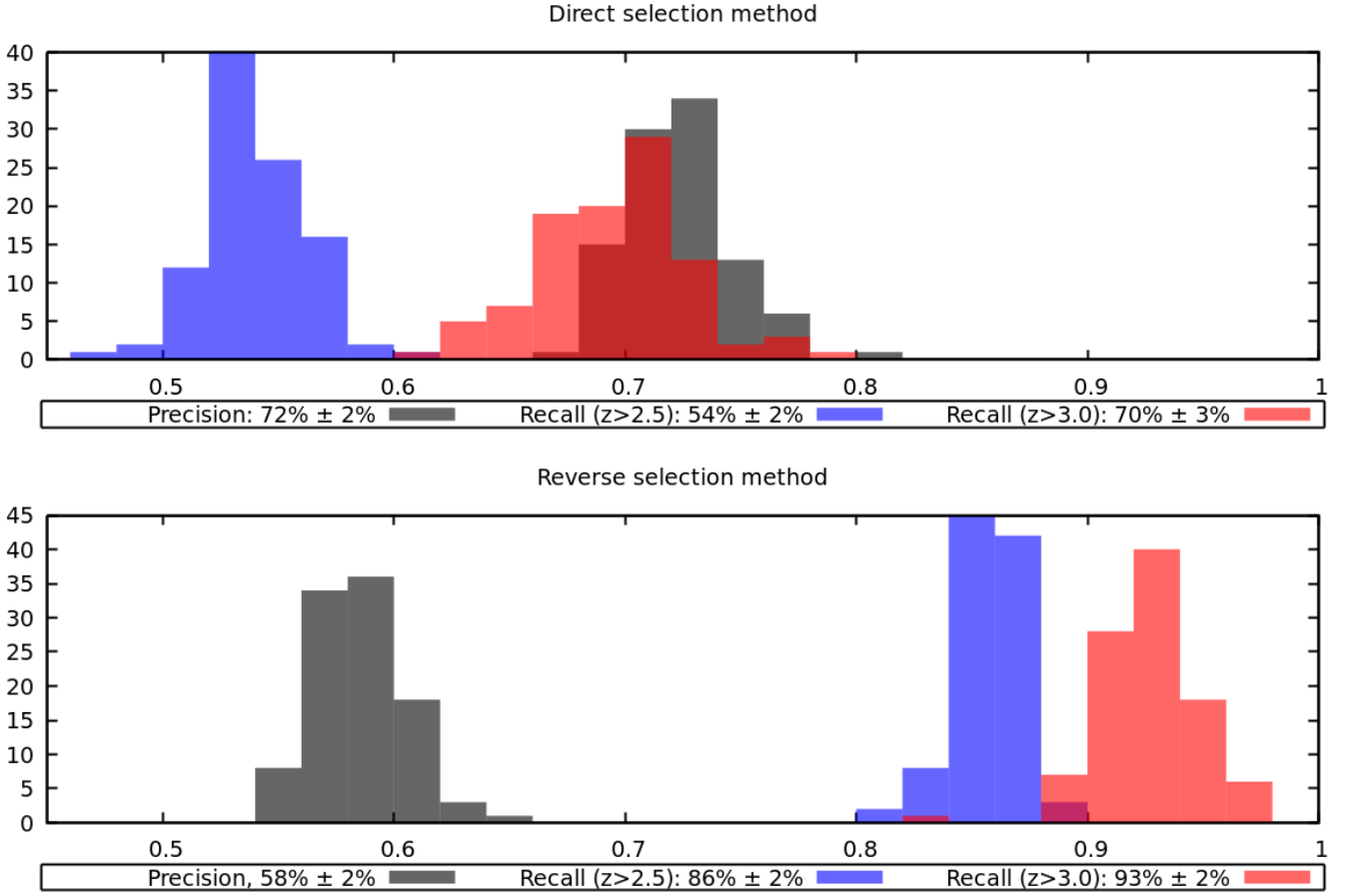
Figure 6 shows a comparison between the high- $z$  QSO precision (dashed black line) and recall (solid blue line) for the reverse selection method, as a function of the probability threshold  $\tau$ . Each point represents the mean and standard deviation over five runs (with randomly chosen training-test splits). We note that the trends of the precision and recall ( $z > 2.5$ ) curves are symmetric to those observed in Fig. 4: as  $\tau$  increases the recall is boosted only for the reverse selection method. On the other hand, precision is somewhat reduced since in the last multi-label step the training set contains high- $z$  QSOs as well as all those sources whose classification were uncertain in the previous steps, hence they are hardly representative of the remaining classes (stars, galaxies, etc.). The recall of the reverse selection method is maximized, and has a smaller scatter, for  $\tau = 0.9$ , hence this is the value adopted in this work Sects. 3.2 and 5.6.

## 6. Redshift estimation

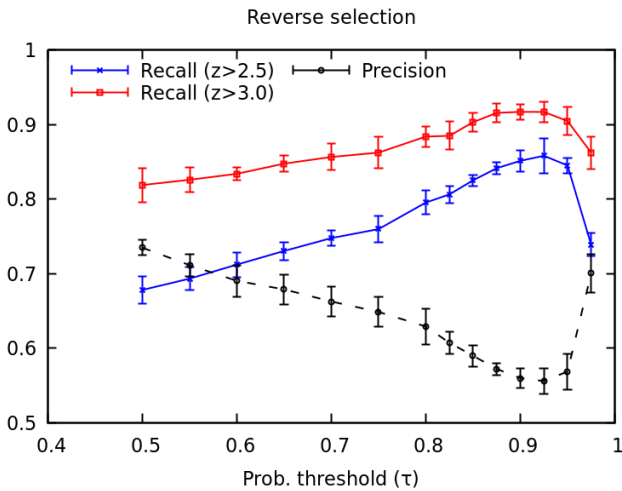
Besides classification, we also need to estimate the redshift of the QSO candidates in order to prioritize the observations (i.e., giving higher priority to higher redshift candidates, Sect. 7). Many approaches are available to estimate a photometric redshift (e.g., Brescia et al. 2021) with different level of complexity and resulting accuracy. The requirement for the prioritization, however, is not particularly tight and any method providing redshift accuracy better than  $\sim 0.5$  is enough. Hence we simply used the regression capabilities provided by XGBoost to estimate the redshift for the low- $z$  and high- $z$  QSOs candidates identified with the reverse selection method. Specifically, we used the redshift values and the same features described in Sect. 5 to train a regression model using the XGBoost framework, with the objective set to `reg:squarederror` and both `num_round` and `max_depth` equal to 15. The latter values were chosen by requiring the `rmse` metric (root mean square error) on the test data to reach a minimum.

The comparison of the true and predicted redshift values for the low- $z$  and high- $z$  QSOs candidates in the test dataset is shown in Fig. 7. Only in a few cases the predicted redshift is significantly different than the true one, especially in the upper





**Fig. 5.** Distributions of precision (black histogram) and recall (blue histogram for the sample of high- $z$  QSOs with  $z > 2.5$ , red histogram for the sample with  $z > 3$ ) as measured on 100 randomly selected test datasets (Sect. 5.6). The upper panel shows the results obtained with the direct selection method (Sect. 5.3) while the lower panel shows the same quantities obtained with the reverse selection method (Sect. 5.5).



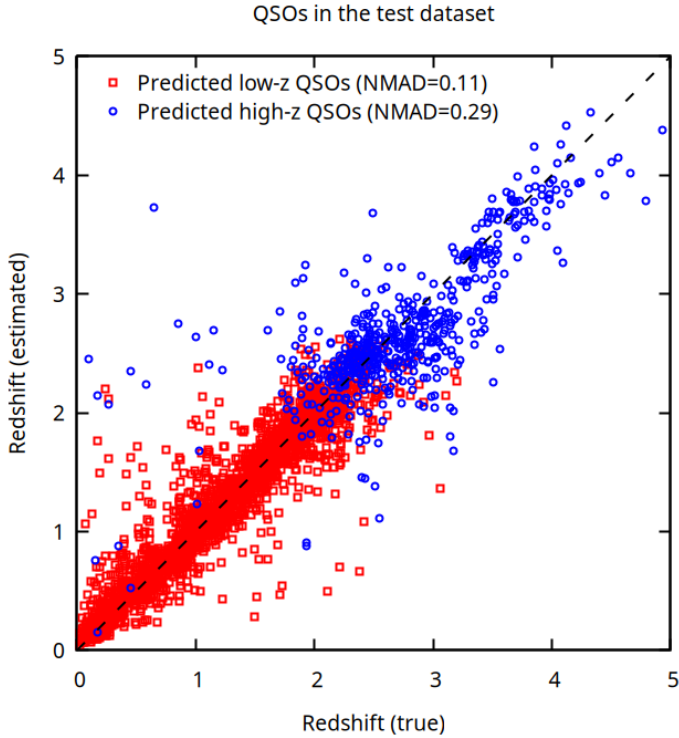
**Fig. 6.** Precision (dashed black line) and recall as a function of the classification probability threshold  $\tau$  for the reverse selection method (Sect. 5.5) over five runs (with randomly chosen training-test splits). The point and error bars represent respectively the mean and the standard deviation over the five runs. The recall lines are calculated considering the whole sample of high- $z$  QSOs with  $z > 2.5$  (blue line) and its subsample of QSOs with  $z > 3$  (red line). The recall curves peak, and have the smaller scatter, at a value of  $\tau = 0.9$ , hence this is the value adopted as threshold for the analysis in Sect. 5.5.

left corner where predicted high- $z$  QSOs have an estimated redshift smaller than 2.5. This suggests that the standard deviation of the difference (0.47 for the high- $z$  QSOs) may be biased by the presence of outliers. A more robust estimator of the scatter in presence of outliers is the NMAD (Leys et al. 2013), whose value for the predicted low- $z$  and high- $z$  QSOs is, respectively, 0.11 and 0.29. Among the 530 sources predicted to have  $z > 2.5$ , we found 35 outliers ( $\sim 6\%$ ) whose estimated redshift lies at more than  $3 \times$  NMAD from the true redshift.

## 7. The high- $z$ QSO candidates sample

The high- $z$  QSO candidates in this sample were selected among the unclassified data sources with no spectroscopic classification (Table 2) using the entire known dataset to train the models in the reverse-selection method (no test dataset is required for this purpose). The selection method identified 3098 sources as high- $z$  QSOs candidates, corresponding to  $\sim 0.1\%$  of the whole unclassified sample.

We selected the targets for spectroscopic follow-up from our list of candidates according to the observing period and giving higher priority to brighter sources in the  $i$  band, as well as to higher estimated redshifts (Sect. 6). So far, we carried out a spectroscopic observations for 121 candidates, and identified 107 new QSOs with  $z > 2.5$  (success rate of 88%), 2 stars and 12 QSOs with  $z < 2.5$ . The new QSO identifications are listed



**Fig. 7.** Comparison of estimated and true redshifts for low- $z$  and high- $z$  QSO candidates in the test dataset.

in Table B.1. Moreover, using data collected from the literature, which we ignored at the time of selection, we obtained 21 new spectroscopic redshifts (all  $>2.5$ ) in our unclassified sample. In total, we could identify 143 QSOs among our candidate sample, which adds to the previously known QUBRICS QSOs (see Fig. 1).

We also compared our list of candidates with the catalog of QSO redshifts obtained by a newly developed spectral energy distribution fitting technique exploiting both photometric information and *Gaia* DR3 spectroscopy (Cristiani et al. 2023). Such catalog contains 1672 new redshift estimates, out of which 1142 fall in the same footprint as the dataset described in Sect. 4, and 717 have a redshift greater than 2.5. We cross-matched this 717 sources with our candidates and found 658 sources in common (92%). If we consider only the sources with estimated redshift greater than 3, our candidates sample includes 103 of the 110 sources also selected by Cristiani et al. (94%). Such high overlaps between QSO candidates identified with independent methods (although on similar initial samples) show that both are potentially able to reach high level of completeness. Our method, however, relies on photometric estimates and does not require the *Gaia* spectra.

The Cristiani et al. sample also includes  $1142 - 717 = 425$  sources with  $z \leq 2.5$ , corresponding to 381 sources in our sample. These have to be considered “failures” of our selection method since their redshift is smaller than 2.5. They are present because of the limitations of our algorithm in correctly classifying QSOs whose redshift is close to the threshold of  $z = 2.5$ . In fact, our redshift estimates for such spurious source is in the range  $2 \lesssim z_{\text{XGBoost}} \lesssim 3$ , with only 2 sources having an estimated redshift greater than  $z = 3$ . As already mentioned in Sect. 2, this issue is the reason to introduce the recall metrics at  $z > 3$  (rather than  $z > 2.5$ ) to estimate the performance of our method. Concerning the other metrics, the presence of these spurious

sources with  $z < 2.5$ , besides the above mentioned 658 ones with  $z > 2.5$ , in our candidate sample implies an upper limit of  $658/(658+381) = 63\%$  for the precision of our method, which is in line with the precision estimated with the analysis described in Sect. 5.6. We note that we may improve the precision by simply neglecting the sources with estimated redshift smaller than 2.5. In this case, we would approximately halve the candidate sample (1563 sources rather than 3098), but the recall with respect to the Cristiani et al. sample of QSOs with  $z > 3$  would fall to  $\sim 78\%$ . The joint adoption of multiple selection algorithms, such as reverse selection and SED fitting (Cristiani et al. 2023) may improve the redshift estimates and the precision while keeping a recall  $\sim 90\%$ . This will be the subject of a future work.

## 8. Conclusions

We presented a novel heuristic method, dubbed reverse selection method, designed to improve the recall (i.e., the completeness over the considered dataset) of a classifier algorithm, even in the presence of a highly imbalanced dataset, at the expense of a slight decrease in precision. The method relies on the adoption of a classification probability threshold for the validation of the outcome of a binary classifier, in order to improve its precision. When applied repeatedly following the class size order (i.e., starting from stars, then galaxies, low- $z$  QSOs, etc.) this allows us to identify and remove noninteresting objects in order to rebalance the datasets toward the less common sources (high- $z$  QSOs).

We applied the reverse selection method to search for high- $z$  QSOs in a highly imbalanced dataset where most of the sources are stars or galaxies, and compared the precision and recall to its simple, direct selection multi-label classifier counterpart, both with and without random undersampling. Our results confirm that the reverse selection method provides a significant boost in recall, up to 90% (for QSOs with  $z > 3$ ) with only a small decrease in precision ( $\sim 60\%$  rather than  $\sim 70\%$ ). In order to show that the improvement in recall is not due to a particular split of the datasets, we tested the robustness of our results by randomly generating the training and test datasets at each run, confirming that our method is capable of achieving a recall of  $\sim 90\%$  (for QSOs with  $z > 3$ ).

Our heuristic method relies on an external classifier algorithm, which is XGBoost for the analysis discussed here. However, the method is agnostic with respect to the underlying classifier, and an alternative algorithm providing a classification probability estimate can in principle be used. The method relies on the probability threshold  $\tau$  whose interpretation is straightforward and whose optimal value for a specific case can be easily tuned. Also, the boost in recall metrics can be easily quantified following the procedure outlined in Sect. 5.6.

Finally, we applied our method to a sample of objects without any known spectroscopic classification, and identified a sample of 3098 new QSO candidates among them. For 121 candidates we obtained a follow-up spectroscopy, and identified 107 new QSOs with  $z > 2.5$ . A comparison with the recently released catalog of Cristiani et al. (2023) shows that both selection methods are able to achieve similar recall rates up to  $\sim 90\%$ , but our method shows such performance with no need for the *Gaia* spectroscopic data. On the other hand, our method still needs to be complemented with a reliable redshift estimate algorithm in order to be able to reduce the number of spurious sources while keeping the same recall rate.

**Acknowledgements.** We thank the anonymous Referee for the insightful comments which helped us improving the manuscript. We acknowledge financial contribution from the grant PRIN INAF 2019 (RIC) 1.05.01.85.09: “New light on the Intergalactic Medium (NewIGM)”. A.G. and F.F. acknowledge support from PRIN MIUR project “Black Hole winds and the Baryon Life Cycle of Galaxies: the stone-guest at the galaxy evolution supper”, contract 2017-PH3WAT. A.G. acknowledges the support of the INAF Mini Grant 2022 “Learning Machine Learning techniques to dig up high-*z* AGN in the Rubin-LSST Survey”. We thank Società Astronomica Italiana (SAIt), Ennio Poretti, Gloria Andreuzzi and Marco Pedani for the observation support at TNG. Part of the observations discussed in this work are based on observations made with the Italian Telescopio Nazionale *Galileo* (TNG) operated on the island of La Palma by the Fundación Galileo Galilei of the INAF (Istituto Nazionale di Astrofisica) at the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. This paper includes data gathered with the 6.5 meter *Magellan* Telescopes located at Las Campanas Observatory, Chile. 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 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, funded by the National Aeronautics and Space Administration. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen’s University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation. We acknowledge the support from the LBT-Italian Coordination Facility for the execution of observations, data distribution and reduction. The LBT is an international collaboration among institutions in the United States, Italy and Germany. The LBT Corporation partners are: The University of Arizona on behalf of the Arizona university system; Istituto Nazionale di Astrofisica, Italy; LBT Beteiligungsgesellschaft, Germany, representing the Max Planck Society, the Astrophysical Institute Potsdam, and Heidelberg University; The Ohio State University; The Research Corporation, on behalf of The University of Notre Dame, University of Minnesota and University of Virginia.

## References

- Allwein, E., Schapire, R., & Singer, Y. 2000, *J. Mach. Learn. Res.*, **1**, 113
- Atlee, D. W., & Gould, A. 2007, *ApJ*, **664**, 53
- Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, *MNRAS*, **490**, 5615
- Baqui, P. O., Marra, V., Casarini, L., et al. 2021, *A&A*, **645**, A87
- Barbisan, E., Huang, J., Dage, K. C., et al. 2022, *MNRAS*, **514**, 943
- Batista, G., Prati, R., & Monard, M.-C. 2004, *SIGKDD Explor.*, **6**, 20
- Boutsia, K., Grazian, A., Calderone, G., et al. 2020, *ApJS*, **250**, 26
- Boutsia, K., Grazian, A., Fontanot, F., et al. 2021, *ApJ*, **912**, 111
- Brescia, M., Cavuoti, S., Razim, O., et al. 2021, *Front. Astron. Space Sci.*, **8**
- Calderone, G., Boutsia, K., Cristiani, S., et al. 2019, *ApJ*, **887**, 268
- Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]
- Chen, T., & Guestrin, C. 2016, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (New York, NY, USA: ACM), 785
- Colless, M., Dalton, G., Maddox, S., et al. 2001, *MNRAS*, **328**, 1039
- Cristiani, S., Porru, M., Guarneri, F., et al. 2023, *MNRAS*, **522**, 2019
- Cupani, G., Calderone, G., Selvelli, P., et al. 2022, *MNRAS*, **510**, 2509
- D’Abrusco, R., Massaro, F., Paggi, A., et al. 2014, *ApJS*, **215**, 14
- D’Abrusco, R., Álvarez Crespo, N., Massaro, F., et al. 2019, *ApJS*, **242**, 4
- De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, *A&A*, **674**, A2
- Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. 2021, *J. Chem. Inform. Model.*, **61**, 2623
- Fernández, A., García, S., Galar, M., et al. 2019, *Learning from Imbalanced Data Sets* (Cham: Springer)
- Friedman, J. H. 2001, *Ann. Stat.*, **29**, 1189
- Grazian, A., Giallongo, E., Boutsia, K., et al. 2022, *ApJ*, **924**, 62
- Guarneri, F., Calderone, G., Cristiani, S., et al. 2021, *MNRAS*, **506**, 2471
- Guarneri, F., Calderone, G., Cristiani, S., et al. 2022, *MNRAS*, **517**, 2436
- Hughes, A. C. N., Bailer-Jones, C. A. L., & Jamal, S. 2022, *A&A*, **668**, A99
- Jin, X., Zhang, Y., Zhang, J., et al. 2019, *MNRAS*, **485**, 4539
- Jin, J.-J., Wu, X.-B., Fu, Y., et al. 2023, *ApJS*, **265**, 25
- Johnson, J. M., & Khoshgoftaar, T. M. 2021, in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1182
- Jones, D. H., Read, M. A., Saunders, W., et al. 2009, *MNRAS*, **399**, 683
- Khorunzhev, G. A., Burenin, R. A., Meshcheryakov, A. V., & Sazonov, S. Y. 2016, *Astron. Lett.*, **42**, 277
- Khramtsov, V., Sergeev, A., Spiniello, C., et al. 2019, *A&A*, **632**, A56
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. 2013, *J. Exp. Soc. Psychol.*, **49**, 764
- Liske, J., Grazian, A., Vanzella, E., et al. 2008, *MNRAS*, **386**, 1192
- Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, *ApJS*, **250**, 8
- McQuinn, M. 2016, *ARA&A*, **54**, 313
- Meiksin, A. A. 2009, *Rev. Mod. Phys.*, **81**, 1405
- Murphy, M. T., Molaro, P., Leite, A. C. O., et al. 2022, *A&A*, **658**, A123
- Nakazono, L., Mendes de Oliveira, C., Hirata, N. S. T., et al. 2021, *MNRAS*, **507**, 5847
- Nakoneczny, S. J., Bilicki, M., Pollo, A., et al. 2021, *A&A*, **649**, A81
- Niculescu-Mizil, A., & Caruana, R. 2005, in *ICML '05: Proceedings of the 22nd international Conference on Machine Learning*, 625
- Onken, C. A., Wolf, C., Bian, F., et al. 2022, *MNRAS*, **511**, 572
- Parmar, A., Kataria, R., & Patel, V. 2019, in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, eds. J. Hemanth, X. Fernando, P. Lafata, & Z. Baig (Cham: Springer International Publishing), 758
- Péroux, C., & Howk, J. C. 2020, *ARA&A*, **58**, 363
- Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. 2004, in *MICAI 2004: Advances in Artificial Intelligence*, eds. R. Monroy, G. Arroyo-Figueroa, L. E. Sucar, & H. Sossa (Berlin, Heidelberg: Springer Berlin Heidelberg), 312
- Prati, R., Batista, G., & Monard, M.-C. 2009, in *Paper presented at the IICAI*, 359
- Provost, F. J. 2000, in *AAAI Technical Report WS-00-05, Workshop on Imbalanced Data Sets*
- Provost, F., & Fawcett, T. 1997, in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43
- Reis, I., Baron, D., & Shahaf, S. 2019, *AJ*, **157**, 16
- Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, *ApJS*, **180**, 67
- Rodrigues, N. V. N., Raul Abramo, L., Queiroz, C., et al. 2023, *MNRAS*, **520**, 3494
- Rousseeuw, P. J., & Croux, C. 1993, *J. Am. Stat. Assoc.*, **88**, 1273
- Schindler, J.-T., Fan, X., Huang, Y.-H., et al. 2019a, *ApJS*, **243**, 5
- Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2019b, *ApJ*, **871**, 258
- Smith, M. R., Martinez, T. R., & Giraud-Carrier, C. G. 2013, *Mach. Learn.*, **95**, 225
- Trakhtenbrot, B. 2021, in *Nuclear Activity in Galaxies Across Cosmic Time*, eds. M. Pović, P. Marziani, J. Masegosa, H. Netzer, S. H. Negu, & S. B. Tessema, *IAU Symp.*, **356**, 261
- Véron-Cetty, M. P., & Véron, P. 2010, *A&A*, **518**, A10
- Wenzl, L., Schindler, J.-T., Fan, X., et al. 2021, *AJ*, **162**, 72
- Wolf, C., Hon, W. J., Bian, F., et al. 2020, *MNRAS*, **491**, 1970
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Yang, J., Wang, F., Wu, X.-B., et al. 2016, *ApJ*, **829**, 33
- Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. 2016, *Big Data Res.*, **5**, 2



## Appendix A: Confusion matrices

This section reports the confusion matrices for three specific runs of the direct selection (Sect. 5.3), direct selection with undersampling (Sect. 5.4), and the reverse selection (Sect. 5.5) methods. In all cases we used exactly the same training-test split. Relevant metrics highlighted in bold are (from top to bottom): the true positives (TP), the number of predicted positives (pP), the precision and the recall at  $z > 2.5$ . The latter two are also reported in Table 3 for an easy comparison of the performance of the three methods.

## Appendix B: List of newly classified high- $z$ QSO

The 107 new QSO classifications and redshift identified with the reverse selection method are reported in Table B.1 (available at the CDS).

**Table A.1.** Confusion matrix, precision, and recall for the direct selection method (Sect. 5.3).

		Predicted class:					TOTAL
		Galaxy	QSOhighZ	QSOlowZ	Star	other	
True class:	<b>Confusion matrix:</b>						
	Galaxy	26399	1	282	3428	6	30116
	QSOhighZ	0	<b>218</b>	129	35	0	382
	QSOlowZ	516	82	2571	79	6	3254
	Star	84	8	22	5597069	0	5597183
	other	408	1	15	5	5	434
	<b>TOTAL</b>	<b>27407</b>	<b>310</b>	<b>3019</b>	<b>5600616</b>	<b>17</b>	<b>5631369</b>
True class:	<b>Precision:</b>						
	Galaxy	96.3%	0.3%	9.3%	0.1%	35.3%	
	QSOhighZ	0.0%	<b>70.3%</b>	4.3%	0.0%	0.0%	
	QSOlowZ	1.9%	26.5%	85.2%	0.0%	35.3%	
	Star	0.3%	2.6%	0.7%	99.9%	0.0%	
	other	1.5%	0.3%	0.5%	0.0%	29.4%	
	<b>TOTAL</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	
True class:	<b>Recall:</b>						
	Galaxy	87.7%	0.0%	0.9%	11.4%	0.0%	100%
	QSOhighZ	0.0%	<b>57.1%</b>	33.8%	9.2%	0.0%	100%
	QSOlowZ	15.9%	2.5%	79.0%	2.4%	0.2%	100%
	Star	0.0%	0.0%	0.0%	100.0%	0.0%	100%
	other	94.0%	0.2%	3.5%	1.2%	1.2%	100%

**Table A.2.** Confusion matrix, precision, and recall for the direct selection method with undersampling (Sect. 5.4).

		Predicted class:					TOTAL
		Galaxy	QSOhighZ	QSOlowZ	Star	other	
True class:	<b>Confusion matrix:</b>						
	Galaxy	21622	41	793	807	6853	30116
	QSOhighZ	1	<b>362</b>	13	5	1	382
	QSOlowZ	67	346	2598	18	225	3254
	Star	183999	32456	8528	5371001	1199	5597183
	other	113	2	18	0	301	434
	<b>TOTAL</b>	<b>205802</b>	<b>33207</b>	<b>11950</b>	<b>5371831</b>	<b>8579</b>	<b>5631369</b>
True class:	<b>Precision:</b>						
	Galaxy	10.5%	0.1%	6.6%	0.0%	79.9%	
	QSOhighZ	0.0%	<b>1.1%</b>	0.1%	0.0%	0.0%	
	QSOlowZ	0.0%	1.0%	21.7%	0.0%	2.6%	
	Star	89.4%	97.7%	71.4%	100.0%	14.0%	
	other	0.1%	0.0%	0.2%	0.0%	3.5%	
	<b>TOTAL</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	
True class:	<b>Recall:</b>						
	Galaxy	71.8%	0.1%	2.6%	2.7%	22.8%	100%
	QSOhighZ	0.3%	<b>94.8%</b>	3.4%	1.3%	0.3%	100%
	QSOlowZ	2.1%	10.6%	79.8%	0.6%	6.9%	100%
	Star	3.3%	0.6%	0.2%	96.0%	0.0%	100%
	other	26.0%	0.5%	4.1%	0.0%	69.4%	100%

**Table A.3.** Confusion matrix, precision, and recall for the reverse selection method (Sect. 5.5).

<b>Confusion matrix:</b>		<b>Predicted class:</b>					<b>TOTAL</b>
		<b>Galaxy</b>	<b>QSOhighZ</b>	<b>QSOlowZ</b>	<b>Star</b>	<b>other</b>	
<b>True class:</b>	Galaxy	25628	8	657	3389	434	30116
	QSOhighZ	0	<b>327</b>	29	24	2	382
	QSOlowZ	277	203	2615	77	82	3254
	Star	772	54	31	5596316	10	5597183
	other	322	1	18	2	91	434
	<b>TOTAL</b>	<b>26999</b>	<b>593</b>	<b>3350</b>	<b>5599808</b>	<b>619</b>	<b>5631369</b>
<b>Precision:</b>		<b>Galaxy</b>	<b>QSOhighZ</b>	<b>QSOlowZ</b>	<b>Star</b>	<b>other</b>	
<b>True class:</b>	Galaxy	94.9%	1.3%	19.6%	0.1%	70.1%	
	QSOhighZ	0.0%	<b>55.1%</b>	0.9%	0.0%	0.3%	
	QSOlowZ	1.0%	34.2%	78.1%	0.0%	13.2%	
	Star	2.9%	9.1%	0.9%	99.9%	1.6%	
	other	1.2%	0.2%	0.5%	0.0%	14.7%	
	<b>TOTAL</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	
<b>Recall:</b>		<b>Galaxy</b>	<b>QSOhighZ</b>	<b>QSOlowZ</b>	<b>Star</b>	<b>other</b>	<b>TOTAL</b>
<b>True class:</b>	Galaxy	85.1%	0.0%	2.2%	11.3%	1.4%	100%
	QSOhighZ	0.0%	<b>85.6%</b>	7.6%	6.3%	0.5%	100%
	QSOlowZ	8.5%	6.2%	80.4%	2.4%	2.5%	100%
	Star	0.0%	0.0%	0.0%	100.0%	0.0%	100%
	other	74.2%	0.2%	4.1%	0.5%	21.0%	100%