

Wide Area VISTA Extragalactic Survey (WAVES): Selection of targets for the Wide survey using decision-tree classification

G. Kaur¹, M. Bilicki¹, S. Bellstedt², E. Tempel³, W. A. Hellwing¹,
I. Baldry⁴, B. Bandi⁵, S. Barsanti^{6, 7}, S. Driver², N. Guerra-Varas⁸,
B. Holwerda⁹, C. Lagos², J. Loveday⁵, and A. Robotham²

¹Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników, 02-685, Warsaw
e-mail: [gursharanjitk, bilicki]@cft.edu.pl

²ICRAR, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

³Tartu Observatory, University of Tartu, Observatooriumi 1, 61602 Tõravere, Estonia

⁴Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool, L3 5RF, UK

⁵Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK

⁶Sydney Institute for Astronomy, School of Physics, University of Sydney, NSW 2006, Australia

⁷Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia

⁸European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching near Munich, Germany

⁹Physics & Astronomy Department, University of Louisville, Louisville, KY 40292, USA

ABSTRACT

The Wide-Area VISTA Extragalactic Survey (WAVES) on the 4-metre Multi-Object Spectroscopic Telescope (4MOST) includes two flux-limited subsurveys with very high (95%) completeness requirements: Wide over $\sim 1200 \text{ deg}^2$ and Deep over $\sim 65 \text{ deg}^2$. Both are Z-band selected, respectively as $Z < 21.1$ and $Z < 21.25$ mag, and additionally redshift-limited, while the true redshifts are not known a priori but will be only measured by 4MOST. Here, we present a classification-based method to select the targets for WAVES-Wide. Rather than estimating individual redshifts for the input photometric objects, we assign probabilities of them being below $z = 0.2$, the redshift limit of the subsurvey. This is done with the supervised machine learning approach of eXtreme Gradient Boosting (XGB), trained on a comprehensive spectroscopic sample overlapping with WAVES fields. Our feature space is composed of nine VST+VISTA magnitudes from u to K_s and all the possible colors, but most relevant for the classification are the g -band and the $u - g$, $g - r$ and $J - K_s$ colors. We check the performance of our classifier both for the fiducial WAVES-Wide limits, as well as for a range of neighboring redshift and magnitude thresholds, consistently finding purity and completeness at the level of 94-95%. We note, however, that this performance deteriorates for sources close to the selection limits, due to deficiencies of the current spectroscopic training sample and the decreasing signal-to-noise of the photometry. We apply the classifier trained on the full spectroscopic sample to 14 million photometric galaxies from the WAVES input catalog, which have all 9 bands measured. We then check what XGB probability threshold should be adopted to obtain a feasible number of target candidates, finding that $P_{\text{XGB}}(z \leq 0.2) > 0.91$ would give about 1 million targets selected with our approach. Our work demonstrates that a machine-learning classifier could be used to select a flux- and redshift-limited sample from deep photometric data. Within WAVES, the probabilities we have derived are combined with a number of other approaches for the final Wide target selection employed for the forthcoming 4MOST observations.

Key words. Target Selection – Machine Learning – photometric redshift

1. Introduction

The Wide-Area VISTA Extragalactic Survey (WAVES, Driver et al. 2019) is one of the observational campaigns to be undertaken on the 4-metre Multi-Object Spectroscopic Telescope (4MOST, de Jong et al. 2019) – a major endeavor of the European Southern Observatory (ESO) to spectroscopically map the entire southern sky, starting from 2026. WAVES aims to study the growth of structure, mass, and energy on scales from 1 kpc to 10 Mpc over the last 7 Gyr. At the largest scales (1-10 Mpc), it will examine structures like groups, filaments, and voids, and their recent development, comparing results with numerical simulations to evaluate the Cold Dark Matter paradigm. At intermediate scales (10 kpc - 1 Mpc), WAVES will assess sizes, mass distribution, and merger rates of galaxy groups to understand dark matter halo and stellar mass assembly. At the smallest scales (1-10 kpc), it will provide precise distance and environ-

mental data to complement high-resolution space imaging for studying the evolution of galaxy bulges, discs, and bars. Overall, WAVES will create a comprehensive spectroscopic dataset of 1.6 million galaxies, connecting the low- ($z < 0.1$) and intermediate- ($z < 0.8$) redshift Universe.

To achieve its goals, WAVES is designed in two main subsurveys: Wide and Deep. WAVES-Wide (WW) aims to obtain spectra of a highly complete ($> 90\%$) sample of galaxies, limited by the Z-band magnitude $Z \lesssim 21.1$ and redshift $z \lesssim 0.2$, over most ($\sim 1200 \text{ deg}^2$) of the joint area of the Kilo-Degree Survey (KiDS, de Jong et al. 2013) and VISTA Kilo-Degree Infrared Galaxy Survey (VIKING, Edge et al. 2013). WAVES-Deep (WD) is planned to reach to $Z \lesssim 21.25$ mag and $z \lesssim 0.8$ over $\sim 65 \text{ deg}^2$ area composed of the Galaxy And Mass Assembly (GAMA) G23 field and four ‘Deep Drilling Fields’ (DDF) of the Legacy Survey

of Space and Time (LSST) at the Rubin Observatory (Ivezić et al. 2019).

In both Wide and Deep cases, the WAVES target selection is flux- and redshift-limited. The former condition is straightforward to achieve, as fluxes are readily available from internal WAVES processing of the KiDS+VIKING (KV) data (Bellstedt et al. in prep.). The challenge lies in assigning redshift limits in surveys, which by design aim at measuring the redshifts themselves. To accomplish this, WAVES uses multiband KV photometry as a starting point.

The selection of WAVES targets has two stages. First, extragalactic sources are separated from stars. Then the selection of the redshift-limited galaxy samples follows. The star-galaxy separation was originally performed in a similar way as for the GAMA fields (Bellstedt et al. 2020), i.e. by locating the objects on color/magnitude/size planes. This has been revised for WAVES purposes by Cook et al. (2024) via an unsupervised machine-learning approach. For the final target catalogs, this is now being updated by using flags from the most recent runs of photo- z estimation codes. In this work, we will assume that we are given a photometric catalog of galaxies and we will not be dealing with possible star contamination or missing galaxies. The second stage of WAVES selection is to choose the photometric galaxy targets for spectroscopic observations within specific redshift ranges. This is typically based on multi-band magnitudes and colors (e.g. Eisenstein et al. 2011; Guzzo et al. 2014; DESI Collaboration et al. 2016), which is especially efficient if specific galaxy types, such as luminous red galaxies, are required (e.g. Prakash et al. 2016). The scientific goals of WAVES require, however, all possible galaxies to be observed, without any additional color- or galaxy-type preselections, as long as they are within the assigned magnitude and redshift limits. What is more, the science goals of the survey put stringent requirements of over 90% spectroscopic completeness to be achieved (Driver et al. 2019). Especially for WW, this would be very difficult to meet with color selection, unless one allows for many observed targets to be beyond the $z \approx 0.2$ limit. Due to limited fibre-hour allocation, such contamination cannot be excessive without affecting both WAVES and the entire 4MOST survey efficiency.

Multi-color information could be gathered into the prediction of photometric redshift (photo- z , e.g. Newman & Gruen 2022), which could then be used for more efficient redshift-dependent target selection. In 4MOST, this is employed for one of the WAVES subsurveys, Optical, Radio Continuum and HI Deep Spectroscopic Survey (ORCHIDSS, Duncan et al. 2023) in addition to radio selection. Another 4MOST community survey, CHileAN Cluster galaxy Evolution Survey (CHANCES, Haines et al. 2023) makes use of photo- z based selections for its low- z survey. In WW, photo- z s are employed by various approaches for target selection: Gaussian processes for photometric redshifts (GPz, Duncan 2022), scaled flux matching (SFM, Baldry et al. 2021), and Tartu Observatory Photo- z s (TOPz, Laur et al. 2022), the results of which are then combined, together with our method, into the final target probability (Tempel et al. in prep.).

However, even the best photo- z methodologies using broad bands, as we have here, do not give sufficient precision to meet the requirements of WW in terms of completeness while preserving purity. Indeed, state-of-the-art results employing KV magnitudes by Bilicki et al. (2021) give $\sigma_z \sim 0.018(1+z)$ at a median $z \sim 0.23$ albeit at brighter magnitudes ($r < 20$) than in WW. Deeper flux-limited samples would likely have even larger photo- z scatter (e.g. Porredon et al. 2021). This could be improved by employing deep learning using galaxy images for photo- z predictions (e.g. Treyer et al. 2024; John William et al. 2025), but the overall

scatter is unlikely to be much smaller than $\sigma_z \sim 0.015(1+z)$. At the WW limiting redshift of $z = 0.2$, this amounts to some 10% relative error, which could be crudely translated to $\sim 90\%$ completeness and purity if such photo- z s are employed for target selection.

In this work, we take a different approach towards assigning WW targets. As the problem reduces to answering the question of whether a source is above or below the redshift threshold, this is basically a classification problem. It could be regarded as binary classification, or just one-class assignment, as what is relevant is to correctly assign a galaxy to the class where $z < z_{\text{lim}}$. Our method to select the WW targets therefore classifies photometric galaxies as either meeting the redshift threshold criterion or not. We solve this classification problem by using an empirical approach of supervised ML. A similar idea was applied in Wu et al. (2022), where convolutional neural networks were employed on DESI Legacy images to identify $z < 0.03$ galaxies, and more recently in Payerne et al. (2025), where a random forest (RF) was used to select high-redshift Lyman-Break Galaxies from current and future broadband wide photometric surveys.

Our goal is to assign class "0" to the desired WW targets (below the chosen Z -magnitude and redshift thresholds) and class "1" otherwise. For that purpose, we use a tree-based algorithm, which is fast and not very demanding computationally, while at the same time well-suited to such binary classification problems. Moreover, our approach allows us to estimate the output probability of an object being assigned to a given class. Working with a flux-limited sample, we build a supervised ML classifier that is trained on galaxies having both the KV photometry and redshift labels from external spectroscopy. The classifier indirectly learns the color-redshift relation, but rather than providing point redshift estimates, it gives the probability that a given galaxy lies below the desired redshift cut.

Tree-based methods are abundantly used for classification and regression tasks in a similar context. Some examples include Gerdes et al. (2010), who designed a photo- z algorithm based on boosted decision trees; Carrasco Kind & Brunner (2013), who employed RF to estimate photo- z s and their probability density functions; Zhou et al. (2021), where RF was used to estimate photo- z s for the DESI Legacy galaxy sample; Signor et al. (2024), where the reddest high-redshift galaxies in the Euclid Deep Fields were identified with gradient-boosted trees; or already mentioned Payerne et al. (2025). Here we employ the XGBoost algorithm (Chen & Guestrin 2016) trained on data with 9-band photometric magnitudes together with the corresponding 36 colors (45 features in total) and true (spectroscopic) redshift labels that are turned into the two classes (targets or non-targets). Our choice follows after testing various other ML approaches, such as RF, k-nearest neighbors (kNN), or support vector machines. Similarly, as for some other methods, e.g. kNN, an advantage of tree-based algorithms is their ability to provide not only a discrete class assignment but also its probability. This could be folded into the target selection, for instance, by giving a higher chance of observation to those galaxies where such probabilities are the largest (Tempel et al. 2020). Another reason for the choice of a tree classifier was the need to get target probability assignments for different selection criteria, as well as short training time, making it computationally efficient.

XGBoost that we employ is a supervised classifier. A general drawback of such methods is their dependence on the training data with true labels. If the training data are incomplete, then in the areas of the parameter space which they do not cover well, the classification might be unreliable. This does affect us as well, and we will discuss the implications below, in relation to the

probabilities assigned by the final classifier. In particular, if a too permissive cut on classification probability is applied, the number of resulting target candidates becomes too large to be practically observed. This is likely driven by particular contamination from higher-redshift galaxies whose colors are similar to those of true targets, but their color-redshift relation is not well mapped in the training set. Such contaminants can, however, be removed by raising the classification probability threshold, which leads to lowering the proposed target number.

The XGBoost-based classification pipeline we present here has been adopted by WAVES for the selection of WW targets within a joint framework employing also other approaches (Baldry et al. 2021; Duncan 2022; Laur et al. 2022). The details are/will be presented in a separate/companion paper by Tempel et al. (in prep.), while here we focus only on our particular method without comparing it with the others.

The paper is organized as follows. In Sec. 2, we describe the WAVES photometric data along with the spectroscopic compilation, which is used to train the XGBoost classifier. In Sec. 3, we outline the challenges associated with target selection for WAVES Wide. The gradient tree-based target selection pipeline and the classifier training are presented in Sec. 4. The performance of the classification pipeline is shown in Sec. 5. In Sec. 6, we apply the trained XGBoost classifier to label all the galaxies in the WAVES photometric catalog and assign target probabilities. Finally, Sec. 7 provides a summary and conclusions.

2. Data

2.1. WAVES photometric data

Our feature space is built from 9-band optical+near-infrared (NIR) magnitudes available in the WAVES photometric catalog. While for the final (inference) dataset we use only the WW coverage, for the training phase we supplement that with photometric measurements in the WD fields. This is necessary because most of the key deep spectroscopic surveys required for our calibration are located only in the Deep fields, while the Wide ones are covered mostly by brighter-end samples.

The WW catalog is based on ESO VST KiDS plus ESO VISTA VIKING public surveys covering the $ugriZYJHK_s$ bandpasses. The input imaging is basically the same as was used to produce the KiDS final Data Release 5 (Wright et al. 2024). However, the image processing and source detection for WAVES are separate from KiDS. The general framework and tools are the same as those used in GAMA (Bellstedt et al. 2020), with some modifications to masking and fragmenting procedures. Detection images, deeper than those used in Bellstedt et al. (2020), are created through an $r + i + Z + Y$ stack, and then passed through the ProFound package (Robotham et al. 2011) for source detection and characterization. A ‘defragmenting’ process is then conducted mostly automatically (with manual inspection and intervention when required) to piece segments back together for sources that have been fragmented. ProFound is then run in a ‘measure’ mode to extract photometry for each source in all 9 bands separately. Artefacts in the image are screened out based on a series of logical arguments based on non-physical colors and sizes. More details will be presented in Bellstedt et al. (in prep.).

The magnitudes (fluxes) in WAVES photometric catalogs come in two flavors: ‘total’ and ‘color’, the latter derived using the detection segment, whereas the total photometry is generated using a dilated segment to capture the full flux of the galaxy. Total photometry is more appropriate for e.g. source selection, whereas the color fluxes have slightly lower noise and hence are usually

better for tasks like photometric redshifting. Similarly as was the case for the other WW target selection approaches mentioned in Sec. 1, we tested both photometry versions and found that the ‘color’ option gives consistently improved results for our classification, by about 1 percent point in all the metrics, we will therefore be using it henceforth. Avoiding ‘total’ photometry makes our approach also immune to possible issues with flux extrapolation that could be present for such measurements.

The WD fields consist of four individual pointings, each spanning 4 square degrees within the VRO LSST DDF Regions and the G23 field. These four fields – WD01, 02, 03, and 10 – correspond to ELAIS-S1, XMMLSS, CDFS, and COSMOS, respectively. The data in these fields are more complex than in WW, and sourced from a diverse range of public and private survey programs conducted with VST and VISTA, including KiDS, KIDZ, WAVES-VST, VOICE (Vaccari et al. 2016), VIDEO (Jarvis et al. 2013), ultraVISTA (McCracken et al. 2012), and WAVES-VISTA (Bellstedt et al. in prep.).

Our ML model requires all the bands to be measured; therefore, in the photometric data, we keep only those objects that have all nine fluxes above 0. This affects primarily the shallowest u -band: in the entire input catalog limited to $Z < 21.1$ mag (14 million galaxies), about 9.8% of the sources do not have flux reported in this filter, while only 1.3% miss at least one of the other bands. Our final choice of requiring all the 9 bands is driven by the fact that first, we use both magnitudes and colors, so one missing band affects 8 further features, and second, we tested that ignoring the u -band worsens our classification metrics. This is because one of the key features for selecting $z < 0.2$ galaxies with our approach is the $u - g$ color, as quantified later in Sec. 5.

We also note that, although we use the ‘color’ magnitudes, the flux limits themselves are imposed according to the ‘total’ photometry. Moreover, we use only objects flagged as galaxies according to Cook et al. (2024). As a result, we are left with a dataset containing 12,781,357 galaxies with 9-band detections, for the $Z < 21.1$ mag limit. This will be our inference sample to which we will apply the trained ML classifier. The inference set is referred to as the photo-cat hereafter.

2.2. Spectroscopic data with true labels

Our ML model detailed below in Sec. 4 is a supervised one, which means that it requires labels with ground truth to be trained and validated on. These labels – class “0” for galaxies below the desired redshift threshold, and class “1” otherwise, both for the same fixed magnitude limit – are derived from spectroscopic samples cross-matched with WAVES imaging. These samples cover both the main KiDS/VIKING/WW area and the WAVES-DDFs, where appropriate optical and NIR photometry is available. The input spectroscopic surveys and datasets are the same as discussed in Jan et al. (2024) and Wright et al. (2024) and we join them into one parent catalog in the same way as described in this former paper¹. The procedure includes in particular appropriate handling of objects present in more than one source spectroscopic dataset as well as redshift quality assignment. In what follows, we use only galaxies with quality flag $NQ \geq 3$ to ensure reliable redshifts, where NQ is a redshift quality flag similar to that employed by GAMA, homogenized based on information from the input surveys (see Wright et al. 2024 for details). The resulting

¹ Please note that our compilation does not include DESI DR1 (DESI Collaboration et al. 2025), which appeared in March 2025, when this project was already at final stages. These new data from DESI will be used by WAVES for further analyses.

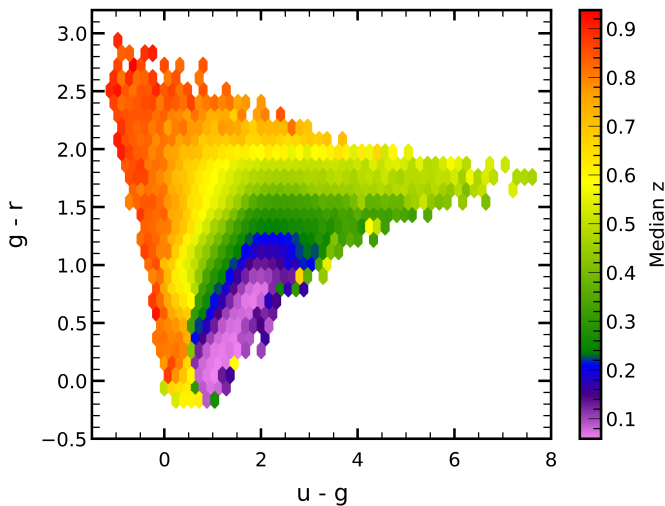


Fig. 1. Distribution of galaxies from the spectroscopic catalog, that we use to calibrate our classifier, on the $u - g$ vs. $g - r$ color plane. Each pixel is colored according to its median redshift. We show only the pixels with a minimum of 10 galaxies each.

imaging dataset cross-matched with spectroscopic labels is hereafter referred to as spec-z-cat. At the fiducial Z-band magnitude limit of 21.1, the median redshift for the spec-z-cat is 0.24, while the median Z-band magnitude is 18.6. Key information on the constituent spectroscopic surveys for this magnitude limit is presented in Table 1, while the details of the selection criteria in the particular datasets and how they were combined are provided in sec. 5 of Wright et al. (2024) and table 1 of Jalan et al. (2024).

We note that our spec-z-cat includes also objects at fainter magnitudes, and we use these to study our classifier's performance at various flux limits in Sec. 5.3. Figure 1 shows an example color-color distribution ($u - g$ vs. $g - r$) of the spec-z-cat mapped by the median spectroscopic redshift. We see that low-redshift galaxies occupy a specific area on this diagram, although with some overlap with the high- z ones which is due to the complicated color-redshift relation. We choose to display this particular plane as these two colors are among the most important features for our classification (Sec. 5). We note however that the ML model uses the entire magnitude-color space and similar separation could be found also in other cross-sections. One further point to note regarding such diagrams is that their appearance is also driven by selection effects in the spectroscopic data that we have available; see below for more discussion.

3. Challenges in selecting WAVES-Wide targets

One of the main challenges of supervised classification is the appropriate selection of training data. Typically, when working with real observational samples, the possibilities to obtain training sets that are sufficiently representative for the target (inference) catalog, are limited. More common are situations when the former is a biased, often brighter and color-selected, subsample of the latter. This is indeed what we are facing here as well for the target selection in WW. Namely, at the planned WW photometric depth of $Z \sim 21.1$ mag, galaxies with $z \lesssim 0.2$ constitute a minority, which is not correctly reflected in the labeled data we have available.

To quantify this effect, we start with a simulation dataset where we know both magnitudes and redshifts. For that we employ a mock galaxy catalog built with the Shark semi-analytic

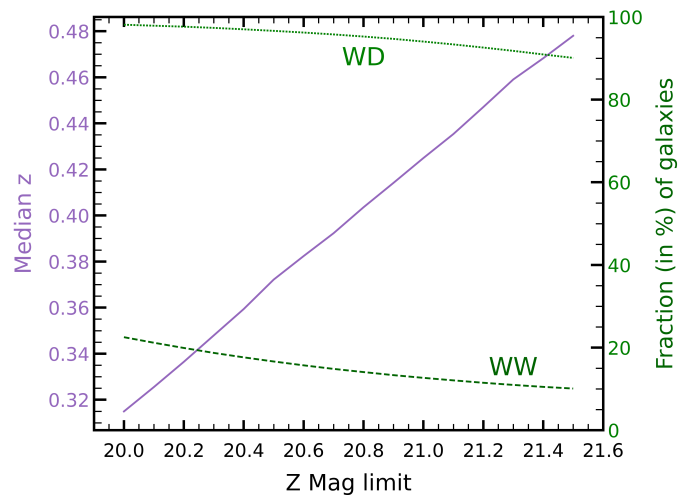


Fig. 2. Dependence of the median redshift (left-hand axis, violet diagonal line) of galaxies selected as flux-limited up to the Z-band magnitude indicated on the x-axis, as derived from a Shark/SURFS mock galaxy catalog. The right-hand axis and the green lines show what percentage of galaxies would constitute the WAVES targets for the Z-band flux limit as on the x-axis and the fiducial redshift limits: $z < 0.2$ for WW (bottom green line) and $z < 0.8$ for WD (top green line).

model (Lagos et al. 2018) applied to the SURFS N-body simulations (Elahi et al. 2018), which was tailored to match the WAVES input data. The Shark semi-analytic model of galaxy formation generates self-consistent predictions from the far-ultraviolet to the far-infrared. Shark covers a wide range of baryon physics effects that are believed to play an important role in the formation and evolution of galaxies. For this paper, we use a lightcone generated from the default model presented in Lagos et al. (2018) using the method described in Chauhan et al. (2019).

Lagos et al. (2019) and Lagos et al. (2020) presented the FUV-to-FIR multi-wavelength predictions of Shark, by combining the model with the generative SED model ProSpect (Robotham et al. 2020), and using a novel method to compute dust attenuation parameters based on the radiative transfer analysis of the EAGLE hydrodynamical simulations of Trayford et al. (2020). The model was shown to produce excellent agreement with the observed number counts from the FUV to the FIR (Lagos et al. 2019), and the redshift distribution of FIR-selected sources (Lagos et al. 2020). Here, we use a lightcone produced for WAVES following the same pipeline described in Lagos et al. (2019).

In particular, our mock sample includes the 9-band flux information mimicking the KV measurements, as well as the "true" galaxy redshifts, although with a caveat that it is limited to $Z = 21.5$ mag. We use these to generate Fig. 2, where we show two pieces of information. The violet diagonal line, corresponding to the left-hand y-axis, illustrates the median redshift of a galaxy sample selected as flux-limited to the Z-band magnitude provided on the x-axis. We can see that at $Z \sim 21$ mag, this median z is well above the WW target limit of $z \lesssim 0.2$, which means that the majority of galaxies at this magnitude cut are at redshifts larger than our targets should have. Indeed, this is quantified with the lower green line, corresponding to the right-hand axis, and showing the percentage of $z < 0.2$ galaxies brighter than the Z-band limit of x-axis. Even for the brightest considered threshold of $Z < 20$ mag, this ratio is close to 20% and decreases for fainter flux limits. In contrast, as shown with the top green line, for WD with $z < 0.8$ selection, more than 90% of galaxies at the considered flux limits would be appropriate targets. Therefore,

Table 1. Input surveys included in the spectroscopic compilation that we use as the training and test sets for our machine-learning classification.

Survey	Median redshift	Median Z-mag	Count	References
2dFGRS	0.113	17.26	82552	Colless et al. (2001)
GAMA	0.217	18.47	228163	Baldry et al. (2010); Liske et al. (2015)
2dFLenS	0.297	18.30	35704	Blake et al. (2016)
HCOSMOS	0.305	19.45	1008	Damjanov et al. (2018)
G15-DEEP	0.325	20.26	1479	Driver et al. (2022)
C3R2	0.363	20.62	165	Masters et al. (2017, 2019)
DEVILS	0.372	20.07	9212	Davies et al. (2018)
DESI EDR	0.382	19.71	84915	DESI Collaboration et al. (2024)
G10-COSMOS	0.391	20.42	7971	Davies et al. (2015)
VVDS	0.440	20.51	1213	Le Fèvre et al. (2005, 2013)
SDSS DR14	0.441	18.65	85631	Abolfathi et al. (2018)
OzDES	0.467	19.78	5460	Lidman et al. (2020)
DEIMOS	0.517	20.37	642	Hasinger et al. (2018)
GOODS	0.521	20.41	178	Popesso et al. (2009); Balestra et al. (2010)
ACES	0.521	20.39	1384	Cooper et al. (2012)
WIGGLEZ	0.583	20.49	21027	Drinkwater et al. (2010)
zCOSMOS	0.610	20.13	553	priv. comm. (M. Salvato)
VIPERS	0.629	20.62	4907	Scoddeggio et al. (2018)
LEGA-C	0.755	20.60	148	van der Wel et al. (2016)
full spec-z-cat	0.242	18.60	572325	

The table is sorted in ascending median redshift of sources in the spec-z-cat. All the numbers apply after cross-matching the input spec-z samples with the photometric data limited at $Z < 21.1$ total mag and removing duplicates between surveys. We list only those surveys that have at least 100 cross-matched objects each.

the task of target selection for WW is much more challenging than for WD.

One additional caveat we need to highlight is that our mocks reach only to $z = 1$, so in fact the violet diagonal one will be slightly shifted upwards in real data, while the green lines are overestimates. These should not be big effects, though, as the fraction of $z > 1$ galaxies at the flux limits of interest is small.

The second challenge we are facing when designing the WW target selection is related to the limited completeness of currently available calibration data with true spectroscopic labels. This will be relevant for any methods of WW selection; even if ‘training’ data are not needed (e.g. in template-fitting approaches), one still needs to verify if a given target assignment indeed yields correct results. In the best-case scenario, one would use a complete flux-limited galaxy subsample with true (spectroscopic) redshift labels reaching at least to the WAVES photometric limit on a sufficiently large sky area, to provide good statistics and minimize cosmic variance. As discussed in Sec. 2.2, at present such samples are not available. We have either wide-angle datasets that are relatively deep but color-selected, or flux-limited but shallow, or very deep but small-area spectroscopic samples. Figure 3 illustrates the resulting difference in coverage of the $u - g$ vs. $g - r$ color-color plane by the full photometric catalog and the spec-z-cat.

We illustrate the incompleteness of the spec-z-cat with respect to the photo-cat in Figs. 4 and 5. In the former, we compare the Z-mag distribution (number counts) of our spectroscopic calibration set (yellow line) with that of the full parent WAVES photometric catalog (blue line). As the spec-z-cat has complicated geometry and we do not know its exact mask, we do not normalize the number counts by sky area; instead, we roughly match them at the bright-end where both spec- and photo-data follow a power-law. We see that the spec-z-cat starts to depart from this behavior already at $Z \sim 19$ mag, and its amplitude at the limit of WW is for such normalization more than an order of magnitude smaller than that of the photo-cat.

In Fig. 5 we compare the redshift distribution of the spec-z-cat with that of the mock, as (obviously) we do not have true redshifts for the entire WAVES input catalog. Both the datasets are flux-limited to $Z < 21.1$ mag. The normalization is again not by the sky area but roughly matched at low redshift where both for mocks and the spec-z-cat we have $dN/dz \propto z^2$ approximately. Similarly to the magnitudes, also the redshift distribution of the spec-z-cat follows the ‘true’ (mock) one only at the low- z end and starts departing already at $z \sim 0.15$. As mentioned in Sec. 2.2, the median redshift of the spec-z-cat for $Z < 21.1$ mag is equal to 0.24, while, as shown in Fig. 2, a complete galaxy sample at this flux limit would have $\langle z \rangle \simeq 0.43$. As Figs. 4 and 5 indicate, this mismatch is driven by an underabundance of faint, higher-redshift sources in the spec-z-cat.

The lower completeness of the spectroscopic training data at fainter, higher-redshift range, as compared to the photometric catalog, will have implications for our supervised learning approach, as we will show below. One particular effect is that the classifier would perform better if a brighter flux cut was adopted instead of the WW fiducial one. In addition, in the final inference, when target probabilities are assigned to the entire photo-cat, our classifier will turn out rather permissive by giving probabilities larger than 0.5 to 1.7 million more galaxies than the original estimate of 0.9 million (Driver et al. 2019). This is likely driven by the fact that many of the photo-cat galaxies lie outside the color space probed by the training data. A direct solution to this problem is by taking a target probability cut well above the fiducial value of 0.5. In the longer term, a possible mitigation is by providing more training examples populating different regions of the color space than what our current spec-z-cat offers, in a similar way as done in the Complete Calibration of the Color-Redshift Relation (Masters et al. 2015; Gruen et al. 2023; McCullough et al. 2024).

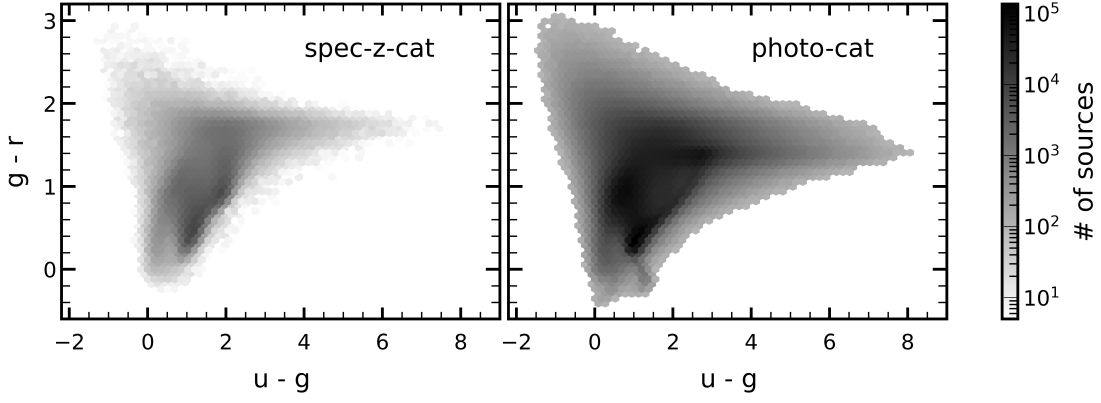


Fig. 3. Number density of galaxies projected on the $u - g$ vs. $g - r$ color plane, respectively for the spec-z-cat (left) that we use to calibrate our ML model, and for the entire WAVES photo-cat (right), from which WW targets are selected. We show only the pixels with a minimum of 5 galaxies each for the spec-z-cat (left) and a minimum of 100 sources each for the photo-cat (right).

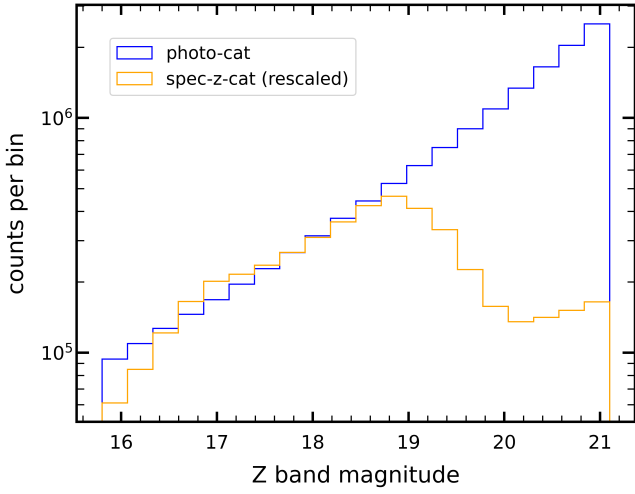


Fig. 4. Comparison of number counts as a function of the Z-band magnitude in the full WAVES photometric catalog (blue line) and the training set with spectroscopic redshift labels (yellow line). The scale on the y-axis gives actual counts per bin for the photo-cat, while for the spec-z-cat, the counts have been artificially rescaled to give a good match at the bright-end.

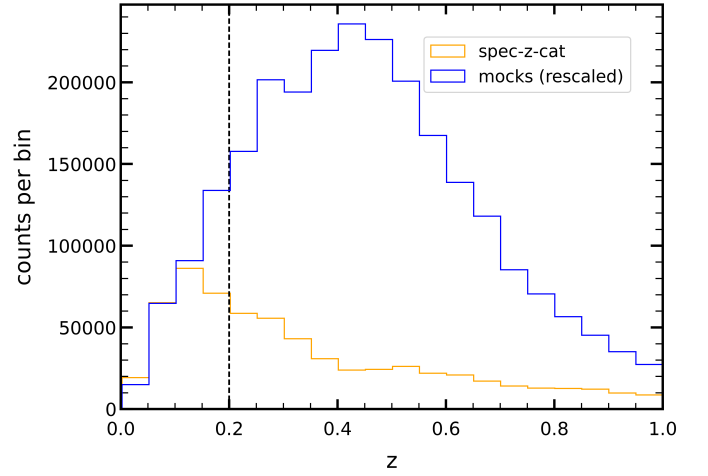


Fig. 5. Comparison of redshift distributions between our spectroscopic training set (yellow) and the SHARK mock catalog (blue), for galaxies at the flux limit of $Z < 21.1$ mag. The vertical line indicates the fiducial WW redshift limit for the targets. The redshift distribution of the mock catalog is rescaled to match the counts of the spec-z-cat at low redshift.

4. Methodology

In this Section, we describe how we selected the ML algorithm for the classification pipeline – XGBoost – as well as the algorithm and the pipeline itself.

4.1. Decision Trees for classification

Decision trees are a class of supervised ML algorithms that have a hierarchical structure. Standard decision trees consider all input features and split the data into subsets based on these features. In the most commonly used CART (Classification And Regression Trees, Breiman et al. 1984), the split is decided by the Gini impurity, which calculates the probability of misclassifying a random data point into the wrong subset or class label. The Gini impurity is given by (Breiman et al. 1984):

$$\text{Gini Impurity} = 1 - \sum_i P_i^2. \quad (1)$$

Here, P_i is the probability of the data point belonging to class " i ", the classes in our case being galaxies with $z \leq z_{\text{lim}}$ for

class "0", or $z > z_{\text{lim}}$ for class "1". The probability P_i (and $1 - P_i$) is calculated as the fraction of galaxies with $z_{\text{true}} \leq z_{\text{lim}}$ (and $z_{\text{true}} > z_{\text{lim}}$ respectively) within the subset generated after the split. For the true redshift label we use the spectroscopic measurement from the calibration sample.

The data is then divided into two subsets (in the case of binary classification) based on this best split, and each subset forms a child node. This process of evaluating and splitting continues recursively, creating a tree-like structure with branches and nodes. The recursion terminates when a stopping criterion is met, such as reaching a predefined maximum depth, having fewer than a predefined minimum number of samples per node, achieving a decrease in impurity below a certain threshold, or when all the samples in a node belong to the same class for classification. In such a situation, the node is designated as a leaf node. In a classification tree, the leaf node assigns a class label, typically the majority class of the samples in that node. The probability of the classification for galaxies to be in a particular class in the leaf node is assigned as the fraction of galaxies falling within the class out of the total number of galaxies in the node.

When looking for the most suitable approach for our binary classification, we tested several methods. Following early tests, we focused on two popular tree-based ML algorithms: Random Forest (RF, Breiman 2001) and XGBoost (XGB, Chen & Guestrin 2016). After a number of experiments, we decided to adopt the latter for several reasons. First, XGB-based classification models turned out to be less prone to overfitting as compared to RF. Second, XGB did not require much hyperparameter tuning, and it worked well already with default parameters. Third, the training time for XGB classifiers was significantly shorter than for the RF ones. Apart from supervised ML methods, we also tried the unsupervised approach k -nearest neighbors (Fix & Hodges 1951). However, the results were not satisfactory. The reason could be that a classification like this is not "inherent" for the photometric feature space, unlike the star-galaxy separation.

The models were implemented using the `scikit-learn` library of Python (Pedregosa et al. 2011). For XGB, the `xgboost` library was applied through the Scikit Learn API.

4.2. XGBoost

Like other supervised learning algorithms, XGB attempts to learn a function, $f(x)$, that maps input features x to output labels y . The function f is discrete for classification, and output labels are just (0, 1) for our binary case. XGB creates an ensemble of CART to minimize a loss function. The default loss function for XGB binary classification, which is also used in this work, is Log-Loss or binary cross-entropy. In the XGB library, this is specified as 'binary:logistic' and is given by:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i) \quad (2)$$

where y is the class label 0 or 1, and $\hat{y}$ is the probability of the output label being 0 or 1 assigned by the leaf node of the tree. XGB minimizes the loss function combined with regularization penalties by constructing decision trees sequentially to minimize the gradient of the loss function. This process of adding new CART, or models in general, one after another, is termed as boosting. The regularization terms help prevent overfitting of the model to training data.

In the RF algorithm, all the trees are created simultaneously instead of iteratively, and the majority vote of the 0 and 1 or yes and no decisions of all trees is the final output. In XGBoost, the leaf node holds a real number – class score – used to evaluate potential splits. The sum of scores of all trees for a galaxy having $z \leq z_{\text{lim}}$, S , is then mapped into the probability of the classification at the leaf node via the sigmoid function:

$$P(z \leq z_{\text{lim}}) = \frac{1}{1 + \exp(-S)} \quad (3)$$

If the classification probability is higher than a threshold probability, then the galaxy is assigned class 0 i.e. $z \leq z_{\text{lim}}$. The default probability limit is 0.5, but in Sec. 6 we will also explore higher thresholds which reduce the number of galaxies assigned as targets. We will in particular find that more strict cuts are needed (even $P > 0.9$) if the size of the targeted sample in WW should not exceed the expected ~ 1 million.

4.3. Classification and metrics

The model's performance in classifying galaxies according to their redshifts is evaluated using a two-class accuracy metric.

However, since the survey specifically targets galaxies with redshifts less than z_{lim} (with the fiducial value $z_{\text{lim}} = 0.2$), we mainly utilize purity and completeness metrics for class 0, corresponding to galaxies with redshifts below this threshold, as well as their harmonic mean, the F1 score, to evaluate the performance of the classifier.

The two-class accuracy is defined as:

$$\text{Accuracy} = \frac{(0, 0) + (1, 1)}{(0, 0) + (0, 1) + (1, 0) + (1, 1)} \quad (4)$$

Here, (a, b) refers to ground truth a , for which class b is ascribed by the ML classifier. Purity (also known as *precision*), completeness (also *recall*), and F1 score for class 0 are respectively given as:

$$\text{Purity, } P = \frac{(0, 0)}{(0, 0) + (1, 0)} \quad (5)$$

$$\text{Completeness, } C = \frac{(0, 0)}{(0, 0) + (0, 1)} \quad (6)$$

$$\text{F1 score} = 2 \times \frac{\text{Purity} \times \text{Completeness}}{\text{Purity} + \text{Completeness}} \quad (7)$$

Our goal is to maximize these metrics, i.e. bring them as close to unity as possible. For the WAVES science goals, it is especially desirable to have Completeness above 0.95 (Driver et al. 2019), but for the efficiency of the survey, also Purity should be kept at a high level. In the WW case, the former is much more important, though, and when probabilities from various target allocation methods are combined, a cost function of C^4P is used (Tempel et al. in prep.).

4.4. Training with *specz* compilation

The XGB classification model is trained on the spectroscopically labeled dataset detailed in Sec. 2.2. As discussed in Sec. 2.1, our input photometric catalog has fluxes available in 9 bands (4 from KiDS and 5 from VIKING). For the ML model, the input photometric information is provided in terms of features. For that, we used 9-band magnitudes corresponding to 'color' fluxes as available in the database.

In Sec. 2.1, we also mentioned that for our analysis we only kept galaxies with the full 9-band information by eliminating objects with any missing bands. This is because, although the XGBoost classifier has an implicit way of dealing with missing values, this in-built approach is not the best solution for imputing the missing fluxes in our case. XGBoost assigns missing values to a default direction (left or right) in each decision tree split. The optimal direction is chosen by evaluating the loss function for each potential split. The reason that this sparsity-aware split-finding method is not suited for the missing flux values is that these values are not missing at random. What is more, in our work we also use all the 36 colors (magnitude differences) constructed from the 9 magnitudes. If we wanted to impute those where at least 1 band is missing, this would affect several of our features and could introduce biases. Therefore, in total, we have 45 input features and each has a numerical value associated.

To select the best model and test its performance, a random subset of the dataset with true labels is held out, which is referred to as the test set. In this work, we use randomly sampled 80% of the spec-z-cat as the training set and the rest as the test set. For

a given set of experiments when both the magnitude and redshift cut are at their fiducial WW values, the test set is fixed. We however draw it anew when we change any of these thresholds.

From the ML point of view, our labeled data (and hence a random test set extracted from them) are relatively balanced as the galaxies with spec- z label $z \leq 0.2$ (Class 0) make up its $\sim 45\%$. We would like to reiterate, however, that this does not reflect the actual expected proportion between the true classes in the final data, which should be closer to 15% at the fiducial flux limit (Sec. 3).

To prevent overfitting and ensure the robustness of the model, we employ the popular k -fold cross-validation approach. In this method, the training data is divided into k equally sized subsets (or "folds"). The model is trained k times, each time using a different subset as a validation set and the remaining $k - 1$ subsets as the training set. This process is repeated k times, and the performance metrics are averaged across the k iterations to obtain a more reliable estimate of the model's performance. For our implementation, we use $k = 5$, meaning the procedure is repeated 5 times with each fold being used exactly once as the validation data.

The full photometric catalog defined in Sec. 2.1 is the target set. For the final inference, to find the probability of the galaxies being desired targets from Class 0, i.e. having $z \leq 0.2$, we train the final XGB classifier on the entire spec- z -cat. Then, the trained classifier gives class labels and probabilities of classification for the target set galaxies.

5. Classification Performance

In this Section, we present the classification metrics for the test set. This will allow us to quantify the model performance on data with true labels.

5.1. Feature Importance

We have kept the feature engineering aspect of the model at a minimal level, as we found that the classifier works well with the combination of the 9 KiDS+VIKING magnitudes plus all the possible 36 colors as the input features, as mentioned in Sec. 4.4. Even if this means repetitive information (as most combinations are then non-unique), such redundancy is useful and often desired for ML, as the models, especially tree-based, usually are ignorant to the fact that some features are combinations of others. In fact, in addition to colors, i.e. magnitude differences, extra information could be provided, such as for instance magnitude ratios, as shown in e.g. Nakoneczny et al. (2019), Nakoneczny et al. (2021), even if there is no direct physical meaning of such features. In this work, we do not employ such feature combinations since the basic setup of magnitudes and colors works sufficiently well.

For better understanding which bands and colors matter the most for WW target selection, we perform feature importance analysis. We do it in two different ways. In the first approach, we rank the features based on their effect on the F1 score (Eq. 7) for the class of galaxies with redshift below 0.2 for the fiducial case. We run the classifier 30 times with shuffled values of one of the features and check the decrease in the F1 score. The relative mean of this decrease for each feature gives the importance of the particular feature for the classification which we express in per cent. Table 2 (middle column) shows that out of the total 45 input features, the $g - r$ and $u - g$ colors are the most important for our classification, followed by the g -band magnitude and some near-IR colors. The fact that the g band appears in the 3 top features

Table 2. Feature importance ranking for our XGBoost classification model from the permutation feature importance and the Shapley analysis. For the permutation feature importance, the scores are obtained from the relative decrease in the F1 score when the values of the feature are shuffled. Feature importance scores from the SHAP analysis are based on the mean of the absolute SHAP values for the features. The features are ordered in the importance ranking from the permutation feature importance. We show the 10 most important features from each method.

Feature name	Standard importance	SHAP score
$g - r$	7.26%	1.51
$u - g$	5.66%	0.94
mag_gc	2.01%	0.94
$J - K$	1.83%	1.59
$Z - Y$	0.83%	0.29
$H - K$	0.63%	0.56
$g - i$	0.47%	0.54
$r - i$	0.38%	0.20
$Y - K$	0.28%	0.27
$r - Y$	0.23%	0.14

most relevant for our classifier is likely related to the fact that the redshifted 4000 Å break passes through the VST g filter close to $z = 0.2$, which is the WW selection threshold.

In order to better understand the impact of each photometric feature on the XGB classification, we perform an additional feature importance analysis called SHAP (SHapley Additive exPlanations, Lundberg et al. 2020). This approach takes into account the contribution of every instance of the feature to the model's final prediction of the probability, $P(z \leq z_{\text{lim}})$ in our case. The Shapley value of a feature is the average of the individual contributions in different feature combinations, considering every possible order in which features could be added. Such analysis is useful because, unlike the permutation-based method above, SHAP takes into consideration the correlations between the features. The SHAP values for the ten most important features for classification are given in the right-hand-side column of Table 2. It is interesting to note that, although their order and relative importance differ from what was given in the middle column of Table 2, all but one of the most important features are the same between the two approaches. Also the top four, although shuffled, are the same, only that now the $J - K_s$ color is the most important in terms of its SHAP value. This color is sensitive to the shift of the 'stellar bump' which is likely the reason why it is at least as important as the features including the g -band.

Based on the above analysis, we could in principle trim the least important features and build a new model with fewer magnitudes/colors. This would however be relevant if it allowed us to reduce the number of required bands and/or it considerably reduced the computational time. However, as visible in Table 2, the top 10 features include combinations of all 9 KiDS+VIKING magnitudes. What is more, even with the full set of 45 features the XGB model runs very fast and computations can be done on a laptop. We therefore chose not to re-run the whole framework and keep all the magnitudes and colors throughout.

5.2. Results for the default WAVES-Wide thresholds

Here we will quantify the performance of our classifier for the default WAVES-Wide selection thresholds, which are $Z < 21.1$ mag and $z < 0.2$. The classification metrics for the test set, computed according to the definitions from Sec. 4.3 (Eqs. 4-7), are presented in Table 3. Here and below, the errors on the metrics

Table 3. Classification performance for the test set. The first three metrics are computed for class '0', i.e. WW targets, while the accuracy is a two-class metric including both positives and negatives (targets and non-targets). See Sec. 4.3 for metric definitions and Sec. 5.2 and Appendix A for details on error calculation.

Metric	Value
Purity (Eq. 5)	94.71% $\pm$ 0.11%
Completeness (Eq. 6)	94.68% $\pm$ 0.12%
F1 score (Eq. 7)	94.70% $\pm$ 0.07%
Accuracy (Eq. 4)	95.64% $\pm$ 0.06%

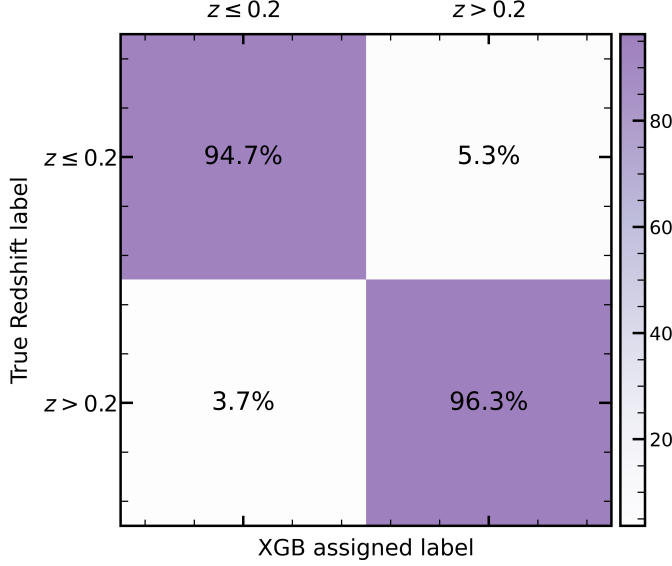


Fig. 6. Confusion matrix showing the percentages of Class '0' (Class '1') sources correctly and incorrectly labeled by the XGB along the top (bottom) row.

are calculated by propagating the Poissonian errors on the number of (0,0), (0,1), (1,0), and (0,1) objects, where 0 is the target class and 1 are non-targets, and first position in parentheses is the true class while the second is the class assigned by the classifier. The results of classification are also shown as a normalized confusion matrix in Fig. 6. The matrix is normalized along each row. We can see that, for instance, about 3.7 % of higher-redshift (class 1) galaxies are incorrectly labeled as $z \leq 0.2$, i.e. as WW targets.

It is important to note that the number of objects identified as a certain class varies with 5-fold training and testing. Hence, the errors in these numbers are correlated. We calculate this correlation via the Pearson coefficient from the 5-fold testing. For more details, see Appendix A. As can be seen, the one-class metrics (purity, completeness and F1) are consistently close to 95%, while the two-class accuracy is even higher. These high values indicate very good performance of our classifier for WW targets.

The values of the metrics discussed above were provided for a cut in classifier probability at 0.5, i.e. objects with $P_{\text{XGB}} > 0.5$ were allocated to class 0 and the others to class 1. This probability threshold could, however, be varied (in particular, increased) for more reliable classification. It is then informative to look at the distribution of probabilities assigned to the full test sample (with both the positive and negative true class, i.e. galaxies at any true redshift); this is presented in Fig. 7. A reliable classifier should have two sharp peaks at $P = 0$ and $P = 1$, which correspond to

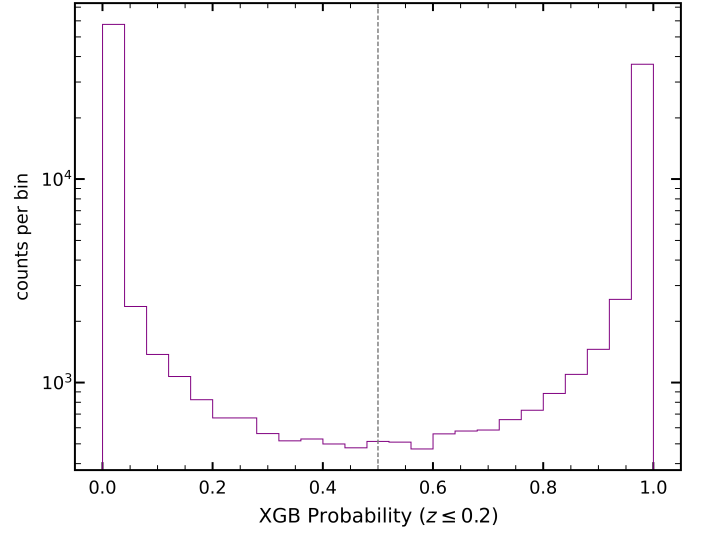


Fig. 7. Distribution of XGB-assigned probability that test sample sources are the desired WAVES-Wide targets, i.e. have $z \leq 0.2$. The default classification probability threshold, shown with the vertical line, is 0.5.

class 1 and 0, respectively. This is what we observe indeed; in addition, relatively few sources are assigned probabilities close to 0.5, which would indicate uncertain predictions. From the plot, we see that varying the probability threshold, based on which the targets are assigned, between even $P \sim 0.3 \div 0.7$ would hardly affect the classification.

To further verify the appropriate threshold for classification probability, in Fig. 8 we display the dependence of the performance metrics as a function of P_{XGB} . We observe that both accuracy and F1 are relatively stable for most of the P_{XGB} while completeness and purity change almost linearly except for a quick steepening at the extreme end of respectively $P \sim 1$ and $P \sim 0$. Overall, the four presented curves suggest that our originally adopted cut of $P = 0.5$ to differentiate between WW targets and non-targets is a good compromise. At the same time, Fig. 8 could be further used to decide how to move the probability threshold in case a higher value of completeness or purity is required. For instance, to obtain completeness of 99%, one would need to adopt a very permissive $P < 0.09$, while for purity at the same level, an aggressive cut of $P > 0.91$ would be necessary. We further discuss the probability cuts that could be taken in the final inference set in Sec. 6.

We further visualize the performance of our classifier by mapping the XGB-assigned probabilities on the redshift-color plane. For the latter, we choose $g - r$, which was identified as the most important feature for our classifier. This is illustrated in Fig. 9. Each dot in this plot is one galaxy from the test set, and they are color-coded by the assigned probability of being a target (dark being high and light being low P). The grey contours illustrate the distribution of the sources where their density is the largest. The bimodality in the contours is due to both red (larger $g - r$) and blue galaxies being present in our test set. The vertical dashed line indicates the $z = 0.2$ threshold for the WW targets. An ideal classifier would label all the galaxies to the left of this line as targets (giving them $P > 0.5$, i.e. dark color) while those to the right would all have low P and light color. Although our classifier is not perfect, we see that consistently with previously quoted values of completeness and purity metrics, a very small fraction of objects end up as false negatives (5.3%) or false positives (3.7%). In addition, these incorrect predictions

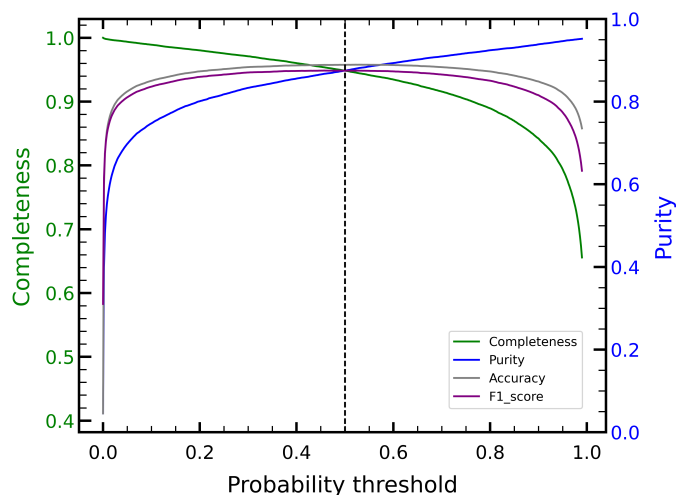


Fig. 8. Classification metrics as a function of the probability threshold of the XGB classifier to select WAVES-Wide targets.

are mostly located close to the threshold line. Some outliers exist, but they are not numerous. For instance, only about 100 galaxies at $z > 0.4$ from the test set have $P(z < 0.2) > 0.7$.

The final test we provide in this section is to quantify the performance of our classifier as a function of the (true) redshift, Z-band apparent magnitude, and the $g - r$ color, for the fiducial WW redshift and magnitude thresholds and using the $P = 0.5$ cut between the classes. For that, we partition the test sample into narrow bins of the respective quantities and compute the statistics per bin. Figures 10 and 11 illustrate respectively redshift and magnitude dependence of completeness (dark green points) and purity (olive points, only for magnitudes). Clear trends emerge: the XGB classification performs best at the bright, low-redshift end, deteriorating for galaxies close to the WW redshift threshold and magnitude limit. This fall-off is more abrupt for redshifts: up to $z \lesssim 0.15$, completeness stays at a level of $\sim 99\%$ and quickly plummets to $\sim 70\%$ at $z \sim 0.2$. As a function of magnitude, the deterioration is more gradual, although the completeness reaches the value of $\sim 70\%$ close to the magnitude limit of the sample. It is also worth noting that, as the median Z-band magnitude of the test set is 18.6 mag, the averaged metrics are biased towards the brighter end. The trends as a function of redshift and magnitude are likely driven by two aspects: one is the limitations of the training set, whose very complete flux-limited component from GAMA reaches to $z < 0.2$ at $r < 20$ mag (equivalent to roughly $z < 19.3$ mag). Secondly, our classifier seems to be most uncertain in its predictions close to the redshift threshold (cf. Fig. 9).

As far as the color dependence is concerned, no strong trends are visible, as illustrated in Fig. 12. For $g - r \lesssim 1.2$, both completeness and purity stay at similarly high levels above 90%. Only for the reddest galaxies of $g - r \sim 1.3$ mag, we observe considerable deterioration in the two metrics. This is likely driven by the fact that such reddened galaxies are mostly above the WW threshold redshift of $z = 0.2$ (cf. Fig. 9), there is then a large imbalance between the true and false class in the training for this color range. However, as very few galaxies at $z < 0.2$ have $g - r > 1.3$, the classification results could be improved by simply removing a posteriori all such red objects from the sample.

Another factor that will contribute to the classifier's performance will be the signal-to-noise ratio (S/N) of the photometry. We have checked the trends in that respect and found that, as

expected, both completeness and purity deteriorate with lowered S/N. For instance, in the g -band they go down to levels of $\sim 85\%$ at S/N of roughly 50, and worsen further with lowered S/N. As magnitudes (fluxes) and their errors are highly correlated, these trends are qualitatively very similar to what is observed as a function of magnitude itself.

We have also looked into any possible trends in magnitude and color distributions for the misclassified sources (i.e., those missed out – false negatives, FN, and contaminants – false positives, FP). While such objects are typically fainter and bluer than the correctly assigned WW targets (true positives), we do not see significant differences between FP and FN that would point to a particular sensitivity of our classifier.

5.3. Going beyond the default WAVES-Wide selections

As we specified previously, the currently planned WAVES-Wide selections are $Z < 21.1$ mag and $z < 0.2$. However, these thresholds could still be revised by the survey depending on the efficiency and time availability of the observations. It is then interesting to explore how our approach will perform if we vary one or both of the limits, still keeping them close to the fiducial values. In particular, we would like to find out if for our binary classifier there exists a more optimal combination of the redshift-magnitude thresholds which would still provide a similar number of WW targets as for the default cuts. For this purpose, we ran a set of new XGB classifications where each time the samples with true labels (train, validation, and test) were changed after imposing new Z-band magnitude and/or redshift thresholds.

We start by varying the flux limit in the Z-band for selection, while the redshift cut is the fiducial one i.e. $z < 0.2$. For that purpose, from the spec-z-cat described in Sec. 2.2 we select subsamples at different Z-mag limits ranging from 20.0 to 21.8 in steps of 0.3 mag (i.e. 7 possibilities). The magnitude range we chose here starts from roughly 1 mag brighter than the fiducial cut, while the upper limit is driven by the input catalog going only to $Z = 22$ mag². We then train and test separate XGB classifiers for these flux-limited catalogs. The results are illustrated in Fig. 13 for the statistics introduced in Sec. 4.3. In the plot, each datapoint with error bars is derived from the training and testing at a particular magnitude limit. We notice that all the 1-class statistics consistently deteriorate when the flux limit is decreased, although not by a large margin: from roughly 0.95 at $Z < 20$ limit to about 0.945 at $Z < 21.8$ mag. This deterioration is likely driven by two factors. First, as illustrated earlier in Fig. 2, increasing the magnitude threshold for the targets increases the mean redshift of the entire input sample to select from, and makes the percentage of the required $z < 0.2$ targets smaller. This means that it becomes more difficult to properly select the targets as the flux limit is lowered. Second, our spectroscopic training set is most complete at the bright end (cf. Fig. 4), therefore, a supervised classifier like ours will perform worse the more faint objects we add.

We now turn to the second set of experiments, where we fix the Z-mag cut at the fiducial value of 21.1 but move the redshift threshold for the targets. This means that we now have one selection of galaxies with spec-z labels for training and tests – the same as in the fiducial case – but we train 4 additional classifiers in addition to the one analyzed in the previous section, namely, we look at target selection for cases with $z_{\text{lim}} = 0.18, 0.19, 0.21, 0.22$. As previously for varied magnitude limits, also here we compute

² The KiDS+VIKING source data reach fainter than this, but for practical reasons we work with a catalog limited to $Z = 22$ mag.

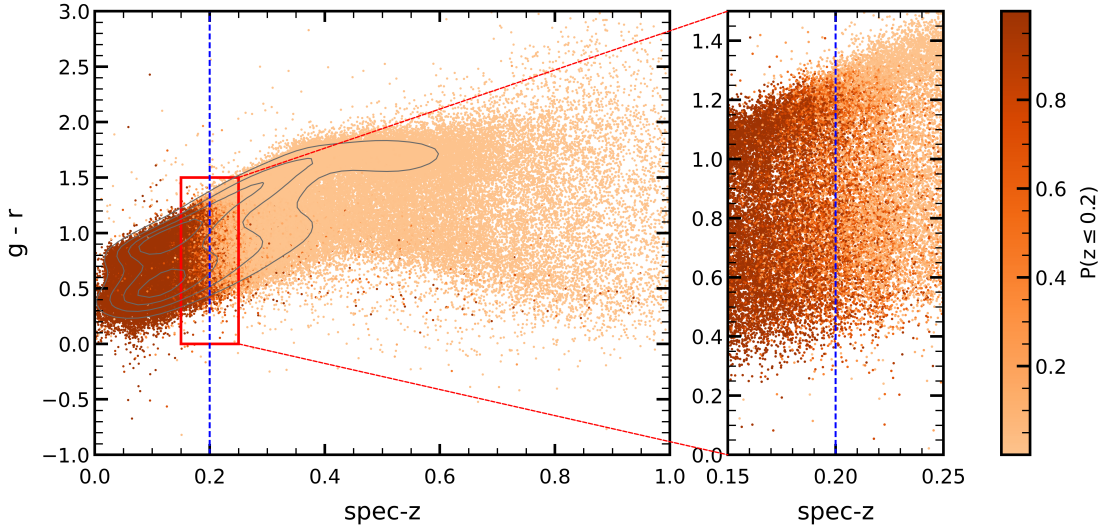


Fig. 9. Distribution of the galaxies from the test set on the true redshift – observed $g - r$ color plane, color coded by the probability of being a WAVES-Wide target as assigned by our classifier. Dark indicates a high probability of having $z < 0.2$, while light means assignment as non-target. The vertical dashed line shows the WW division threshold. Grey contours indicate the region of the largest density of galaxies in the diagram. On the right-hand side, we enlarge the area close to the threshold to highlight the region where most of the misclassification happens.

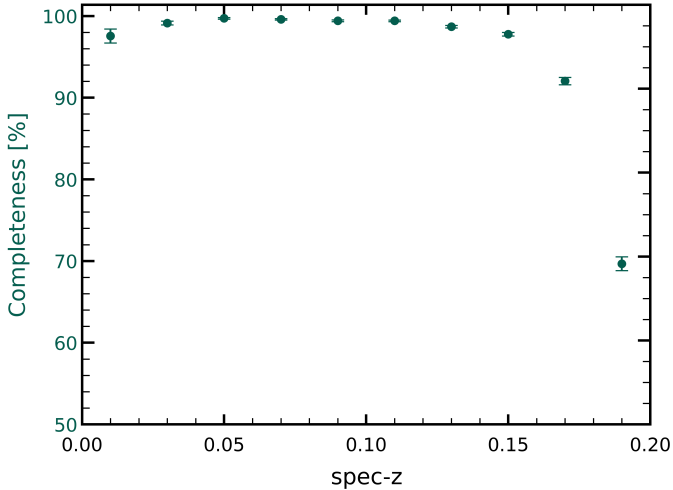


Fig. 10. Completeness as a function of true redshift for the desired target (Class '0') galaxies, computed per redshift bin of width 0.02. The scatter on the metric is from the Poissonian error on the number of correctly identified (0,0) and missed (0,1) galaxies.

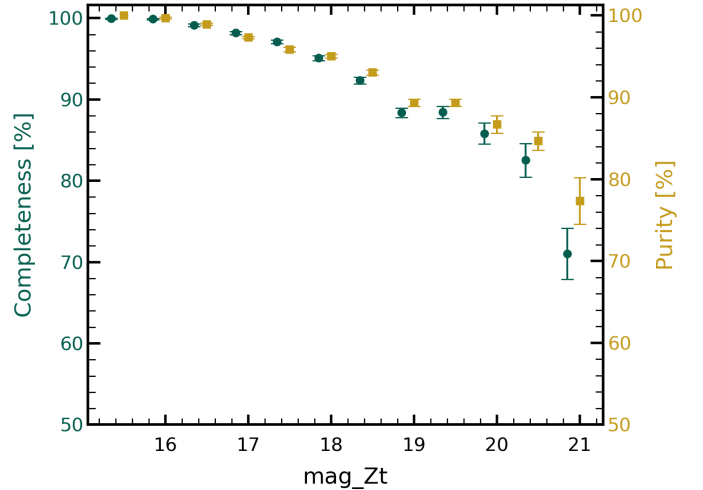


Fig. 11. Completeness (dark green) and purity (olive) as a function of Z-band magnitude for sources in the test set, computed per magnitude bin of width 0.5. For better visualization, olive dots representing purity are shifted by 0.15 along the x-axis. The scatter in the metrics arises from Poissonian errors associated with the counts of correctly identified (0,0), contaminant (1,0), and missed (0,1) galaxies.

the same metrics and show them in Fig. 14, where this time the changing redshift threshold for WW targets is shown on the x-axis. We find that, for the fiducial magnitude cut, increasing the redshift limit between $z = 0.18$ and $z = 0.22$ improves all the analyzed statistics, although again the differences are not substantial, roughly between 0.945 and 0.955. A likely explanation for this improvement is that, as can be appreciated from the violet line in Fig. 4, the median redshift of galaxies flux-limited to $Z = 21.1$ is about 0.43. Therefore, increasing the redshift selection threshold in our analyzed range reduces the imbalance between targets and non-targets, which improves the performance of the binary classifier. The closer we are to the median redshift with the threshold, the easier it should be for the classifier.

As the final test, we vary jointly the magnitude and redshift threshold for the WW target selection. Here the number of possibilities is very large if we keep the same ranges and steps as

in the two previous paragraphs; therefore we discuss only a sub-selection for 3 different redshift limits ($z < 0.18$, $z < 0.20$, $z < 0.22$) and 4 magnitude thresholds ($Z \leq 20.0$, $Z \leq 20.6$, $Z \leq 21.2$, $Z \leq 21.8$), which gives 12 combinations in total. We report the metrics for the XGB classifiers in Table 4, in which the first column is the redshift limit for the target selection, column two provides particular Z-mag cuts, and the remaining columns include the metrics computed for the particular redshift-magnitude threshold combination. The results confirm our earlier findings: the metrics are maximized if the redshift limit is increased and the magnitude limit decreased. In other words, our supervised classifier would work best, for the currently available training data, if WW target selection went deeper in terms of redshift, but shallower in terms of flux, than the fiducial adopted values.

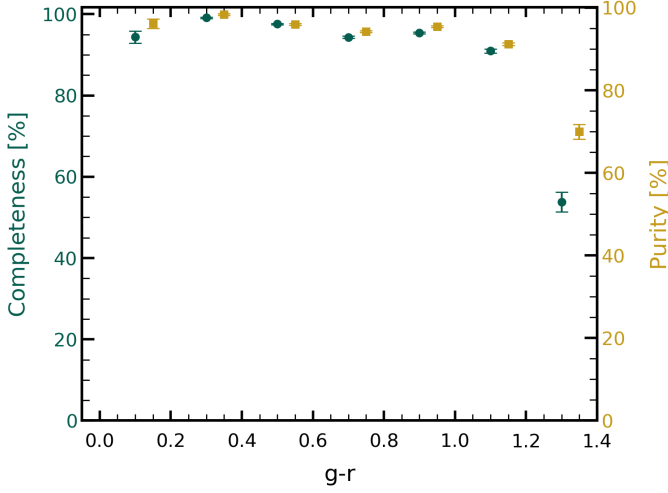


Fig. 12. Completeness (dark green) and purity (olive) as a function of the $g-r$ color for sources in the test set, computed per bin width of 0.2. For better visualization, olive dots representing purity are shifted by 0.05 along the x-axis. The scatter in the metrics arises from Poissonian errors associated with the counts of correctly identified (0,0), contaminant (1,0), and missed (0,1) galaxies.

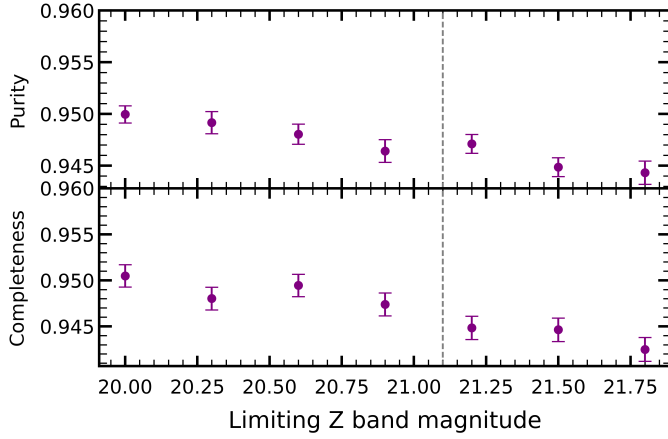


Fig. 13. Performance of XGB-based selection of WAVES-Wide targets for different thresholds of the Z-mag cut (x-axis) with the redshift limit fixed at the fiducial $z < 0.2$ value. The scatter on the metrics is from the Poissonian error on the number of True Positives, False Positives, and False Negatives. The vertical dashed line shows the WW fiducial Z-band cut.

We note, however, that the differences in the metrics between the most optimal and the fiducial case are not very big, roughly at the 1% level, even if they are statistically significant, taking into account the typical scatter of $\sim 0.1\%$.

6. Selection of WAVES-Wide targets from the entire input catalog

We will now employ our approach to select WAVES-Wide targets from the entire input photometric catalog. For that purpose, we train the XGB classifier on the full spec-z set of 572,325 galaxies (at the fiducial $Z < 21.1$ mag limit). As previously, our true class ‘0’ are galaxies at $z < 0.2$. For the probability threshold of $P > 0.5$ our model allocates above 2.6 million target candidates out of over 12 million of all objects with 9-band detections and pre-assigned to be galaxies according to the Cook

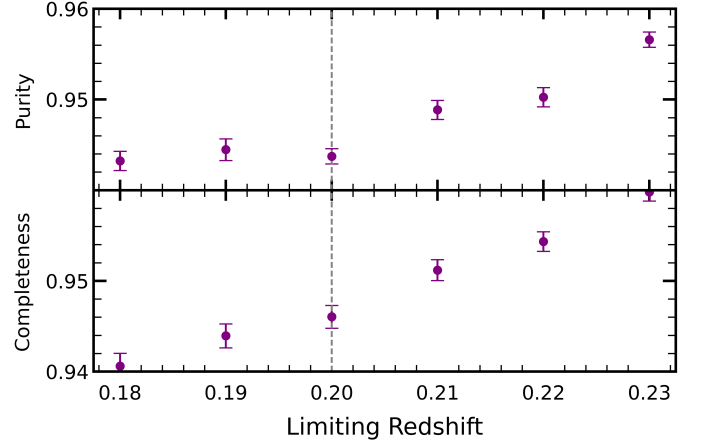


Fig. 14. Performance of XGB-based selection of WAVES-Wide targets for different thresholds of the redshift limit (x-axis) with the Z-mag cut fixed at the fiducial $Z = 21.1$ mag value. The scatter on the metrics is from the Poissonian error on the number of True Positives, False Positives, and False Negatives. The vertical dashed line shows the WW fiducial redshift cut.

Table 4. Performance of XGB-based selection of WAVES-Wide targets for different combinations of the redshift limit for the selection (first column) and Z-band magnitude limit (second column). Columns 3-5 present, respectively, purity, completeness and F1 score for the given combination of cuts. We highlight the instances when the given metric is maximized on this particular grid. The scatter of all the metrics takes the value of about 0.1%, so we do not show it here.

Redshift cut	Z-mag cut	Purity	Completeness	F1 Score
0.18	20.0	94.5 %	94.5 %	94.5 %
	20.6	94.4 %	94.5 %	94.5 %
	21.2	94.1 %	94.2 %	94.2 %
	21.8	94.0 %	93.7 %	93.9 %
0.20	20.0	94.9 %	94.9 %	94.9 %
	20.6	94.8 %	94.8 %	94.8 %
	21.2	94.7 %	94.8 %	94.8 %
	21.8	94.6 %	94.1 %	94.8 %
0.22	20.0	95.4 %	95.7 %	95.6 %
	20.6	95.1 %	95.8 %	95.4 %
	21.2	95.1 %	95.6 %	95.3 %
	21.8	94.8 %	95.1 %	95.0 %

et al. (2024) procedure. This gives about 20.7% of input galaxies being allocated as targets with our approach, as compared to 12.1% predicted from the SHARK mocks (Sec. 3). The considerably larger percentage obtained in the real catalog can mean some degree of contamination by galaxies at higher redshifts than required. This could be reduced by increasing the probability cut of the classifier. It is also possible that the mocks underestimate the fraction of the $z < 0.2$ galaxies at our fiducial flux cut.

For the resulting catalog of WW target candidates, we cannot compute such performance statistics as in the previous sections, as these require true labels (redshifts) to be known. We can, however, still analyze how the classifier works by looking into the properties of the resulting catalog. The first relevant quantity to check is the probability of being a target as assigned by XGB. This is illustrated in Fig. 15, which shows the histogram of the XGB probabilities for the entire photometric catalog in the same way as was done previously in Fig. 7 for the spectroscopic test set. The allocated targets are to the right of the vertical $P = 0.5$

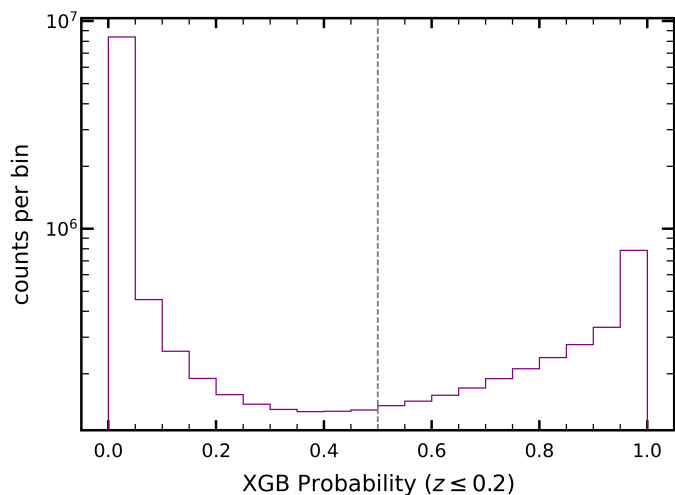


Fig. 15. Distribution of XGB-assigned probability of inference set galaxies being the desired WAVES-Wide targets, i.e. having $z \leq 0.2$. The default classification probability threshold, shown with the vertical line, is 0.5.

line. The first observation to make is the very pronounced peak at $P \sim 0$ and the generally larger ‘confidence’ of the classifier to assign non-targets than targets, visible in the minimum of the probability distribution shifted to the left with respect to the middle 0.5 value. Also, the slope of the histogram at the high-probability end (for the possible targets) is less steep than for non-targets, although there still is a clear peak at $P \sim 1$. The fact that the shape of the probability histogram is here quite different from that for the test set (Fig. 7) may indicate that the classifier is less ‘certain’ for a number of galaxies in terms how they should be allocated and in particular, the probability cut at 0.5 may not be optimal in the final photometric set.

As discussed above, the high percentage of targets allocated by our classifier among all the input photometric galaxies may mean some degree of contamination by objects with true redshifts $z > 0.2$. A way to reduce these, and generally decrease the number of possible targets – which might be preferable from the point of view of time allocation on the 4MOST instrument – is by raising the probability cut over the fiducial $P > 0.5$. In Fig. 16 we show how the number of the targets would change when sliding through the P threshold all the way from 0 to 1. Obviously, at the extremes, we have either the entire photometric dataset assigned as ‘targets’ ($P_{\text{cut}} = 0$) or none ($P_{\text{cut}} = 1$). We also see that the target number changes very rapidly close to these extremes, while it varies much more slowly over the range of $0.1 \lesssim P_{\text{cut}} \lesssim 0.7$. This diagram can also be used to decide what probability cut to take in order to obtain a fixed number of target candidates. As already mentioned, for the fiducial cut of $P > 0.5$, we have almost 2.6 million targets; this would be reduced to 2 million if we take $P > 0.71$ or further to 1 million for $P > 0.91$.

In Figure 17, we take a closer look at how our classifier maps the probabilities onto the color space. For this, we show the observed $u - g$ vs. $g - r$ color-color planes for the test set (left panel) and the entire photometric inference set (right panel). Pixels are color-coded by the median XGB probability of having $z < 0.2$. We notice that the classifier seems more confident for the test set, where there is a clearer separation between dark (targets) and light (non-targets) pixels; this is more diluted for the entire inference sample. Secondly, in the photometric dataset from which the observational targets will be chosen, there is a considerable number of galaxies having colors not well-represented in the test

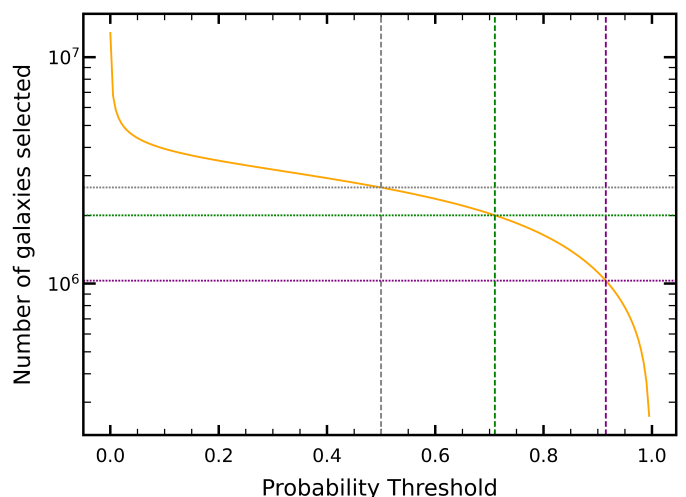


Fig. 16. The number of galaxies selected as targets for WAVES-wide as a function of the $P(z \leq 0.2)$ classification threshold. The grey lines show the number of the targets selected when the classification probability threshold is the default one i.e. $P > 0.5$. Green and purple lines show the probability criteria needed to select 2 million and 1 million targets (respectively $P > 0.71$ and $P > 0.91$).

set, the latter being statistically consistent with the training. The same will be the case in the general feature space, i.e. when other colors are also taken into account. This is an expected outcome of the fact that our spectroscopic training set is complete (in terms of flux-limited selection) only within the GAMA limit of $r < 20$ mag.

As we already discussed in Sec. 2.2, at the faint end of the WAVES selection we only have color- or galaxy-type-selected datasets with true redshift labels for the training. This complication will affect any empirical methodology that aims to map the color space to redshift for the purpose of WW target assignment. In our case, we can, however, conclude that our classifier assigns relatively low probabilities ($P \lesssim 0.6$) of being a target to the objects whose colors are not well represented in the current spectroscopic training. A cut on probability above the fiducial $P = 0.5$ threshold will mitigate the contamination by false positives (galaxies chosen as targets which in fact have $z > 0.2$). For instance, cutting at $P > 0.9$ removes all the false positives with true redshifts $z > 0.35$ in the test set. While we cannot check this directly in the full photometric sample as we do not know the true redshifts, we can safely assume that similar improvement will be observed there.

An additional effect visible in Fig. 17 is that the photo-cat, from which targets are selected, has a much broader color coverage than the spec-z-cat used for training. This is expected first due to the much larger size of the former dataset, but also because of specific selections of spectroscopic surveys that are included in our calibration sample. The large extent of photometric data in the color space can be also partly due to noisy photometry at the faint end, especially for the shallowest u -band. In the particular color-color plane that we present in Fig. 17, the data well outside of the scope of the spec-z-cat do not seem to influence the performance of the classifier: objects with high probability of being a target lie in an area where galaxies indeed have $z < 0.2$ (cf. Fig. 1). However, the situation might be different if other colors are inspected. Due to high dimensionality of the problem, a quantitative comparison of training and inference set coverage would require using an appropriate tool, such as self-organizing maps (e.g. [Jalan et al. 2024](#)). This is beyond the scope of our work,

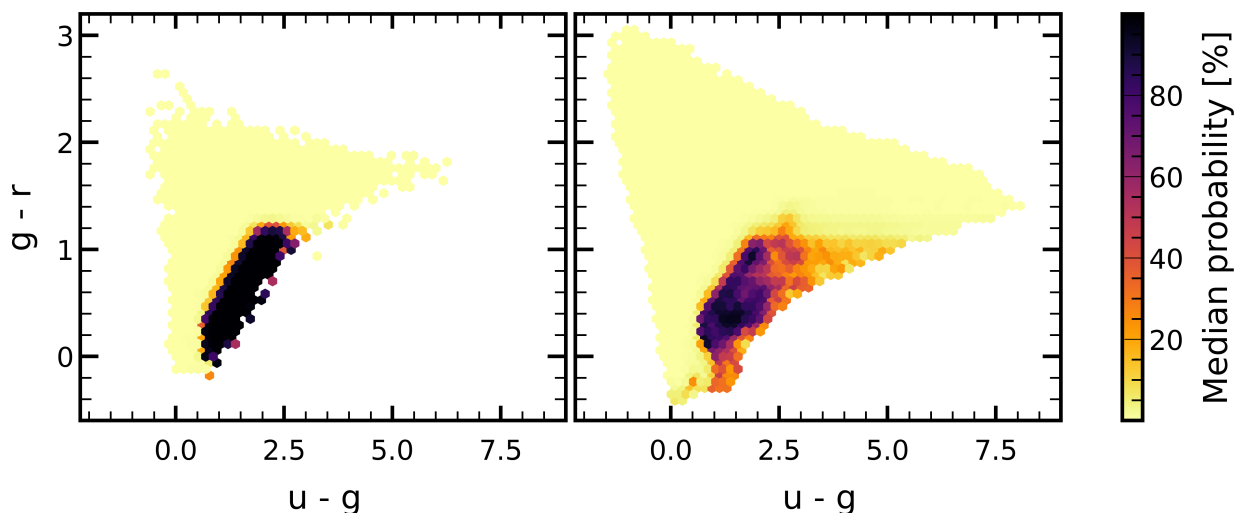


Fig. 17. (left panel) Distribution of test set galaxies on the observed $u - g$ vs. $g - r$ color-color plane. Each pixel is color-coded with the median probability of galaxies being assigned as targets by XGBoost, i.e. having $z < 0.2$. (right panel) Similar, but for the entire photometric sample at the WAVES-Wide fiducial flux limit of $Z < 21.1$ mag. For the test set, we only show pixels with at least 5 galaxies each, while for the inference sample, this threshold is 100. Both panels share the same y-axis range for the $g - r$ color.

as the eventual WAVES target selection combines our approach with others developed by the team (Tempel et al. in prep.). We note, however, that independently of the method, color cuts on the inference catalog could be applied a priori to remove highly unlikely targets. This would reduce the size of the photometric catalog as well as possibly the number of mis-classified targets.

7. Summary and conclusions

In this paper, we presented a supervised machine-learning (ML) approach to select observational targets for the forthcoming WAVES-Wide galaxy survey on 4MOST. The targets will be both flux- and redshift-limited, while their redshifts are not known a priori. In our method, rather than estimating individual redshifts from multi-band photometry, we performed binary classification to decide if a galaxy should be a target or not, based on object locations in magnitude-color space. In particular, we used the fast and efficient XGBoost classifier, trained on a labeled sample derived from a compilation of relevant spectroscopic surveys overlapping with WAVES photometry, from which we use 9 broad-band magnitudes and 36 associated colors. XGB, an ensemble tree-based ML model, is not very demanding computationally and also provides output probabilities of objects belonging to a given class, rather than just discrete labels.

Using the labeled sample separated into train, validation and test sets, we quantified the performance of our XGB classifier. This was done for the default WW limits of $Z < 21.1$ mag & $z < 0.2$, but also when one or both of these thresholds were varied. In all cases the classification metrics such as completeness, purity or accuracy were at the level of $\sim 95\%$. Small improvement in such statistics is visible if the magnitude limit of the selection is lowered and/or the redshift cut is increased, but the differences are within 1%. We also examined the probabilities of objects belonging to the desired class (i.e. being a target) that XGB provides. For the test set we found that the default limit of $P > 0.5$ is optimal as far as the various metrics are concerned.

As with any supervised ML model, our classification method has its limitations related to the composition of the training set. Here in particular, our approach requires spectroscopic redshifts for the class labels in the calibration set, overlapping on the

sky with input photometry for the color-based features we use. At present, within the general WAVES area, surveys with true redshifts do not provide complete coverage of the color-redshift space to fully represent WW targets on sufficiently wide areas. As a result, our training set is biased towards brighter, redder and lower-redshift galaxies as compared to a general galaxy sample magnitude-limited to $Z < 21.1$, from which targets will be selected. This leads to deterioration of the classifier's performance close to the flux and redshift cuts (i.e. at $z \gtrsim 0.18$ and $Z \gtrsim 19$). On the other hand, we do not see strong trends in classification metrics as a function of color, except for deterioration at the reddest end, where most of the galaxies are redshifted beyond the WW limit.

Biases in the training set may also affect classification results once the trained model is applied to the overall inference sample. Predictions need to be extrapolated in the regions of the color space that were not well covered by training. In our case, this extrapolation seems to be leading to a considerably different distribution of probabilities of being a target when evaluated on the entire output catalog as compared to the test set. Namely, many of the target candidates have XGB probabilities above the fiducial threshold of 0.5 but well below values indicating high confidence of the classifier, such as 0.9 or so. This, however, can be used to our advantage, and by adopting a higher probability threshold, we can remove the less certain target candidates. Indeed, direct application of the trained classifier to the entire WW photometric catalog gives us almost 2.6 million objects with a target classification probability over 0.5. This is much larger than previous estimates of ~ 0.9 million (Driver et al. 2019), but can be however reduced to any desired figure by increasing the probability cut. For instance, to obtain 1 million target candidates with our approach, a cut of $P > 0.91$ can be applied.

The results of our XGBoost classification are used together with other approaches for WAVES-Wide target selection to prepare the final list of galaxies to be observed by the 4MOST instrument. For that purpose, weights are computed using a cost function of the form C^4P , where C is completeness and P purity, the latter quantities being obtained from the particular techniques (GPz, SFM, TOPz and XGB). These weights are then used to calculate the weighted probability of an object to be a WW target,

employing the p_{target} delivered by the different approaches. As a side note, all four techniques provide similar overall completeness values, while ours presented here gives the highest purity. More details on the particular approaches and how they are combined will be documented elsewhere (Tempel et al. in prep).

The performance of our supervised classification approach could be further improved if the training set with true redshift labels is extended by adding more of fainter galaxies at higher redshifts, to better map the color-redshift relation. Some such galaxies are available from DESI DR1 (DESI Collaboration et al. 2025) which covers in particular the WW northern (equatorial) field. In the long term, WAVES itself, and 4MOST more generally, will deliver spectroscopic redshifts for such objects. However, as the target selection for 4MOST observations is already fixed, these new redshifts will not serve to revise the target selection but rather to verify a posteriori the proposed approaches such as ours.

Acknowledgements. We would like to thank Matthijs Mars and Julia Krajecka for their contributions at the early stages of this project. GK, MB & WAH are supported by the Polish National Science Center through grant no. 2020/39/B/ST9/03494. SB acknowledges funding by the Australian Research Council (ARC) Laureate Fellowship scheme (FL220100191). ET acknowledges funding from the HTM (grant TK202), ETAg (grant PRG1006, PRG3034) and the EU Horizon Europe (EXCOSM, grant No. 101159513). WAVES is a joint European-Australian project based around a spectroscopic campaign using the 4-metre Multi-Object Spectroscopic Telescope. The WAVES input catalogue is based on data taken from the European Southern Observatory's VST and VISTA telescopes (ID 179.A-2004 and ID 177.A-3016). Complementary imaging of the WAVES regions is being obtained by a number of independent survey programmes, including GALEX MIS, VST, WISE, Herschel-ATLAS, and ASKAP, providing UV to radio coverage. WAVES is funded by the ARC (Australia) and the participating institutions. The WAVES website is <https://wavesurvey.org>.

References

- Abolfathi B., et al., 2018, *ApJS*, **235**, 42
- Baldry I. K., et al., 2010, *MNRAS*, **404**, 86
- Baldry I. K., Sullivan T., Rani R., Turner S., 2021, *MNRAS*, **500**, 1557
- Balestra I., et al., 2010, *A&A*, **512**, A12
- Bellstedt S., et al., 2020, *MNRAS*, **496**, 3235
- Bilicki M., et al., 2021, *A&A*, **653**, A82
- Blake C., et al., 2016, *MNRAS*, **462**, 4240
- Breiman L., 2001, *Machine learning*, 45, 5
- Breiman L., Friedman J., Stone C., Olshen R., 1984, *Classification and Regression Trees*. Taylor & Francis, <https://books.google.pl/books?id=JwQx-W0mSyQC>
- Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, **432**, 1483
- Chauhan G., Lagos C. d. P., Obreschkow D., Power C., Oman K., Elahi P. J., 2019, *MNRAS*, **488**, 5898
- Chen T., Guestrin C., 2016, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery, New York, NY, USA, p. 785–794, doi:10.1145/2939672.2939785, <https://doi.org/10.1145/2939672.2939785>
- Colless M., et al., 2001, *MNRAS*, **328**, 1039
- Cook T. L., et al., 2024, *MNRAS*, **535**, 2129
- Cooper M. C., et al., 2012, *MNRAS*, **419**, 3018
- DESI Collaboration et al., 2016, *arXiv e-prints*, p. arXiv:1611.00036
- DESI Collaboration et al., 2024, *AJ*, **168**, 58
- DESI Collaboration et al., 2025, *arXiv e-prints*, p. arXiv:2503.14745
- Damjanov I., Zahid H. J., Geller M. J., Fabricant D. G., Hwang H. S., 2018, *ApJS*, **234**, 21
- Davies L. J. M., et al., 2015, *MNRAS*, **447**, 1014
- Davies L. J. M., et al., 2018, *MNRAS*, **480**, 768
- Drinkwater M. J., et al., 2010, *MNRAS*, **401**, 1429
- Driver S. P., et al., 2019, *The Messenger*, **175**, 46
- Driver S. P., et al., 2022, *MNRAS*, **513**, 439
- Duncan K. J., 2022, *MNRAS*, **512**, 3662
- Duncan K., et al., 2023, *The Messenger*, **190**, 25
- Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, *The Messenger*, **154**, 32
- Eisenstein D. J., et al., 2011, *AJ*, **142**, 72
- Elahi P. J., Welker C., Power C., Lagos C. d. P., Robotham A. S. G., Cañas R., Poulton R., 2018, *MNRAS*, **475**, 5338
- Fix E., Hodges J., 1951, *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, <https://books.google.pl/books?id=4XwyTAECAAJ>
- Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, *ApJ*, **715**, 823
- Gruen D., et al., 2023, *The Messenger*, **190**, 28
- Guzzo L., et al., 2014, *A&A*, **566**, A108
- Haines C., et al., 2023, *The Messenger*, **190**, 31
- Hasinger G., et al., 2018, *The Astrophysical Journal*, **858**, 77
- Ivezić Ž., et al., 2019, *ApJ*, **873**, 111
- Jalan P., et al., 2024, *A&A*, **692**, A177
- Jarvis M. J., et al., 2013, *MNRAS*, **428**, 1281
- John William A., Jalan P., Bilicki M., Hellwing W. A., Thuruthipilly H., Nakoneczny S. J., 2025, *A&A*, **698**, A276
- Lagos C. d. P., Tobar R. J., Robotham A. S. G., Obreschkow D., Mitchell P. D., Power C., Elahi P. J., 2018, *MNRAS*, **481**, 3573
- Lagos C. d. P., et al., 2019, *MNRAS*, **489**, 4196
- Lagos C. d. P., da Cunha E., Robotham A. S. G., Obreschkow D., Valentino F., Fujimoto S., Magdis G. E., Tobar R., 2020, *MNRAS*, **499**, 1948
- Laur J., et al., 2022, *A&A*, **668**, A8
- Le Fèvre O., et al., 2005, *A&A*, **439**, 845
- Le Fèvre O., et al., 2013, *A&A*, **559**, A14
- Lidman C., et al., 2020, *MNRAS*, **496**, 19
- Liske J., et al., 2015, *MNRAS*, **452**, 2087
- Lundberg S. M., et al., 2020, *Nature Machine Intelligence*, **2**, 56–67
- Masters D., et al., 2015, *ApJ*, **813**, 53
- Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, *ApJ*, **841**, 111
- Masters D. C., et al., 2019, *ApJ*, **877**, 81
- McCracken H. J., et al., 2012, *A&A*, **544**, A156
- McCullough J., et al., 2024, *MNRAS*, **531**, 2582
- Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, *A&A*, **624**, A13
- Nakoneczny S. J., et al., 2021, *A&A*, **649**, A81
- Newman J. A., Gruen D., 2022, *ARA&A*, **60**, 363
- Payerne C., et al., 2025, *J. Cosmology Astropart. Phys.*, **2025**, 031
- Pedregosa F., et al., 2011, *Journal of Machine Learning Research*, **12**, 2825
- Popesso P., et al., 2009, *A&A*, **494**, 443
- Porredon A., et al., 2021, *Phys. Rev. D*, **103**, 043503
- Prakash A., et al., 2016, *ApJS*, **224**, 34
- Robotham A. S. G., et al., 2011, *MNRAS*, **416**, 2640
- Robotham A. S. G., Bellstedt S., Lagos C. d. P., Thorne J. E., Davies L. J., Driver S. P., Bravo M., 2020, *MNRAS*, **495**, 905
- Scodreggio M., et al., 2018, *A&A*, **609**, A84
- Signor T., et al., 2024, *A&A*, **685**, A127
- Tempel E., et al., 2020, *A&A*, **635**, A101
- Trayford J. W., Lagos C. d. P., Robotham A. S. G., Obreschkow D., 2020, *MNRAS*, **491**, 3937
- Treyer M., Ait Ouahmed R., Pasquet J., Arnouts S., Bertin E., Fouchez D., 2024, *MNRAS*, **527**, 651
- Vaccari M., et al., 2016, in *The 4th Annual Conference on High Energy Astrophysics in Southern Africa (HEASA 2016)*. p. 26 (arXiv:1704.01495), doi:10.22323/1.275.0026
- Wright A. H., et al., 2024, *A&A*, **686**, A170
- Wu J. F., et al., 2022, *ApJ*, **927**, 121
- Zhou R., et al., 2021, *MNRAS*, **501**, 3309
- de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, *Experimental Astronomy*, **35**, 25
- de Jong R. S., et al., 2019, *The Messenger*, **175**, 3
- van der Wel A., et al., 2016, *ApJS*, **223**, 29

Appendix A: Errors on classification metrics

Here we describe our method for estimating the errors on the classification metrics, based on 5-fold cross-validation. Here we will discuss purity as an example.

As defined in Equation 5, purity is computed using the number of true positives (0,0) and false positives (1,0). Given that we employ 5-fold cross-validation, as described in Section 4.4, applying these models to the test set produces an array of five values, denoted as TP_i and FP_i , where i ranges from 1 to 5.

The classification performance metrics introduced in Section 4.3 (Eqs. 4-7) are derived from the means of the true positives (0,0), false positives (1,0), false negatives (0,1), and true negatives (1,1) obtained from the test performance of the 5-fold models. We define these means as $\overline{TP}$, $\overline{FP}$, $\overline{FN}$, and $\overline{TN}$.

Consequently, the purity equation (Eq. 5) can be rewritten as:

$$\text{Purity, } P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \quad (\text{A.1})$$

The Poissonian errors associated with (0,0), (1,0), (0,1), and (1,1) are given by $\sqrt{\overline{TP}}$, $\sqrt{\overline{FP}}$, $\sqrt{\overline{FN}}$, and $\sqrt{\overline{TN}}$, respectively. The corresponding uncertainty in purity using standard error propagation is then computed as:

$$\sigma_P = \left[\left(\frac{dP}{d(\overline{TP})} \sigma_{TP} \right)^2 + \left(\frac{dP}{d(\overline{FP})} \sigma_{FP} \right)^2 + 2 \left(\frac{dP}{d(\overline{TP})} \right) \left(\frac{dP}{d(\overline{FP})} \right) \rho_{TP,FP} \right]^{1/2} \quad (\text{A.2})$$

Here, $\rho_{TP,FP}$ represents the Pearson correlation between True Positive values [$TP/(0,0)$] and False Positives [$FP/(1,0)$], which is computed as:

$$\rho_{TP,FP} = \frac{\sum_{i=1}^5 (FP_i - \overline{FP})(TP_i - \overline{TP})}{\sqrt{\sum (FP_i - \overline{FP})^2} \sqrt{\sum (TP_i - \overline{TP})^2}} \quad (\text{A.3})$$

The partial derivatives $\frac{dP}{d(\overline{TP})}$ and $\frac{dP}{d(\overline{FP})}$ are given by:

$$\frac{dP}{d(\overline{TP})} = \frac{\overline{FP}}{(\overline{TP} + \overline{FP})^2} \quad (\text{A.4})$$

$$\frac{dP}{d(\overline{FP})} = -\frac{\overline{TP}}{(\overline{TP} + \overline{FP})^2} \quad (\text{A.5})$$

Finally, the scatter in the purity metric, as presented in Table 3, is determined using Equations A.2-A.5. Similarly, the uncertainties for Completeness, F1 score, and Accuracy are worked out following the same approach.