Astronomy & Astrophysics manuscript no. euclid_photo_selection
March 14, 2024
©ESO 2024

arXiv:2403.08726v1 [astro-ph.CO] 13 Mar 2024

Euclid Testing photometric selection of emission-line galaxy targets⋆

M. S. Cagliari⋆⋆[1], B. R. Granett[2], L. Guzzo[1,2,3], M. Bethermin[4,5], M. Bolzonella[6], S. de la Torre[5], P. Monaco[7,8,9,10], M. Moresco[11,6], W. J. Percival[12,13,14], C. Scarlata[15], Y. Wang[16], M. Ezziati[5], O. Ilbert[5], V. Le Brun[5], A. Amara[17], S. Andreon[2], N. Auricchio[6], M. Baldi[18,6,19], S. Bardelli[6], R. Bender[20,21], C. Bodendorf[20], E. Branchini[22,23,2], M. Brescia[24,25,26], J. Brinchmann[27], S. Camera[28,29,30], V. Capobianco[30], C. Carbone[31], J. Carretero[32,33], S. Casas[34], M. Castellano[35], S. Cavuoti[25,26], A. Cimatti[36], G. Congedo[37], C. J. Conselice[38], L. Conversi[39,40], Y. Copin[41], L. Corcione[30], F. Courbin[42], H. M. Courtois[43], A. Da Silva[44,45], H. Degaudenzi[46], A. M. Di Giorgio[47], J. Dinis[45,44], F. Dubath[46], C. A. J. Duncan[38,48], X. Dupac[40], S. Dusini[49], A. Ealet[50], M. Farina[47], S. Farrens[51], S. Ferriol[41], S. Fotopoulou[52], M. Frailis[8], E. Franceschi[6], S. Galeotta[8], B. Gillis[37], C. Giocoli[6,53], A. Grazian[54], F. Grupp[20,55], S. V. H. Haugan[56], H. Hoekstra[57], I. Hook[58], F. Hormuth[59], A. Hornstrup[60,61], K. Jahnke[62], E. Keihänen[63], S. Kermiche[64], A. Kiessling[65], M. Kilbinger[66], B. Kubik[41], M. Kümmel[21], M. Kunz[67], H. Kurki-Suonio[68,69], S. Ligori[30], P. B. Lilje[56], V. Lindholm[68,69], I. Lloro[70], D. Maino[1,31,3], E. Maiorano[6], O. Mansutti[8], O. Marggraf[71], K. Markovic[65], N. Martinet[5], F. Marulli[11,6,19], R. Massey[72], S. Maurogordato[73], H. J. McCracken[74], E. Medinaceli[6], S. Mei[75], Y. Mellier[76,74], M. Meneghetti[6,19], E. Merlin[35], G. Meylan[42], L. Moscardini[11,6,19], E. Munari[8], R. C. Nichol[17], S.-M. Niemi[77], C. Padilla[32], S. Paltani[46], F. Pasian[8], K. Pedersen[78], V. Pettorino[79], S. Pires[51], G. Polenta[80], M. Poncet[81], L. A. Popa[82], L. Pozzetti[6], F. Raison[20], R. Rebolo[83,84], A. Renzi[85,49], J. Rhodes[65], G. Riccio[25], E. Romelli[8], M. Roncarelli[6], E. Rossetti[18], R. Saglia[21,20], D. Sapone[86], B. Sartoris[21,8], P. Schneider[71], M. Scodeggio[31], A. Secroun[64], G. Seidel[62], M. Seiffert[65], S. Serrano[87,88,89], C. Sirignano[85,49], G. Sirri[19], J. Skottfelt[90], L. Stanco[49], C. Surace[5], A. N. Taylor[37], H. I. Teplitz[16], I. Tereno[44,91], R. Toledo-Moreo[92], F. Torradeflot[33,93], I. Tutusaus[94], E. A. Valentijn[95], L. Valenziano[6,96], T. Vassallo[21,8], A. Veropalumbo[2,23], J. Weller[21,20], G. Zamorani[6], J. Zoubian[64], E. Zucca[6], C. Burigana[97,96], V. Scottez[76,98], M. Viel[10,8,99,9,100], and L. Bisigello[97,85]

(Affiliations can be found after the references)

ABSTRACT

Multi-object spectroscopic galaxy surveys typically make use of photometric and colour criteria to select their targets. That is not the case of *Euclid*, which will use the NISP slitless spectrograph to record spectra for every source over its field of view. Slitless spectroscopy has the advantage of avoiding defining a priori a specific galaxy sample, but at the price of making the selection function harder to quantify. In its Wide Survey, *Euclid* aims at building robust statistical samples of emission-line galaxies with fluxes brighter than 2×10^{-16} erg s^{-1} cm^{-2}, using the Hα-[N II] complex to measure redshifts within the range [0.9, 1.8]. Given the expected signal-to-noise ratio of NISP spectra, at such faint fluxes a significant contamination by wrongly measured redshifts is expected, either due to misidentification of other emission lines, or to noise fluctuations mistaken as such, with the consequence of reducing the purity of the final samples. This can be significantly ameliorated by exploiting the extensive *Euclid* photometric information to identify emission-line galaxies over the redshift range of interest. Beyond classical multi-band selections in colour space, machine learning techniques provide novel tools to perform this task. Here, we compare and quantify the performance of six such classification algorithms in achieving this goal. We consider the case when only the *Euclid* photometric and morphological measurements are used, and when these are supplemented by the extensive set of ancillary ground-based photometric data, which are part of the overall *Euclid* scientific strategy to perform lensing tomography. The classifiers are trained and tested on two mock galaxy samples, the EL-COSMOS and Euclid Flagship2 catalogues. The best performance is obtained from either a dense neural network or a support vector classifier, with comparable results in terms of the adopted metrics. When training on *Euclid* on-board photometry alone, these are able to remove 87% of the sources that are fainter than the nominal flux limit or lie outside the $0.9 < z < 1.8$ redshift range, a figure that increases to 97% when ground-based photometry is included. These results show how by using the photometric information available to *Euclid* it will be possible to efficiently identify and discard spurious interlopers, allowing us to build robust spectroscopic samples for cosmological investigations.

Key words. methods: statistical – methods: data analysis – techniques: photometric – surveys – galaxies: distances and redshifts

1. Introduction

The ESA *Euclid* mission will carry out an imaging and spectroscopic survey over one third of the sky (Laureijs et al. 2011). The imaging channel will enable measurements of cosmic shear providing a tomographic view of the matter distribution, while the spectroscopic redshift survey will map the large-scale structure in three dimensions. Jointly, the two probes will yield unprecedented constraints on the cosmological model (Euclid Collaboration: Blanchard et al. 2020).

⋆ This paper is published on behalf of the Euclid Consortium.
⋆⋆ e-mail: marina.cagliari@unimi.it

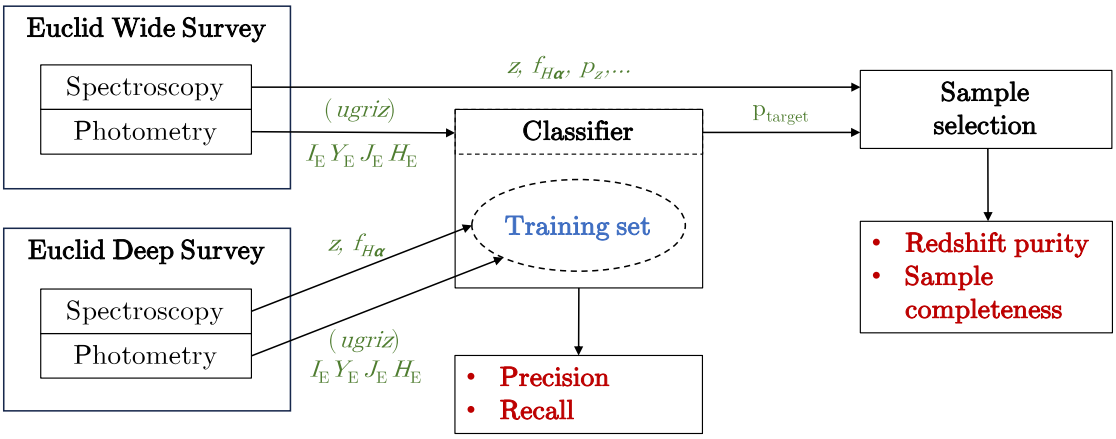


Fig. 1. Schematic description of the spectroscopic sample selection pipeline. The flowchart shows where a photometric target selection would be inserted in the spectroscopic selection pipeline. The photometric classifier performance is quantified by its precision and recall (defined in Sect. 2.1), while the final spectroscopic sample is characterised by the redshift purity and sample completeness.

The *Euclid* near-infrared spectrograph and photometer (NISP; Maciaszek et al. 2022) has three broadband filters for imaging, Y_E, J_E, and H_E (Euclid Collaboration: Schirmer et al. 2022) and a set of grisms for spectroscopy, while the visual instrument (VIS; Cropper et al. 2016) images through a single broad pass band, I_E, spanning the range [530, 920] nm, with high spatial resolution of 0.1 arcsec/pixel. Jointly, these two instruments will carry out the Euclid Wide and Deep Surveys (Euclid Collaboration: Scaramella et al. 2022). The NISP instrument operates as a slitless spectrograph, to record the dispersed light of all sources in the field of view to a nominal emission-line flux limit of 2×10^{-16} erg s^{-1} cm^{-2}, which corresponds to a 3.5σ detection of a 0.5 arcsec diameter source in the Wide survey as designed. The use of slitless spectroscopy makes the spectroscopic survey highly efficient, since individual sources do not need to be targeted; however, reliable redshift measurements will only be secured for a fraction of the galaxies that are detected photometrically. The Wide Survey will detect the most luminous Hα emitters over the redshift range $0.9 < z < 1.8$, with typical broadband flux corresponding to $H_E \lesssim 24$; however, it will be sensitive to continuum emission only from the most luminous galaxies and, so, the redshift estimation will be based primarily on the detection of emission lines (Euclid Collaboration: Gabarra et al. 2023). The Wide Survey will be complemented by the Deep Survey, which will reach 2 magnitudes deeper in flux over an area of 50 deg^2 split over three separate fields. In the Deep Survey blue grism ([926, 1366] nm) observations will complement those with the standard red grism ([1206, 1892] nm). Both the grisms have a dispersion of 13 Å/pixel. With greater sensitivity and an extended wavelength range, the Deep Survey will be used to construct a reference galaxy sample with secure spectroscopic redshift measurements, to characterise the selection function and redshift error distribution of the Wide Survey.

The design of the *Euclid* spectroscopic survey poses a particular challenge for sample selection: bright emission-line galaxies for which the redshift can be measured make up a small fraction of all photometrically detected sources and this sample is not known beforehand. We can illustrate our expectations of the *Euclid* spectroscopic sample using the Flagship2 mock galaxy catalogue, which was calibrated against the Hα luminosity function model 3 of Pozzetti et al. (2016). The mock catalogue contains approximately 2×10^5 galaxies/deg^2 to the magnitude limit $H_E < 24$. Out of this sample, only 2% are in the redshift range $0.9 < z < 1.8$ and have Hα emission-line flux greater than 2×10^{-16} erg s^{-1} cm^{-2}. The majority of the photometrically detected sources with $H_E < 24$ will leave no signal on the spectrograph, being either too faint in continuum emission, or not having a detectable emission line in the wavelength range of the red grism. When targeting galaxies at the low signal-to-noise limit, spurious noise features can be mistaken for emission lines leading to wrong redshift measurements. Current end-to-end tests of the data reduction pipeline suggest that the spurious detection rate is even higher than the naive prediction based on Gaussian noise statistics due to artefacts from spectral contamination. If not appropriately treated, such wrong redshifts in the galaxy catalogue degrade the cosmological constraints derived from the two-point correlation function or power spectrum galaxy clustering statistics (Addison et al. 2019).

In principle, when selecting the sample for analysis all available information should be used to minimise the fraction of spurious measurements, while at the same time, maximising the number density of the sample, or another figure of merit. However, the benefits from including additional constraints in the sample selection criteria must be carefully weighed against potential systematic biases. In the case of *Euclid*, including additional information from ground-based photometry modifies the selection function of the survey and could couple the sample with unwanted systematic effects that arise from observations made through the Earth's atmosphere (see, e.g., Ross et al. 2011, for a quantitative discussion of the impact of angular systematics on the measured clustering). The trade off of adding ground-based information will clearly also depend on the scientific analysis being considered. With slitless spectroscopy, since every galaxy in the field is in any case observed, we shall have the important advantage of being able to test a posteriori the impact of any chosen selection on the measured clustering, and evaluate the robustness of the results.

Our aim with this work is to investigate photometric classification criteria that are sensitive to both redshift and emission-line flux, in order to identify the sources that are likely to give successful spectroscopic redshift measurements in the Wide Survey. This strategy is similar to the methods used in ground-based spectroscopic surveys that make use of magnitude and colour selections to build the target sample for spectroscopy. For example, colour selections were applied to build the Sloan Digital Sky Survey Luminous Red Galaxy sample (Eisenstein et al. 2001) and the VIMOS Public Extragalactic Redshift Survey (VIPERS; Guzzo et al. 2014). A sample of emission-line galaxies was tar-

geted by the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) using a colour selection (Comparat et al. 2016), and a similar approach was adopted for the emission-line galaxy sample targeted by the Dark Energy Spectroscopic Instrument (DESI; Raichoor et al. 2023).

As a generalisation of the conventional colour cuts that are made in a two-dimensional colour-colour plane, we apply machine learning-based classification algorithms. These algorithms are well suited to optimising classification tasks in a high-dimensional parameter space. Thus, we expect them to outperform simple selection rules.

An option that is immediately available for such a use are photometric redshifts. *Euclid* will construct an unprecedented photometric redshift catalogue from the combination of ground-based and *Euclid* photometric bands. However, as we will discuss, photometric redshifts alone do not solve the problem. Even if photometric redshifts allow us to select a sample of galaxies at the target redshift range, additional criteria on galaxy physical properties, such as the star-formation rate, will still be needed to identify the population with bright emission lines (see Sect. 4.4).

A schematic representation of the *Euclid* spectroscopic sample selection pipeline is shown in Fig. 1. A redshift measurement will be performed for all sources detected in photometry, and will be accompanied by an assessment of its confidence level, as well as the measurements of spectral features including emission-line fluxes. Sources that do not have a significant detection in spectroscopy should be assigned a low measurement confidence. Additionally, *Euclid* will produce photometric catalogues based on the $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$-band images, which will be augmented with ground-based measurements (u, g, r, i, z) needed particularly for photometric redshift estimation (Stanford et al. 2021).

The photometric classification that we discuss enters as a second input to spectroscopic sample selection. The classifier can be trained on the Deep Field catalogues, which is expected to give robust redshift measurements for the emission-line target galaxies in the Wide Survey. The classifier will be applied to the photometric data of the Euclid Wide Survey, and its results combined with the spectroscopic measurements to build the final selected sample. This can be characterised in terms of its 'redshift purity' and 'sample completeness'. Any photometric criteria will necessarily reduce the number density of the sample; however, if emission-line galaxy targets can be identified from the photometry, this will increase the fraction of correctly-measured redshifts and improve the purity.

We use the terms sample completeness and redshift purity to characterise the quality of the *Euclid* spectroscopic samples. We define completeness with respect to the Hα emission-line galaxy sample that exists in the Universe, which we call the true targets.[1] These are defined by a set of intrinsic properties, including angular position, redshift, size, and flux, that do not depend on the measurement process. Once the observations are made, we construct the sample catalogue which contains the set of measured properties, signal-to-noise estimates, and quality flags for the detected sources. The completeness tells us the fraction of the true targets that have a correct redshift measurement and make it into the sample for analysis,

$$C = \frac{N_{\text{True Targets \& Sample \& Correct-}z}}{N_{\text{True Targets}}} . \quad (1)$$

On the other hand, the redshift purity tells us the fraction of the sample that has a correct redshift measurement,

$$P = \frac{N_{\text{Sample \& Correct-}z}}{N_{\text{Sample}}} . \quad (2)$$

The redshift purity only makes reference to the sample selected for analysis and does not depend on other intrinsic properties of the galaxies besides redshift.[2]

In this paper, we focus on the photometric classification, which is one step of the selection process illustrated in Fig. 1. We consider the potential gain from the photometric classification in terms of its precision and recall (defined in Sect. 2.1), which will impact the final purity and completeness of the spectroscopic redshift sample. The photometric selection reduces the size of the sample in the numerator of completeness (Eq. 1) and thus leads to a lower value of completeness. However, it acts on both the numerator and denominator of purity (Eq. 2), and so is a way to potentially boost the purity. The propagation of the photometric classification to the spectroscopic sample selection and the computation of purity and sample completeness requires full end-to-end simulations of the *Euclid* reduction pipeline. In Sect. 5, we will present results from preliminary simulations based on the *Euclid* spectroscopic pipeline, leaving a more detailed investigation to follow-up work.

The paper is organised as follows. In Sect. 2 we present the different algorithms we tested, and introduce the metrics we used to quantify the classifier performance. In Sect. 3 we discuss the mock catalogues, the noise model we apply to the photometry, and give the target definition. The results of the different analyses are presented in Sect. 4 and discussed in Sect. 4.4. In Sect. 5 we discuss how the photometric selection affects the spectroscopic sample. We conclude in Sect. 6.

2. Classification algorithms

A classifier is an algorithm that outputs the probability of an object of being an element of a given class, or group. For the purpose of this work, which is to identify target galaxies from their photometric properties, we use a binary classifier. In this case, the algorithm simply outputs the probability p of the object being a target, and $1 - p$ the probability of it being a non-target. A galaxy enters the target sample if $p > p_{\rm thresh}$, where $p_{\rm thresh}$ is a threshold probability value. How the threshold is chosen is discussed in Sect. 2.1.

In this work we tested six different machine learning classifiers. The first three are self-organising maps (SOMs), dense neural networks (NNs) and support vector machine classifiers (SVCs). The other three are voting classifiers based on decision trees: the random forest (RF), the adaptive boosting classifier, or AdaBoost (ADA), and the extremely randomised tree classifier, or extra-tree classifier (ETC). These specific algorithms were chosen for our tests as they are known to perform well in classification tasks and are able to identify non-linear boundaries between classes.

[1] This definition differs from that typically used in ground-based multi-object spectroscopic surveys that define completeness with respect to a known target sample constructed from photometric catalogues. Since the detection in *Euclid* spectroscopy will depend primarily on the signal-to-noise ratio of the emission lines, the sample with spectroscopic redshifts will not be representative of a simple photometric selection.

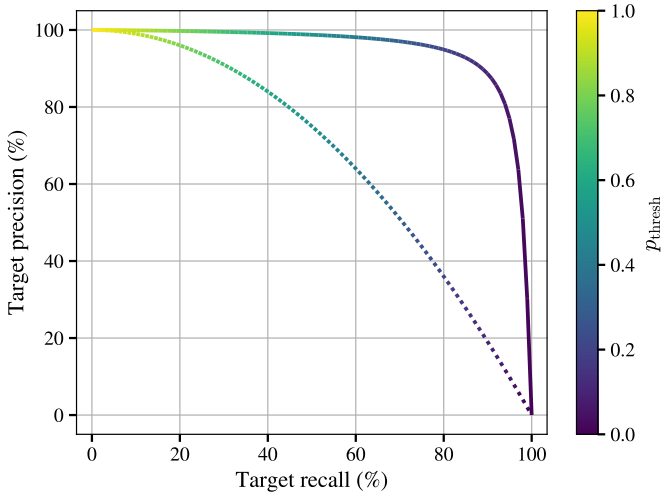


Fig. 2. Relationship between precision and recall of a classifier. The lines are colour-coded as a function of the classification probability threshold. The solid and dotted lines show the behaviour of two classifiers for illustration. The classifier represented by the solid line performs better than the dotted line since it gives higher precision and recall.

2.1. Classification metrics

To compare the results from different classifiers, we adopt three metrics defined from their 'confusion matrix'. The elements of the confusion matrix of a binary classifier are the counts of true positives ($N_{\rm TP}$), true negatives ($N_{\rm TN}$), false positives ($N_{\rm FP}$), and false negatives ($N_{\rm FN}$). Our chosen metrics are the 'precision', 'recall', and 'false positive rate' (FPR), defined respectively as

$$\text{precision} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FP}}\,, \tag{3}$$

$$\text{recall} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FN}}\,, \tag{4}$$

$$\text{FPR} = \frac{N_{\rm FP}}{N_{\rm FP} + N_{\rm TN}}\,. \tag{5}$$

The precision is the fraction of the selected sample that are true targets, i.e., it quantifies the level of contamination due to wrongly classified sources. The recall, also known as 'true positive rate', is the fraction of true targets that are identified correctly (as $N_{\rm TP} + N_{\rm FN}$ corresponds to the total number of targets). The false positive rate, or 'fall-out', is the fraction of non-targets that are mislabelled as targets and enter the selected sample as interlopers. The complement of the false positive rate is the 'true negative rate',

$$\text{TNR} = \frac{N_{\rm TN}}{N_{\rm FP} + N_{\rm TN}} = 1 - \text{FPR}\,, \tag{6}$$

which characterises the fraction of non-targets that are correctly removed from the sample.

These metrics change as functions of the probability threshold chosen for the classifier, i.e., the probability value $p_{\rm thresh}$ above which an object is classified as a target. This is a hyper-parameter of the model, which we set to maximise a chosen metric. In a binary classification, a training set is said to be 'balanced' when it is evenly split between targets and non-targets, and $p_{\rm thresh} \sim 0.5$. When the training set contains a much larger number of targets than non-targets, or vice versa, it is called 'unbalanced', and we refer to this case as an 'unbalanced classification'. In general, in unbalanced classifications the optimal probability threshold is very different from 0.5. Precision and recall can be computed as a function of $p_{\rm thresh}$ and plotted against each other, as shown in the example of Fig. 2. Such a plot is very informative for the photometric selection task that is the scope of our work. In Fig. 2 we present two possible behaviours of this curve. The solid line is an almost ideal classifier that has high precision also when the recall is high, while the dotted curve corresponds to a classifier with worse performance. Since the photometric criteria make up only one step of the spectroscopic sample selection process (see Fig. 1), we want to keep the recall of the photometric classification as high as possible. In other words, we want to get a resulting sample as complete as possible, discarding the minimum number of true targets. Thus, we choose a specific value for the recall and, from this relation, derive the corresponding precision yielded by the algorithm. We use the precision at 95% recall as our benchmark value. A similar plot can be produced in terms of redshift purity and sample completeness. The shape of this curve will depend on the chosen probability threshold, and consequently the recall, of the photometric classification. In Sect. 5 we justify the choice of the 95% recall value and present results for the redshift purity and sample completeness.

Finally, we use the false positive rate as the main metric to compare algorithms trained with different input features (see Sect. 4.4). The false positive rate helps to visualise the fraction of misidentified objects in terms of redshift or emission-line flux and shows the source of the contaminants.

2.2. Self-organising map

Self-organising maps (Kohonen 1982, 1990) use unsupervised learning to project a high dimensional feature space onto a lower dimensional one, usually a two-dimensional space, as the name map suggests. We build a 55×55 map trained for 60 epochs, where an epoch corresponds to an iteration of the algorithm during which the entire training set is processed. To train the self-organising map, in addition to the photometric features used as inputs for all the other methods, we also add the target label (see Sect. 3). Then, when projecting new data onto the self-organising map the target labels are removed. These steps make the implementation of the self-organising map presented here more similar to a supervised learning algorithm. We also introduce a weight, $w_{\rm SOM}$, of the photometric features, which enables us to control the importance of the label in the training. This is a hyper-parameter of the self-organising map model. Finally, the probability of an object of being a target is defined by the target fraction in the cell it has been projected onto. The self-organising maps were implemented using `SOMPY` (Moosavi et al. 2014).

2.3. Neural network

Neural networks are by far the most popular supervised learning algorithms. They can be described as a sequence of layers; when the inputs are processed by a layer they first undergo a linear transformation and then a nonlinear function is applied to

[2] We do not consider the sample purity, which can include other criteria such as flux, since our main objective is to select galaxies with good redshift measurements for the galaxy clustering analysis.

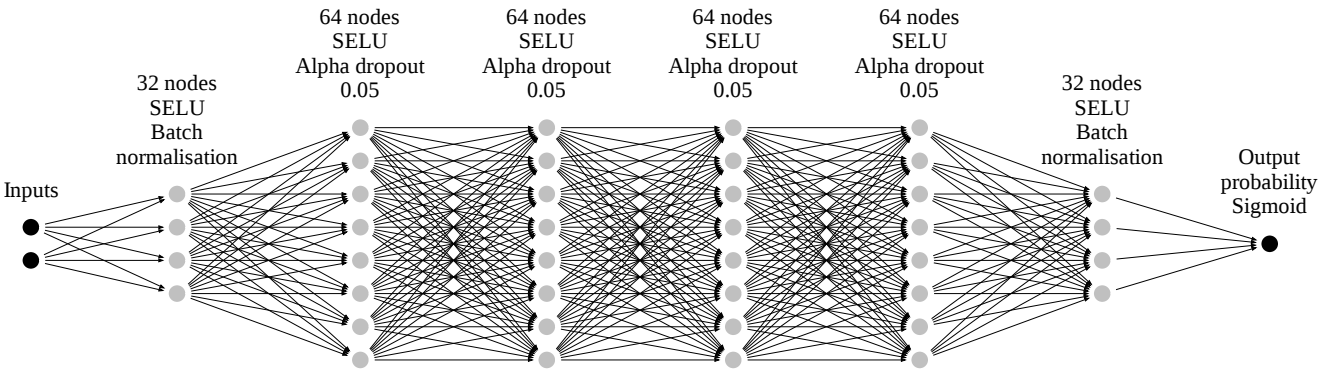


Fig. 3. A schematic representation of the neural network architecture used for classification. Values pass from the input to the output along the connected edges; each node represents a linear combination of the inputs and the application of a non-linear activation function. The value at the output represents the binary classification probability between 0 and 1. The number of input neurons varies for the different configurations (4 for *Euclid*-only, and 8 after adding ground-based photometry, see Sect. 4). For visualisation, the number of neurons in each hidden layer has been divided by 4.

them. During the learning process the neural network updates the coefficients, usually called weights, of the linear transformation of each layer in order to fit the target function $y = f(x)$ that relates the inputs x, to the labels y. This structure enables neural networks to potentially fit any function of the input features (LeCun et al. 2015).

Our neural network architecture was optimised for the problem at hand. Figure 3 shows a schematic representation of the neural network. The input layer is followed by a first block that consists of a dense layer with 32 neurons and a batch normalisation layer (Ioffe & Szegedy 2015). Then, a second block which consists of a dense layer with 64 neurons and an alpha dropout layer (Klambauer et al. 2017) with rate 0.05 is repeated four times. Finally, the first block is repeated before the output layer, which consists of 1 neuron. The activation function of all the layers except for the output is a scaled exponential linear unit (SELU; Klambauer et al. 2017). The last layer, as it has to output a probability, has a sigmoid activation function. Since the ratio between positive and negative examples is very low, we opted for a sigmoid focal cross entropy loss function (Lin et al. 2017),

$$\mathrm{FL}(p) = -\alpha\,(1-p)^{\gamma}\ln(p)\,, \tag{7}$$

where α and γ are two hyper-parameters of the model. We use $\alpha = 0.6$ and $\gamma = 4$. We implemented the neural network in the TensorFlow2 framework (Abadi et al. 2015).

2.4. Support vector machine classifier

Support vector classifiers (Boser et al. 1992) partition the feature space by applying a kernel transformation to map curved boundaries into planes and finding the maximum-margin hyperplane that separates the classes. It is important to note that for our training we weight differently the target and non-target examples. This weighting is necessary in the case of imbalanced classes. Alternatively, one could select a balanced subsample of the original training set. However, such a solution would greatly reduce the size of the training sample. Our approach uses the support vector classifier implementation of `scikit-learn` (Pedregosa et al. 2011), which has an inbuilt functionality to balance the sample via weighting.

We adopt the `scikit-learn` default kernel, which is the radial basis function kernel (RBF),

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma\,\|\mathbf{x}-\mathbf{x}'\|^2\right)\,, \tag{8}$$

where $\|\mathbf{x}-\mathbf{x}'\|^2$ is the Euclidean squared distance and γ is the hyper-parameter that controls the dimension of the region of influence of the training point.

2.5. Decision tree-based classifiers

The last three classifiers are voting or ensemble classifiers. In general, a voting classifier is an algorithm that combines the output of different base classifiers through a vote, which can be weighted or not. In this work we used classifiers based on the same base algorithm, the decision tree. These classifiers differ in how they split the data set to train the trees, how they build the trees, and how they combine together their probability outputs.

A decision tree is a supervised machine learning model that approximates a function with a series of simple decision rules (see Hastie et al. 2001, Chap. 9). Decision trees have the advantages that they can be easily visualised, have high explainability, and require very little data preparation; however, they can easily over-fit the training sample making their output and final structure dependent on the training set. These issues can be reduced by combining the results of different trees (e.g., Bauer & Kohavi 1999).

The first of these voting classifiers are random forests. Random forests (Breiman 2001) are an ensemble of decision trees each one trained with a subsample of the training set. This subsample is a bootstrap sample, which means its elements are randomly selected with replacement from the complete training set. The final output of the random forest for the classification task is a majority voting between all the decision trees of the forest. Random forests very efficiently reduce the over-fitting of single decision trees. To take into account the class imbalance of the sample we weigh the two class examples by the inverse of their frequency. The weights are computed for each bootstrap subsample.

The second ensemble classifier is a discrete adaptive boosting classifier (Freund & Schapire 1995). Differently from the random forest, adaptive boosting classifiers can use different

base classifiers. In this work, we limited the analysis to adaptive boosting classifiers based on decision trees with weighted data to balance the sample examples. Adaptive booster classifiers combine the results of subsequently trained base learners with a weighted majority vote. At each step of the training a new learner is built from the training set, which is re-weighted to reduce the importance of data that have been correctly classified in the previous steps.

Finally, the last algorithm we use is the extra-tree classifier. Extra-tree classifiers are ensemble classifiers based on decision trees (Geurts et al. 2006). An extra-tree classifier is composed of a group of decision trees, which are trained with bootstrap subsamples of the training set, as in random forest training. The difference between a random forest and an extra-tree classifier lies in how the decision rules of the trees are selected. In random forests, the splits of the tree nodes are deterministic and depend on the selection algorithm; in extra-tree classifiers, instead, they are randomly drawn and the final rule is chosen as the best-performing one among them. This helps in reducing even more the variance of the method. All three voting classifiers are implemented in `scikit-learn`, and the function to weigh the data to balance them is part of their built-in functionalities.

3. Benchmark data

3.1. Mock galaxy catalogues

We use two catalogues to benchmark the selection algorithms: the EL-COSMOS catalogue and the Euclid Flagship2 mock galaxy catalogue. These catalogues include broadband photometry, emission-line fluxes and morphological properties. We make use of the *Euclid* photometric bands from VIS, $I_{\rm E}$, and NISP, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$ (Euclid Collaboration: Schirmer et al. 2022), with depths listed in Table 1. Additionally, photometric data from multiple ground-based surveys will be included in *Euclid* analyses to extend the wavelength coverage to the optical with u, g, r, i, and z bands and obtain reliable photometric redshifts that are key for *Euclid* weak lensing science. These include the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST, LSST Science Collaboration et al. 2009), the Dark Energy Survey (DES, Flaugher 2005), and the Ultraviolet Near Infrared Optical Northern Survey (UNIONS).[3] In order to benchmark the photometric selection in this work, we use the *ugriz* filters and UNIONS survey depths, which are listed in Table 1. Hereafter, we refer to the photometry of the four *Euclid* filters as *Euclid* photometry, and to the photometric data from the five optical filters as ground-based photometry. The photometry does not include the effect of Milky Way extinction.

The resolution of *Euclid* NISP spectroscopic observations is not sufficient to separate Hα from its neighbouring [N II] λ6549 and [N II] λ6584 companions. As such, *Euclid* will measure the combined flux of this triplet of emission lines, which we shall use here and indicate for brevity as

$$f_{\rm H\alpha+[N\,II]} = f_{\rm H\alpha} + f_{\rm [N\,II]\lambda6549} + f_{\rm [N\,II]\lambda6584}\,. \tag{9}$$

We refer to the triplet as the 'Hα complex'.

We also investigate the benefit of adding morphological information to the target classification. The two mock galaxy catalogues we use here include morphological model parameters including disk ellipticity, bulge scale, disk scale, and bulge-to-disk ratio; however, since these properties will not be, in general, directly measured from the data, we used them to derive the observable half-light radius, $r_{\rm half}$, and axial ratio, e. To do this, we ran `GALSIM` (Rowe et al. 2015) using the morphological parameters for each mock galaxy to generate a simulated image of the galaxy as it would be observed by VIS, from which we estimated the half-light radius and axial ratio. We carried out this procedure only for the Flagship2 catalogue.

We use only galaxies in the mock catalogue, without accounting for the possibility that stars or active galactic nuclei may be misclassified in real data and enter the sample. Contamination from faint stars, in particular, can potentially reduce the purity of the galaxy sample. The severity of such contamination depends on the performance of the star-galaxy classification, which is a separate step of the *Euclid* data analysis and whose impact is beyond the scope of this work.

3.1.1. EL-COSMOS

The EL-COSMOS catalogue is an extension of the COSMOS 2020 photometric catalogue (Weaver et al. 2021). The COSMOS catalogue is a multi-band data set assembled in the *Hubble* Space Telescope COSMOS field over the past fifteen years (Scoville et al. 2007). The catalogue was extended as described in Saito et al. (2020) with synthetic photometry and emission-line fluxes. To assign the fluxes of the emission lines the authors combined spectral energy fits of the stellar continuum, which correlates with the intrinsic emission line fluxes, with a careful modelling of dust attenuation as a function of redshift. We use an update to the emission-line catalogue produced for the Euclid Consortium (Euclid Collaboration, in prep.). It contains about 2×10^5 galaxies and 2000 active galactic nuclei. This catalogue also contains stars observed in the COSMOS field, which, as explained, we do not consider.

3.1.2. Euclid Flagship

The Euclid Flagship2 mock galaxy catalogue (Euclid Collaboration, in prep.) is based on the Flagship2 N-body simulation, the large reference simulations built by the Euclid Consortium. Galaxies were added to the simulation using an extended halo occupation distribution model. The Flagship2 galaxy mock catalogue represents an improvement with respect to the previous version in terms of modelling of the galaxy properties. The catalogue contains photometric and spectroscopic information, morphological parameters, along with lensing properties. The morphological parameters are correlated with the galaxy properties to reproduce observed trends in galaxy size. For our work, we selected a subsample of $\sim 2 \times 10^5$ objects that contains a number of galaxies comparable to EL-COSMOS. We note that Flagship2 does not contain active galactic nuclei, while EL-COSMOS contains about 2000 of them.

An additional step must be taken to compute the total flux of the Hα complex for Flagship2 mock galaxies. The catalogue gives the flux of Hα and of the [N II] λ6584 line only. Assuming a relative 1:3 ratio for the [N II] doublet, we estimate the total flux as

$$f_{\rm H\alpha+[N\,II]} = f_{\rm H\alpha} + \frac{4}{3} f_{\rm [N\,II]\lambda6584}\,. \tag{10}$$

The emission-line fluxes in Flagship2 were calibrated against the Hα luminosity function model 3 of Pozzetti et al. (2016). We use the line and broadband fluxes with internal dust attenuation applied. From Flagship2 we use both *Euclid* and ground-based photometric data, as well as the morphological parameters derived as discussed earlier.

[3] https://www.skysurvey.cc/aboutus/.

Table 1. Point source magnitude limits at depth $(S/N)_{lim} = 10$ for *ugriz*, and for I_E, Y_E, J_E, and H_E.

Band	$m_{lim,10\sigma}$
u	23.5
g	24.4
r	24.1
i	23.5
z	23.3
I_E	24.6
Y_E	23.0
J_E	23.0
H_E	23.0

Notes. The 10σ point source depth values in AB magnitude adopted for each filter.

The Flagship2 catalogue also provides photometric redshift estimates, hereafter photo-*z*s, obtained with state-of-the art algorithms using both *Euclid* and ground-based photometry (Euclid Collaboration: Desprez et al. 2020). In order to allow the computations of photo-*z*s for billions of *Euclid* sources, a two-stage approach has been adopted. First, `Phosphoros`, a template-fitting code (Paltani et al. in preparation), is used to compute the redshift probability distribution functions on a sample of galaxies selected from reference fields that benefit from very deep observations in a large number of photometric bands (e.g., COSMOS; Weaver et al. 2021). The *k*-nearest neighbour photometric redshift algorithm (Tanaka et al. 2018) is then used to estimate the posterior distributions of redshift for sources in the Euclid Wide Survey. This procedure was replicated in the Flagship2 mock galaxy catalogue. In this work we use the first mode of the posterior redshift distribution as the photo-z estimate. We use the photo-z to select galaxies within the redshift range of interest and compare the metrics with the results from the trained classifiers.

3.2. Noise model

The errors on the broadband photometric measurements were simulated assuming background-limited observations (Euclid Collaboration: Pocino et al. 2021) such that the standard deviation on the measurement is

$$\sigma_f = \frac{f_{lim}}{(S/N)_{lim}}, \tag{11}$$

where f_{lim} is the flux at the specified signal-to-noise limit $(S/N)_{lim}$. In Table 1 we show the AB magnitude limits, m_{lim}, corresponding to f_{lim} for $(S/N)_{lim} = 10$.[4] The observed fluxes were then extracted from a Gaussian distribution with the true galaxy flux, f, as mean, and variance given by σ_f. In order to be able to reproduce the results, we constructed observed catalogues for both EL-COSMOS and Flagship2, which contain realisations of the flux errors produced following the recipe described above.

The driving idea in the application of our selection procedure to the real *Euclid* data is that the training set will be constructed from the higher signal-to-noise data of the Euclid Deep Fields, which will have high completeness and purity at the depth of the Wide Survey. In order to build a training set that matches the noise properties in the Wide Survey, the photometry from the Deep Fields will have to be either measured in Wide-like stacks or degraded appropriately to match the noise level of the Wide Survey.

[4] The magnitude limits for UNIONS in Table 1 were computed from the 5σ limits available at `https://www.skysurvey.cc/survey/`.

3.3. Sample selection and pre-processing

For our analysis, we selected from the EL-COSMOS and Flagship2 catalogues two sub-samples limited to $H_E < 24$ (which corresponds to a 4σ point-source detection limit). In addition, as mentioned earlier, the resulting Flagship2 catalogue was further sparsely sampled in order to match the same number of objects of EL-COSMOS. Each catalogue was then split into three subsets, for training, validation, and testing, containing respectively 75%, 15%, and 10% of the total parent catalogue. In fact, the validation set is needed only for the training of the neural network; for the other algorithms we could use 90% of the total sample as the training set. However, for the sake of a fair comparison, we opted to use the same training and test sets for all methods, by discarding the validation set objects when not needed.

The galaxies we aim to select with the photometric selection have, on top of the $H_E < 24$ cut,

$$\begin{cases} 0.9 < z < 1.8 \\ f_{H\alpha+[N\,II]} > 2 \times 10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}} \end{cases}, \tag{12}$$

where z and $f_{H\alpha+[N\,II]}$ are the true redshift and emission line flux of the galaxies. The objects satisfying this selection are what we call 'target' galaxies. In terms of the classifier training, we assign a label 1 to the target galaxies and a label 0 to the remaining objects, hereafter non-targets. It should be noted that the $H\alpha+[N\,II]$ flux criterion in the target definition is specified to select galaxies with bright emission lines that are likely to give successful spectroscopic redshift measurements. We will see that this target definition does not impose a sharp flux cut in the measured sample; galaxies just below the flux limit still have a high probability of being selected and of giving a correct redshift measurement. Moreover, these galaxies will also contribute to the redshift purity metric.

The percentage of galaxies entering the target sample within the full $H_E < 24$ catalogues is very low: $\sim 8\%$ for EL-COSMOS and $\sim 3\%$ for Flagship2. The difference between the two catalogues is consistent with the current uncertainty in the Hα luminosity function at $z > 1$. The low target fractions of the two catalogues make the classification task extremely unbalanced. The solutions adopted for each classifier were discussed in Sect. 2 and span from weighting schemes to specific loss functions.

Finally, all input training parameters are pre-processed via standard scaling,

$$X = \frac{x - \bar{x}}{\sigma_x}, \tag{13}$$

where $\bar{x}$ is the mean value of input feature x over the training sample, and σ_x its standard deviation. After this normalisation, the sample has zero mean and unit standard deviation, which makes the training of the algorithms more efficient, typically leading to better results.

4. Results and discussion

4.1. Benchmark selections

Before discussing the performance of the machine learning algorithms we present the results from simple classifiers based on

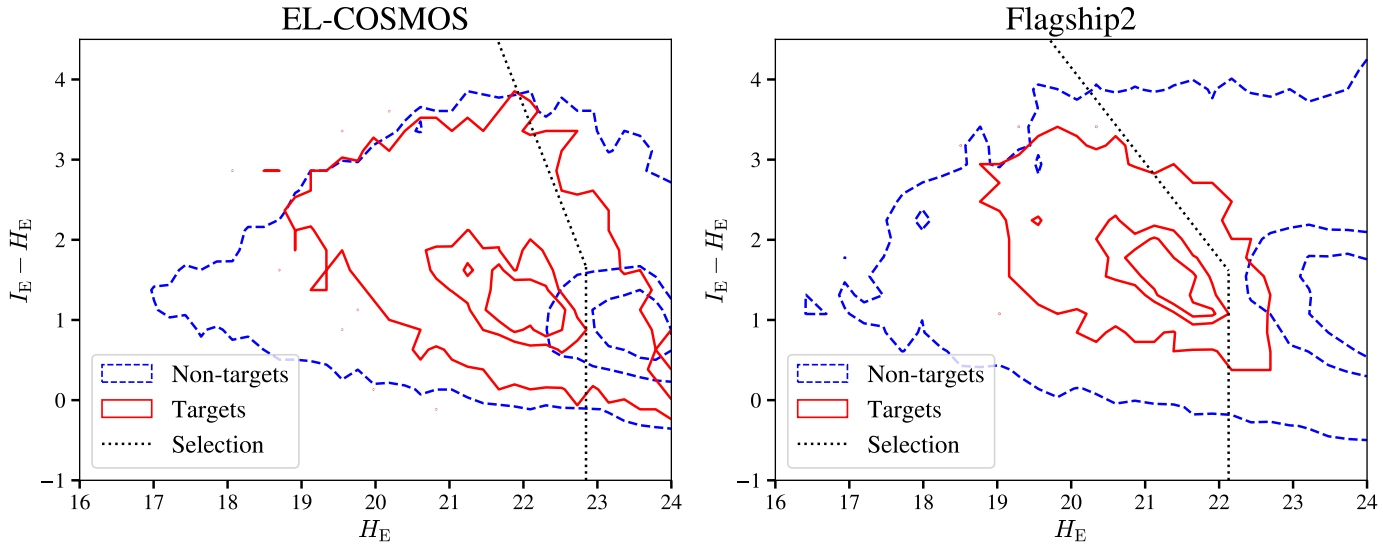


Fig. 4. Optimised colour selection in the $(I_E - H_E)$ versus H_E colour-magnitude plane for EL-COSMOS and Flagship2. The blue dashed lines correspond to the non-target distribution and the solid red lines to the target distribution. The contours contain 99%, 50%, and 25% of the samples. The dotted black segments represent an optimised colour cut in this plane corresponding to recall ∼ 95%.

Table 2. Recall (%) and precision (%) for different H_E cuts.

	EL-COSMOS		Flagship2	
H_E cut	Recall	Precision	Recall	Precision
22.84	95	13.8	-	-
22.06	-	-	95	8.9
21.0	20.2	11.3	47.6	10.0
22.0	63.1	16.3	93.3	9.1
23.0	97.1	12.7	100	4.8
24.0	100	7.8	100	2.6
Colour cut	95	14.3	95	9.9

Notes. Precision and recall for EL-COSMOS and Flagship2 at different H_E magnitude limits. The first two rows correspond to the H_E cut that gives 95% recall respectively for EL-COSMOS and Flagship2.

magnitude and colour with *Euclid* photometry. These tests provide a benchmark for the machine learning classifiers. We focus on the $(I_E - H_E)$ versus H_E plane, which shows the largest displacement between targets and non-targets (see Figs. 4 and A.1). The distributions are seen to be most separated in H_E magnitude. Indeed, the use of H_E is expected to be particularly suited to capture information on the Hα flux, as it covers the $[1.5, 2.0]\,\mu$m band, which encompasses the Hα complex for $1.3 \lesssim z \lesssim 2$. In addition, the $(I_E - H_E)$ colour is sensitive to redshift, since it spans the 4000 Å break at $z > 1$.

We thus begin by considering magnitude-limited samples in H_E. Table 2 gives the resulting recall and precision metrics for H_E cuts ranging from 21 to 24 magnitudes. For the Flagship2 catalogue, all targets have $H_E < 23$ giving 100% recall at that limit, while for EL-COSMOS, 100% recall is reached at $H_E < 24$.

Next, we consider a selection in the $(I_E - H_E)$ versus the H_E plane. The colour-magnitude selection reads as follows,

$$(I_E - H_E) < a\,(H_E - b) \quad \text{AND} \quad H_E < H_E^{\rm cut}\,. \tag{14}$$

We searched for a selection with the form of Eq. (14) that maximises the purity while giving recall ∼ 95%, which we chose as the reference value for comparing the algorithms (see Sects. 2.1 and 5). The best colour cut for EL-COSMOS has slope $a = -2.36$, pivot $b = 23.60$, and $H_E^{\rm cut} = 22.85$. For Flagship2 the slope is $a = -1.90$, $b = 14.74$, and $H_E^{\rm cut} = 22.13$.

Figure 4 shows the targets (solid red) and non-target (dashed blue) distributions in the colour-magnitude plane of interest for EL-COSMOS (left panel) and Flagship2 (right panel). The dotted black line corresponds to the colour-magnitude cut. The two panels show the difference in the target distributions of EL-COSMOS and Flagship2. Flagship2 does not have any targets with $H_E > 23$, in contrast, EL-COSMOS targets reach the magnitude limit of the sample. For this reason we allow the H_E cut to adapt to the training data. We report the precision of the optimised colour cuts in the bottom row of Table 2.

From Table 2, we see that all selections give a higher precision for EL-COSMOS than Flagship2. This can be understood since the fraction of targets is higher in EL-COSMOS than in Flagship2. The colour cut gives a marginal improvement in purity (0.5 – 1) over the H_E cut. We next show the results from machine learning classifiers, which make full use of the high-dimensional parameter space to optimise the selection.

4.2. Using Euclid data only

We first discuss the results obtained training the classifiers using only *Euclid* photometry, comparing the two catalogues EL-COSMOS and Flagship2. The input features for each object are the same for both catalogues, namely its H_E magnitude and near-infrared colours, $(I_E - Y_E)$, $(Y_E - J_E)$, $(J_E - H_E)$.

Figure 5 shows the precision-recall curves produced by the six different classifiers using respectively EL-COSMOS (left panel) and Flagship2 (right panel). We remark that for an ideal classifier the plot would show a close-to-flat precision around unity (see Fig. 2), followed by a sharp drop at the highest possible recall value. To provide a reference baseline, in Fig. 5 we also present (dotted magenta line) the curve one obtains when simply selecting $H_E < H_E^{\rm limit}$ magnitude-limited samples. The

Table 3. Precision values (%) at 95% recall for the different classifiers.

		SOM	NN	SVC	RF	ADA	ETC
Euclid	EL-COSMOS	13.9	17.5	17.3	16.4	12.9	16.7
	Flagship2	12.7	16.0	18.0	15.5	10.4	16.9
Euclid +	Flagship2 morphology	9.6	17.6	16.8	15.3	11.0	14.7
	EL-COSMOS ground	20.7	34.3	34.3	31.5	29.1	28.0
	Flagship2 ground	26.1	47.9	43.5	39.3	39.7	35.6

Notes. The two top rows give the results for training using *Euclid* photometry only, while morphological data and ground-based photometry, respectively, are used in the bottom rows. The relative uncertainty on all values is ~ 6%, estimated from multiple realisations of the training and test sets.

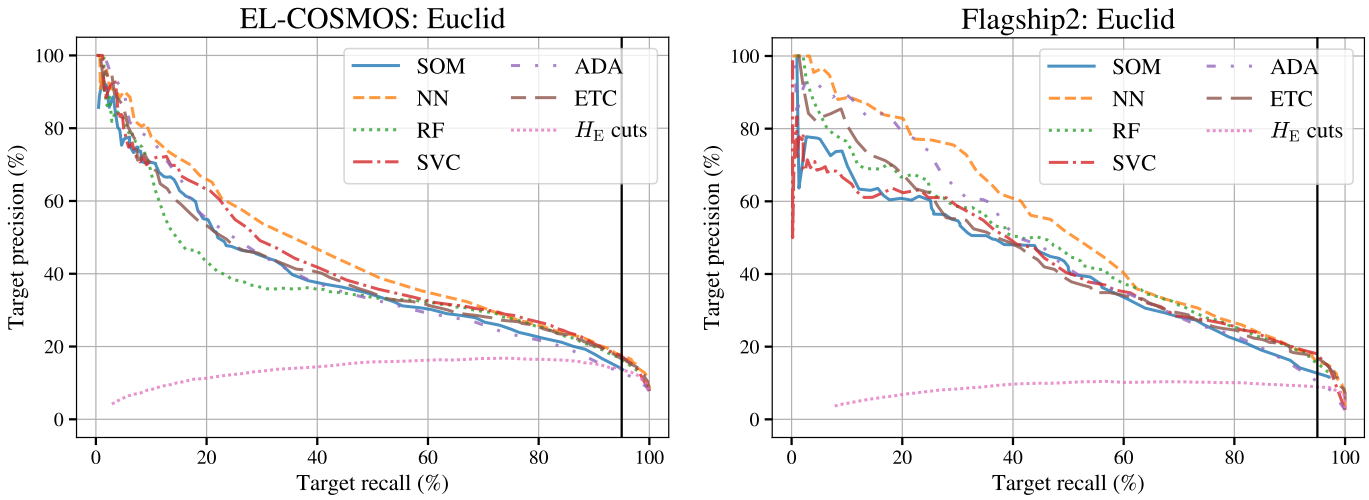


Fig. 5. Precision vs. recall performance of the different classifiers, using *Euclid* photometry alone for the training. The two panels correspond to the two test catalogues as indicated. The vertical solid line gives our reference recall value of 95%.

curve has been computed by smoothly varying H_E^{limit} between 20.0 and 24.0 (see Table 2). The vertical black line corresponds to 95% recall, as reported in Table 3.

Comparing the two panels, the first evident difference is the larger variance in performance over the whole recall range shown by the different algorithms in the case of the Flagship2 sample. Conversely, the classifiers trained with EL-COSMOS show a sharper drop in precision at small recall values. In both cases, the magnitude-limited selection is (not unexpectedly) worse than the machine learning classifiers, but in the case of EL-COSMOS the resulting performance becomes comparable to that of the worse-performing classifiers at 95% recall.

Overall, Fig. 5 and Table 3 show similar performance when training with either Flagship2 or EL-COSMOS, with the former showing a larger variance at the recall threshold. Such an agreement is an encouraging indication of the robustness of the general conclusions that can be drawn from these results. In both cases, the best-performing algorithms are the neural network, the support vector classifier, and the extra-tree classifier. The random forest follows shortly behind, indicating that the bootstrap resampling used in the decision tree training is especially efficient for this task. Last comes the self-organising map, which is not optimised for this kind of task, and the adaptive boosting classifier.

The effect of complementing *Euclid* infrared photometry with morphological information described by the galaxy half-light radius and axial ratio values can be seen in Fig. 6. The

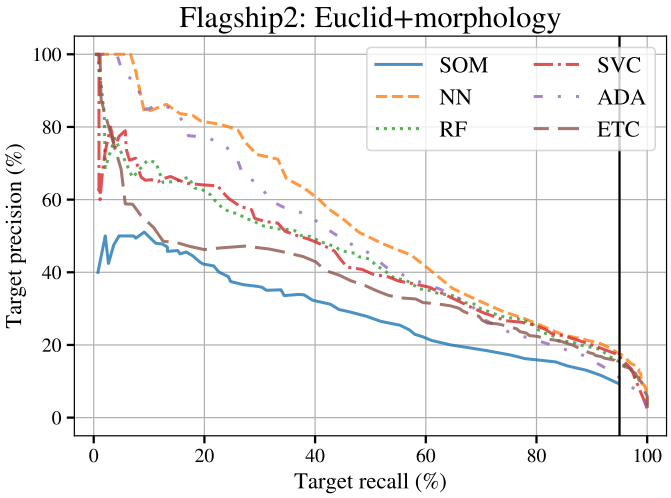


Fig. 6. Same as Fig. 5, but now adding morphological information in terms of half-light radius and axial ratio values.

plot shows no large improvement and some classifiers perform worse. We also observe an even larger variance between the different classifiers, especially at low recall values. The best-performing one is still the neural network, followed by the support vector classifier, the random forest, and the extra-tree

classifier. Again, the adaptive boosting classifier and the self-organising map fare poorly. A more detailed discussion is left for Sect. 4.4.

4.3. Adding ground-based photometry

When we combine *Euclid* and ground-based photometry we substitute the I_E band with the five ground-based filters, *ugriz*. In this case, the input features of the classifiers are the following seven colour combinations $(u-g)$, $(g-r)$, $(r-i)$, $(i-z)$, $(z-Y_E)$, (Y_E-J_E), and (J_E-H_E). In addition, we also use H_E as the last input feature.

Figure 7 shows how the diagnostic plots change when combining *Euclid* and ground-based photometry. We immediately see from Fig. 7 how the ground-based data improves the overall performance, yielding curves that are much closer to the ideal shape (see Fig. 2). In the case of Flagship2 we also mark (green dot) the precision ($\sim 5.5\%$) and recall ($\sim 92.4\%$) values recovered when using photometric redshifts to simply isolate targets with $0.9 \leq z_{\rm photo} \leq 1.8$, with no extra information to constrain the desired Hα line flux. We note that the photometric redshift selection does not reach the 95% recall value. We also consider photometric redshift selections with various H_E magnitude limits, shown by the magenta dotted line. In Appendix B we present a preliminary test that combines the photometric and the redshift information in the training of a neural network.

The improvement in performance appears to be larger when estimated using Flagship2 than with EL-COSMOS with a difference of $\sim 10\%$ in precision for all algorithms. The reason for this can be related to the colour distribution of the targets. In the EL-COSMOS catalogue the distribution functions of magnitudes and colours for targets shows more variance than in Flagship2 where the targets are more localised on colour space. When *Euclid*-only photometry is used, the information is not sufficient for tightly constraining the target region in the parameter space, thus producing similar results from the two catalogues. However, when ground-based photometry is added, in the Flagship2 case it becomes easier to isolate the targets. These differences may be due to the recipes used for assigning spectral energy distributions and synthetic emission lines in the two catalogues.

The relative ranking of the different classifiers derived from the two catalogues is the same. The worst performing algorithm at the recall threshold is the self-organising map, which shows a steeper drop in precision than the others (see Table 3). The remaining algorithms have precision values $> 35\%$ for Flagship2, with the neural network reaching almost 50%. For EL-COSMOS at the recall threshold the values of the precision are always $> 25\%$, peaking at $\sim 34\%$ for both the neural network and the support vector classifiers.

4.4. Comparison of the results

In this section, we focus on the results based on the Flagship2 training and discuss the results obtained with the three configurations. We will then focus on the best-performing classifier, the neural network, and discuss in more detail the three cases. We will also show a comparison with a simpler redshift-only selection based on *Euclid* photometric redshifts.

Figures 5, 6, and 7, together with Table 3, provide a direct quantitative comparison of the three training configurations: the best performance is obtained by the combined *Euclid* and ground-based photometry. For Flagship2, this more than doubles the precision at the recall threshold with respect to the other two configurations, a clear benefit of the extra information on lower redshift objects provided by the optical bands (see discussion in the following). The addition of morphological information through the half-light radius and ellipticity, conversely, does not introduce any significant improvement: the neural network and the adaptive boosting classifier show only a minimal gain, while all others worsen their performance.

The half-light radius, in particular, does show a trend as a function of redshift, but this relation has a large scatter and weak correlation coefficient. It is possible that other morphological measures that we did not consider, such as the Sérsic index, will be more sensitive to galaxy type and have a greater importance for classification; however, we reserve this investigation for future work. When fed uninformative features, the classification algorithms will tend to ignore them. The majority of the tested classifiers have, in fact, built-in mechanisms to ignore a feature. Specifically, the neural network would reduce, during the training, the weight of the specific feature that appears to be uninformative, while the decision tree-based classifier would not introduce decision rules based on it. Similarly, the support vector classifier would only produce boundaries orthogonal to an uninformative feature. The same cannot be said about a standard self-organising map: in this case, the effect of an uninformative feature is to spread the classification targets over a larger number of cells, thus reducing the sensitivity.

In order to understand which features are most relevant for classification, which is known as the 'saliency' in the machine learning literature, in Fig. 8 we show the mean gradients of the network outputs with respect to the input features. We see that the most important feature turns out to be the H_E magnitude, followed by the ground-based colours. The dependence on the optical colours and in particular on $(I_E - Y_E)$ in the *Euclid* photometry configuration has two main reasons. First, the optical bands retain low redshift information (see following discussion); second, the correlation between the I_E and $f_{\rm H\alpha+[N\,II]}$ is even stronger than the correlation of the emission line flux and H_E. The network uses $(I_E - Y_E)$ to extract I_E from the pivot magnitude H_E and infer this correlation. Lastly, as expected, the morphological parameters are the least important inputs for the neural network.

Having identified the H_E magnitude as the most informative feature, we can gain additional intuition about the classifiers by comparing the number counts $N(H_E)$ of the true targets to those of the samples recovered by the neural network. These are shown in Fig. 9. The green histogram gives the number counts for the true targets, i.e., the reference distribution we are trying to reproduce with the classifier. Notably, the counts go to zero for $H_E > 22.5$, hence there are no target galaxies fainter than this magnitude. This explains the rapid gain in precision one obtains by simply cutting the full sample (here shown by the orange histogram) at brighter and brighter values of H_E (see Table 2). Looking at the other histograms, we see that the application of the neural network effectively cuts the distribution down to the correct H_E. When using only *Euclid* bands (blue histogram), this leaves an excess of sources, which are either outside the redshift range or below the chosen Hα + [N II] flux limit, which are significantly reduced by adding the ground-based information (magenta dashed histogram). Note also how a selection over the target redshift range $[0.9, 1.8]$ using photometric redshifts clearly does not effectively cut on the H_E magnitude, leaving a large population of faint objects. Nevertheless, we remind the reader that this discussion is specific to Flagship2. In the case of EL-COSMOS, also galaxies fainter than $H_E \simeq 22.5$ are part of the target sample (see Fig. 4 and Table 2).

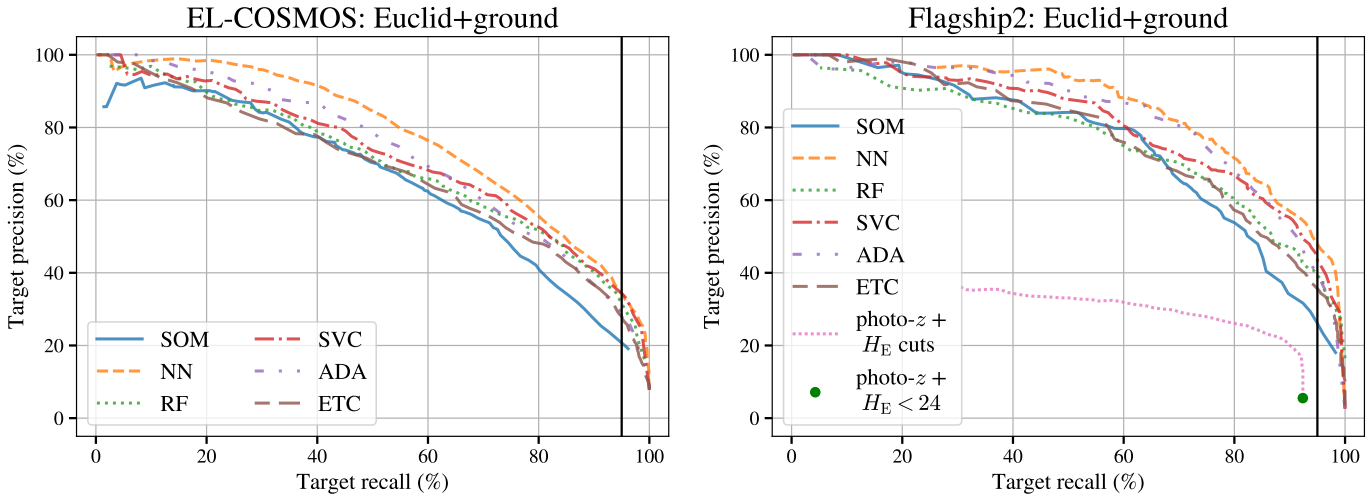


Fig. 7. Precision versus recall curves for the analyses with *Euclid* photometry and ground-based photometry. *Left*: results for EL-COSMOS. *Right*: results for Flagship2. The dotted magenta line represents the photo-z selection for a range of H_E magnitude limits. The green marker indicates the photo-z selection with $H_E < 24$.

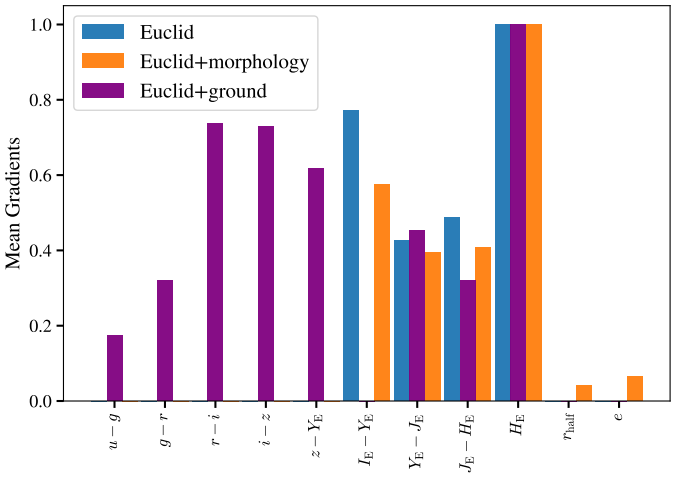


Fig. 8. Mean gradients of the neural network output as a function of the input for the three training configurations. In blue, orange, and purple are respectively plotted the mean gradients of the neural networks trained with *Euclid* photometry, *Euclid* photometry and morphology, and *Euclid* and ground-based photometry. All gradients have been normalised to that corresponding to the *Euclid* H_E magnitude.

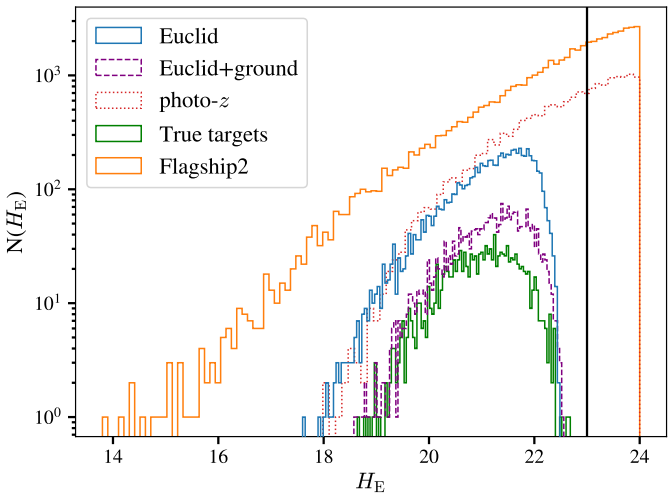


Fig. 9. H_E-band number counts for samples built from the Flagship2 catalogue. The samples selected with the neural network classifier, using *Euclid* photometry only or combined with ground-based photometry are shown, respectively, by the blue-solid and magenta-dashed histograms. As indicated by the legend, the red-dotted histogram corresponds to a sample selected in redshift only, using *Euclid* photometric redshifts. The counts for the full Flagship2 catalogue and the true target sample are also shown for reference, by the orange and green histograms. Note how the distribution of the true targets (green histogram) dies off at magnitudes fainter than $H_E \simeq 22.5$. The targets in the EL-COSMOS catalogue extend to fainter flux.

We can use the false positive rate (see Sect. 2.1) to interpret the origin of misclassified galaxies as a function of redshift and emission-line flux. In the top panel of Fig. 10 this quantity is plotted as a function of redshift. The sample produced using *Euclid* photometry alone shows an excess of false positives at $z < 1$. This explicitly shows the inability with only the *Euclid* bands to properly exclude low-redshift galaxies, as well as some with flux below the flux limit. The addition of ground-based photometry effectively cures this, removing all galaxies at $z < 0.9$, leaving only a fraction of misidentified objects fainter than 2×10^{-16} erg s^{-1} cm^{-2} inside the target redshift range. It is interesting to note that the *Euclid*-only and the *Euclid* plus ground curves become indistinguishable at $z \gtrsim 1.4$. This is consistent with the redshift at which the 4000 Å break enters the Y_E band (at 9600 Å) and indicates that in this range the combination of I_E and Y_E, J_E, and H_E provides, in general, sufficient spectral leverage to break degeneracies to both capture the correct redshift and identify emission-line targets. As also shown, a photometric redshift selection is effective at removing low-redshift galaxies, but keeps in the sample all the low-flux galaxies (as is expected, since we are selecting on redshift alone).

In Fig. 10 bottom panel, instead, we plot the false positive rate as a function of $f_{H\alpha+[N\,II]}$. In this case, the peak and discontinuity evident at the flux limit, 2×10^{-16} erg s^{-1} cm^{-2}, is due to sources just below the flux limit, which enter the sample as false positives. Above the flux limit, instead, false positives arise from galaxies that are outside the redshift range. For this reason, the

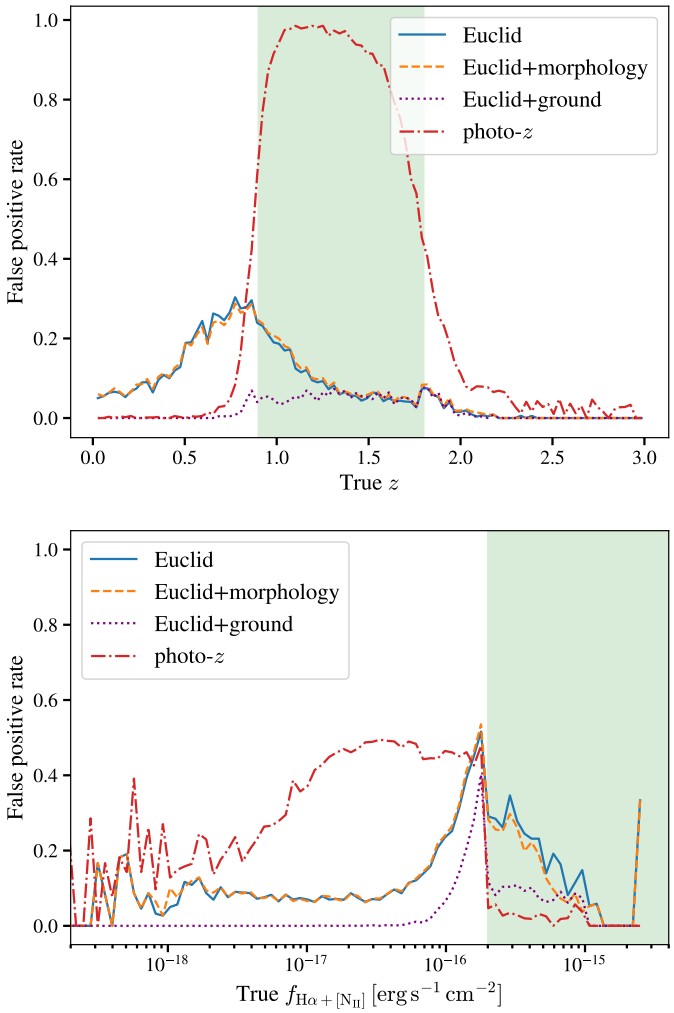


Fig. 10. False positive rate as function of redshift and $f_{H\alpha+[N\,II]}$. The plot allows us to identify the origin of non-targets that enter the selected samples. The solid blue, dashed orange, and dotted purple curves, respectively, correspond to the neural networks trained with *Euclid* photometry, *Euclid* plus morphological data, and *Euclid* plus ground-based photometry. The dash-dotted red line is the false positive rate of the photo-z selection. *Top*: False positive rate as a function of z. The green shaded area marks the target redshift range. *Bottom*: False positive rate as a function of $f_{H\alpha+[N\,II]}$. The green shaded area corresponds to the Hα limiting flux. There is a peak in the false positive rate just below the flux limit used to define the target sample, although we note that these galaxies can still give correct redshift measurements.

photo-z selection gives the lowest false positive rate, followed by the *Euclid* and ground-based classification. This does not tell the full story, however. The photo-z selection includes a number of false positives entering the sample at low fluxes, which are the cause of the very low precision shown by this selection. The classifier trained with ground-based photometry provides the best solution by balancing the two conditions of removing objects below the line flux limit and outside the redshift range.

Complementarily, it is also interesting to look at the true negative rate (Eq. 6) of the whole selected sample, which gives an insight into the fraction of non-targets removed from the sample. When we select galaxies using *Euclid* photometry only, the true negative rate is 87%; the combination with ground-based data increases this metric up to 97%. Conversely, the true negative rate of the photo-z selection is 59%. Again, the better performance of the classifiers in comparison to the photo-z selection reflects the fact that the latter does not make a selection in the emission-line limiting flux.

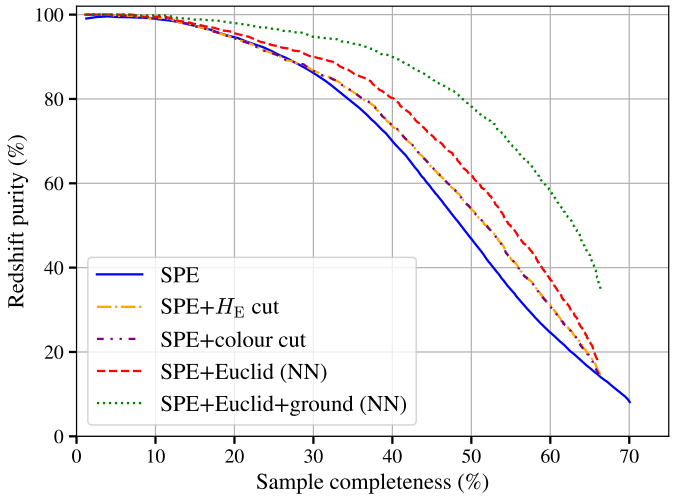
Finally, the machine learning algorithms identify regions in the full colour-magnitude space with a higher density of targets. In the case of the classifier trained on *Euclid* photometry, this is a four-dimensional space. In Appendix C we present slices through the four-dimensional probability maps constructed from each classifier, showing how the selection depends on colour. It is interesting to visualise the boundaries constructed by each classifier. There is no visible separation between target and non-target galaxies in the colour planes and the classification algorithms define complex boundaries in the four-dimensional space. The support vector classifier and the neural network produce particularly smooth boundaries, while the self-organising map and tree-based classifiers do not. The irregular boundary is an indication that the classifier is over-fitting the training set and will not generalise well. In addition, we verified that the 5% of the targets that we lose by imposing the 95% recall value are uniformly distributed in colour and are not part of any particular object class. We note that the lost targets are mainly faint objects.

5. Purity and completeness

The final purity of the spectroscopic sample will depend on the combination of the photometric information with the selection criteria applied to the spectroscopic measurements, as described by the flow diagram of Fig. 1. To provide a concrete, yet preliminary, example, we would like to quantify here the improvement in the final redshift purity and sample completeness produced by our photometric selection process. This work is based on a set of simulated spectra that were processed by the *Euclid* spectroscopic measurement pipeline (the SPE processing function). Although the simulated data were not yet fully realistic, they are nevertheless very useful for understanding how a machine learning-based photometric classification can aid in the sample selection. Also, the simulated spectra were built from the EL-COSMOS sample described in Sect. 3.1.1, which helps in making this test self-consistent. Two-dimensional spectral images were generated using the `FastSpec` code based on the spectral energy distribution and morphological parameters of the galaxies. These images were convolved with the NISP instrumental point spread function and realistic noise was added according to the detector model. Multiple exposures were simulated for each source and stacked with one to four exposures. One-dimensional spectra were extracted from the images and input to the *Euclid* spectroscopic measurement processing function to measure the redshift and spectral features.

The spectroscopic measurement pipeline carries out a likelihood analysis using spectral templates to estimate the redshift. It produces a probability distribution function of the redshift that is typically sharply peaked with a few primary redshift solutions. The integral of the peak provides a useful measure of the reliability of the solution. We vary the threshold in this reliability value to select spectroscopic samples and build the relationship between redshift purity and completeness, as shown by the SPE solid blue line in Fig. 11.

In the following discussion we focus on the results from the neural network classifier applied to the simulated spectroscopic sample. The purity and completeness values should be taken as indicative of the general trends and not as accurate forecasts of the pipeline performance. The values depend on the specific distribution of simulated sources and instrumental configuration.



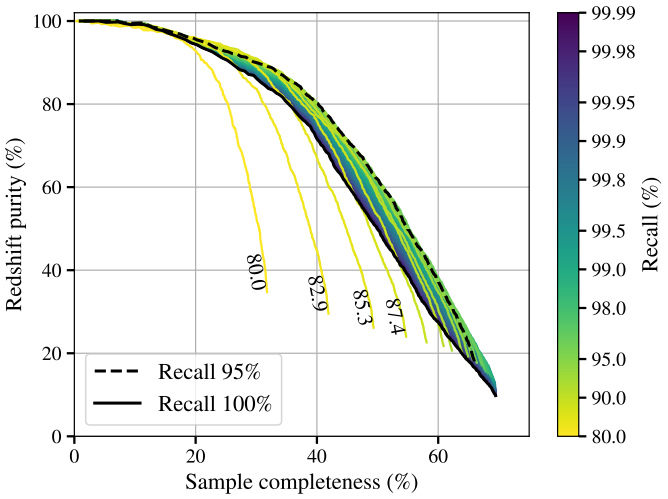


Fig. 11. Redshift purity and sample completeness as a function of spectroscopic reliability threshold. The solid blue, dashed red, dotted green, and dash-dotted orange lines respectively correspond to a selection using only SPE reliability, SPE reliability combined with a photometric classification based on *Euclid* data, with the classification that uses *Euclid* and ground-based photometry, with the H_E magnitude limit selection, and with the colour selection in the $(I_E - H_E)$-H_E plane . In all cases the recall of the photometric classification is set to 95%.

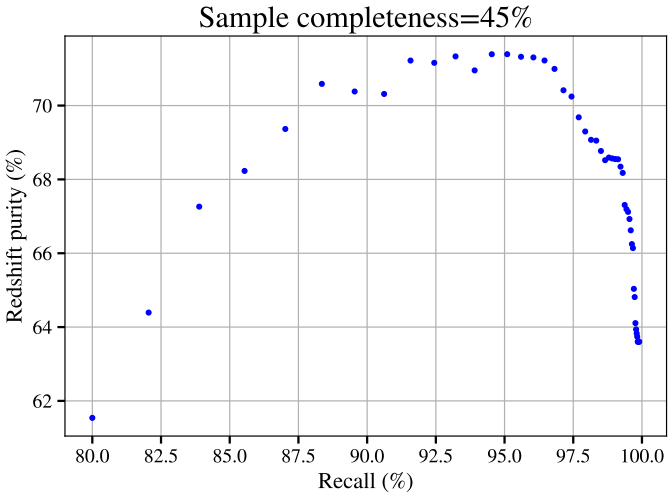


Fig. 12. Spectroscopic redshift purity and completeness with the addition of the photometric classification. *Top*: the curves are colour-coded as a function of the recall of the photometric classification. The spectroscopic reliability threshold varies along each curve, while varying the threshold on the photometric classification probability shifts the curve. The purity improves as recall increases, reaching a maximum for recall ∼ 95% and declining after. For better visualisation, the first lines are labelled with the corresponding recall value. At recall values above 95% the curves are tightly packed. The solid black line corresponds to 100% recall, while the dashed line to 95% recall, the value we chose to benchmark our results. *Bottom*: redshift purity as a function of the recall of the photometric classification, fixing the value of sample completeness to 45%.

The target sample is defined as described in Sect. 3.3, using the total flux of the Hα and N II complex.

Figure 11 shows how redshift purity versus sample completeness plot improves when we complement the pure spectroscopic reliability cut selection (blue solid line) with increasing information provided by the photometric neural network classifier for the two configurations using *Euclid*-only or *Euclid* and ground-based photometry. The curve corresponding to the H_E magnitude-limit selection that gives 95% recall (see Table 2) is also plotted together with the curve corresponding to the colour selection presented in Sect. 4.1. These two curves visually overlap, but the colour cut curve (dash-dot-dotted purple line) is marginally higher than the simple magnitude cut curve (dash-dotted orange line). The figure shows that in the range between 40% and 60% completeness, the photometric classification improves the redshift purity. For example, at a fixed value of 45% sample completeness, the classification based on *Euclid*-only bands improves the purity by ∼ 20%, when we add ground-based photometry the improvement rises to ∼ 45%. The simple H_E magnitude limit selection, at that same completeness value, gives an improvement of a few per cent only ($\lesssim 10\%$), evidencing the importance of exploiting all available photometric information.

To examine the effect of the photometric classification in more detail, in the top panel of Fig. 12 we show the redshift purity and sample completeness as a function of the reliability threshold imposed on the spectroscopic redshift measurement. The photometric classification has its own threshold parameter on the classification probability, which when combined with the spectroscopic selection, produces a family of curves. We label these curves based on their recall values. The bottom panel shows the dependence of redshift purity on the photometric selection recall, when the completeness is fixed to 45%. As we see, a recall value of 95% approximately maximises the purity-completeness curve, which justifies the choice made in Sect. 2.1.

The main conclusion from this exercise is that the impact of properly elaborated photometric information on the final purity and completeness of the *Euclid* spectroscopic sample is very significant, with a major improvement especially when ground-based visible bands are included. The precise gain, however, will depend on the galaxy distribution, the survey configuration and the instrument model.

6. Conclusions

We have investigated the benefits of combining photometric information with the spectroscopic measurement criteria for selecting *Euclid* spectroscopic samples. *Euclid* spectroscopy will give estimates of the galaxy redshifts, fluxes of the emission

lines, and confidence intervals. However, since emission-line galaxies make up only a small fraction of the photometric sample, measurement noise can reduce the redshift purity and completeness of the sample and degrade the figure of merit for the galaxy clustering probe. The addition of photometric criteria in the selection can allow us to improve the purity of the sample by identifying sources that are likely to be bright emission-line galaxies at the target redshift.

To this end, we compared a set of machine learning classification algorithms with the aim of photometrically selecting emission-line target galaxies that are likely to give good redshift measurements in the Euclid Wide Survey. We used two catalogues to benchmark the classification performance, EL-COSMOS and Flagship2. Both catalogues have *Euclid* and ground-based simulated photometry. We produced noisy realisations of the catalogues assuming background-limited observations. The two catalogues yield similar results when using as input *Euclid*-only photometry, but when this is combined with ground-based data, the results using Flagship2 outperform those with EL-COSMOS. This is related to the differences in the Hα luminosity function and colour distribution of the two catalogues. In addition to these two configurations (*Euclid*-only and *Euclid* plus ground) we also considered adding morphological information (half-light radius and the axial ratio). We find that in general, while the addition of ground-based data strongly improves the precision (doubling it in the case of Flagship2), including morphological information (at least in the form provided here) gives negligible improvement.

The purity of the final spectroscopic sample will depend on the combination of the photometric classification with further selection criteria based on the properties of the spectroscopic data (see diagram in Fig. 1). To investigate this requires full end-to-end simulations of the spectroscopic reduction pipeline. We presented a preliminary exercise to assess the relative gain when the spectroscopic data are complemented by the photometric selection discussed here. This will be expanded in future work. We showed that in the range between 40% and 60% completeness the purity is boosted by $\sim$20% when using *Euclid*-only bands, and between 40% and 100% when including ground-based photometry. We consider this a remarkable indication.

The introduction of ground-based data significantly improves the purity of the sample, but in the practical application can also bring additional nuisance in the form of systematic errors. The ground-based photometry will come from multiple surveys and so will not be fully homogeneous. It will also suffer from additional selection effects correlated with the observing conditions that can propagate as systematic errors to the galaxy clustering measurements and cosmological constraints. Thus, the gains in purity from incorporating ground-based data must be carefully weighed against the potential of adding systematic errors, also considering the specific requirements of the science analysis to be carried out. We foresee that ground-based data may be used in analyses where a higher level of purity is desired, such as for studying the galaxy halo occupation distribution or galaxy evolution as a function of environment.

Photometric redshifts can also play a key role in sample selection. We used the *Euclid* photometric redshift estimates to select galaxies in the target redshift range and compared the performance of such a selection to that of the colour-based machine learning classifiers. Figure 10 shows that the photo-z selection is very efficient for redshift classification, especially to remove low redshift interlopers, but is not effective in identifying emission-line galaxies. Indeed, the photo-z selection has the highest fraction of false positives from faint galaxies with $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]} < 2\times10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, but the lowest for bright ones with $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]} > 2\times10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, which means that it makes a better redshift selection than the algorithms presented in this work. Photometric redshifts could be used with additional constraints from spectral energy distribution fits to identify bright emission-line galaxy targets. In particular, the *Euclid* photometric redshift pipeline will output estimates of galaxy physical properties including the star-formation rate and dust attenuation, which will allow us to select emission-line galaxy samples. We expect that a classifier developed based on photometric redshifts and estimates of physical properties from spectral energy distribution fitting would perform similarly to the pure colour and magnitude-based classifiers that we tested, since the underlying photometric information is the same. Analogously, we expect a classifier trained to make a selection in redshift alone to perform similarly to the photo-z selection. Alternative classifiers that use the estimates of galaxy physical properties from the *Euclid* photometric redshift pipeline for sample selection will be investigated in a future work.

It is important to note that in this study some of the complications that will be present in real *Euclid* data were not considered. First, we assume an ideal training set, which is fully representative of the Wide Survey data. In the actual Euclid Wide Survey, the training set will come from the Deep Fields, which will total $\sim 50\,\mathrm{deg}^2$. Shallow and full-depth photometric measurements will be available for the *Euclid* photometry in the Deep Fields; however, we will only have the full-depth measurements for the ground-based photometry. As they are currently trained, the machine learning algorithms learn to classify the targets at a given noise level and it is not necessarily true that they will be able to generalise their results when trained and tested on samples with different noise levels. Therefore, if ground-based photometry is used, it will be necessary to degrade the measurements to match the noise level in the Wide Survey. Since the ground-based photometry will come from multiple surveys, this operation will not be simple, and residual variations in homogeneity in the noise can lead to systematic variations in the classifier performance.

Moreover, the effective emission-line flux limit will vary across the Wide Survey due to foreground emission including zodiacal light and scattered stellar light (Euclid Collaboration: Scaramella et al. 2022). In this study, we used a fixed flux limit to build the training set of emission-line galaxies. In practice, this does not impose a sharp flux cut in the measured sample. However, when developing a classifier on real data, we will be able to use the Deep Survey to define the training set as the set of galaxies that are correctly measured by the *Euclid* pipeline, without imposing any specific constraints on their physical properties. It will also be possible to construct a classifier that accounts for variations in the noise level across the Wide Survey to optimise the sample.

Finally, a further complication that must be considered is contamination from stars in the galaxy catalogue that can impact the purity. The photometric classifier can be trained to maximise the precision in the presence of stars. This work will require us to incorporate a star-galaxy classifier, which is based both on size and photometric colours. Here, the morphological measurements will be important.

In the next stage of this work, we will consider the full set of spectroscopic and photometric selection criteria in order to compute the redshift purity, sample completeness and ultimately cosmologically relevant figures of merit. This requires running the spectroscopic reduction pipeline on mock data in order to produce end-to-end simulations. Such simulations will allow us to optimise the sample selection criteria, possibly with the use

of machine learning classifiers. With *Euclid* observations beginning in fall 2023, we will be able to further tune the selection based on the actual telescope performance and ultimately construct the spectroscopic galaxy sample that will be used to test the cosmological model.

Acknowledgements. The Euclid Consortium acknowledges the European Space Agency and a number of agencies and institutes that have supported the development of *Euclid*, in particular the Agenzia Spaziale Italiana, the Belgian Science Policy, the Canadian Euclid Consortium, the French Centre National d'Etudes Spatiales, the Deutsches Zentrum für Luft- und Raumfahrt, the Danish Space Research Institute, the Fundação para a Ciência e a Tecnologia, the Hungarian Academy of Sciences, the Ministerio de Ciencia, Innovación y Universidades, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Research Council of Finland, the Romanian Space Agency, the State Secretariat for Education, Research and Innovation (SERI) at the Swiss Space Office (SSO), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* web site (`http://www.euclid-ec.org`). This work has made use of CosmoHub (Carretero et al. 2017; Tallada et al. 2020). CosmoHub has been developed by the Port d'Informació Científica (PIC), maintained through a collaboration of the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT) and the Institute of Space Sciences (CSIC & IEEC), and was partially funded by the "Plan Estatal de Investigación Científica y Técnica y de Innovación" program of the Spanish government.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org
Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2019, ApJ, 879, 15
Bauer, E. & Kohavi, R. 1999, Mach. Learn., 36, 105
Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992, in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92 (New York, NY, USA: Association for Computing Machinery), 144–152
Breiman, L. 2001, Mach. Learn., 45, 5
Carretero, J., Tallada, P., Casals, J., et al. 2017, in Proceedings of the European Physical Society Conference on High Energy Physics. 5-12 July, 488
Comparat, J., Delubac, T., Jouvel, S., et al. 2016, A&A, 592, A121
Cropper, M., Pottinger, S., Niemi, S., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9904, Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave, ed. H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, & E. C. Tong, 99040Q
Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, AJ, 122, 2267
Euclid Collaboration: Blanchard, A., Camera, S., Carbone, C., et al. 2020, A&A, 642, A191
Euclid Collaboration: Desprez, G., Paltani, S., Coupon, J., et al. 2020, A&A, 644, A31
Euclid Collaboration: Gabarra, L., Mancini, C., Rodriguez Munoz, L., et al. 2023, arXiv:2302.09372
Euclid Collaboration: Pocino, A., Tutusaus, I., Castander, F. J., et al. 2021, A&A, 655, A44
Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, A&A, 662, A112
Euclid Collaboration: Schirmer, M., Jahnke, K., Seidel, G., et al. 2022, A&A, 662, A92
Flaugher, B. 2005, International Journal of Modern Physics A, 20, 3121
Freund, Y. & Schapire, R. E. 1995, in Computational Learning Theory, ed. P. Vitányi (Berlin, Heidelberg: Springer Berlin Heidelberg), 23–37
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Mach. Learn., 63, 3
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, A&A, 566, A108
Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning, Springer Series in Statistics (New York, NY, USA: Springer New York Inc.)
Ioffe, S. & Szegedy, C. 2015, arXiv:1502.03167
Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. 2017, arXiv:1706.02515
Kohonen, T. 1982, Biological Cybernetics, 43, 59
Kohonen, T. 1990, Proceedings of the IEEE, 78, 1464
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, arXiv:1708.02002
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Maciaszek, T., Ealet, A., Gillard, W., et al. 2022, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 12180, Space Telescopes and Instrumentation 2022: Optical, Infrared, and Millimeter Wave, ed. L. E. Coyle, S. Matsuura, & M. D. Perrin, 121801K
Moosavi, V., Packmann, S., & Vallés, I. 2014, SOMPY: A Python Library for Self Organizing Map (SOM), gitHub.[Online]. Available: https://github.com/sevamoo/SOMPY
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 12
Pozzetti, L., Hirata, C. M., Geach, J. E., et al. 2016, A&A, 590, A3
Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, AJ, 165, 126
Ross, A. J., Ho, S., Cuesta, A. J., et al. 2011, MNRAS, 417, 1350
Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121
Saito, S., de la Torre, S., Ilbert, O., et al. 2020, MNRAS, 494, 199
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Stanford, S. A., Masters, D., Darvish, B., et al. 2021, ApJS, 256, 9
Tallada, P., Carretero, J., Casals, J., et al. 2020, Astronomy and Computing, 32, 100391
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, Publications of the Astronomical Society of Japan, 70, S9
Weaver, J. R., Kauffmann, O., Shuntov, M., et al. 2021, in American Astronomical Society Meeting Abstracts, Vol. 53, American Astronomical Society Meeting Abstracts, 215.06

[1] Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[2] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
[3] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[4] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 67000 Strasbourg, France
[5] Aix-Marseille Université, CNRS, CNES, LAM, Marseille, France
[6] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[7] Dipartimento di Fisica - Sezione di Astronomia, Università di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[8] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy
[9] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste TS, Italy
[10] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy
[11] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, 40129 Bologna, Italy
[12] Centre for Astrophysics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[13] Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[14] Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada
[15] Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St SE, Minneapolis, MN 55455, USA
[16] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[17] School of Mathematics and Physics, University of Surrey, Guildford, Surrey, GU2 7XH, UK
[18] Dipartimento di Fisica e Astronomia, Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[19] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[20] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[21] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany
[22] Dipartimento di Fisica, Università di Genova, Via Dodecaneso 33, 16146, Genova, Italy
[23] INFN-Sezione di Genova, Via Dodecaneso 33, 16146, Genova, Italy

[24] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126, Napoli, Italy
[25] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[26] INFN section of Naples, Via Cinthia 6, 80126, Napoli, Italy
[27] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[28] Dipartimento di Fisica, Università degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
[29] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
[30] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[31] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
[32] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[33] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
[34] Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, 52056 Aachen, Germany
[35] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
[36] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[37] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[38] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[39] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
[40] ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[41] University of Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
[42] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[43] UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, 4 rue Enrico Fermi, 69622 Villeurbanne, France
[44] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisboa, Portugal
[45] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[46] Department of Astronomy, University of Geneva, ch. d'Ecogia 16, 1290 Versoix, Switzerland
[47] INAF-Istituto di Astrofisica e Planetologia Spaziali, via del Fosso del Cavaliere, 100, 00100 Roma, Italy
[48] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[49] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[50] Univ Claude Bernard Lyon 1, CNRS, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
[51] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France
[52] School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, UK
[53] Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy
[54] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, 35122 Padova, Italy
[55] University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 Munich, Germany
[56] Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, 0315 Oslo, Norway
[57] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[58] Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK
[59] von Hoerner & Sulger GmbH, SchloßPlatz 8, 68723 Schwetzingen, Germany
[60] Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark
[61] Cosmic Dawn Center (DAWN), Denmark
[62] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[63] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, 00014 University of Helsinki, Finland
[64] Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille, France
[65] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA
[66] AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, 91191 Gif-sur-Yvette, France
[67] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland
[68] Department of Physics, P.O. Box 64, 00014 University of Helsinki, Finland
[69] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[70] NOVA optical infrared instrumentation group at ASTRON, Oude Hoogeveensedijk 4, 7991PD, Dwingeloo, The Netherlands
[71] Universität Bonn, Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn, Germany
[72] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, DH1 3LE, UK
[73] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice cedex 4, France
[74] Institut d'Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France
[75] Université Paris Cité, CNRS, Astroparticule et Cosmologie, 75013 Paris, France
[76] Institut d'Astrophysique de Paris, 98bis Boulevard Arago, 75014, Paris, France
[77] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
[78] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, DK-8000 Aarhus C, Denmark
[79] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France
[80] Space Science Data Center, Italian Space Agency, via del Politecnico snc, 00133 Roma, Italy
[81] Centre National d'Etudes Spatiales – Centre spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[82] Institute of Space Science, Str. Atomistilor, nr. 409 Măgurele, Ilfov, 077125, Romania
[83] Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, 38204, San Cristóbal de La Laguna, Tenerife, Spain
[84] Departamento de Astrofísica, Universidad de La Laguna, 38206, La Laguna, Tenerife, Spain
[85] Dipartimento di Fisica e Astronomia "G. Galilei", Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[86] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
[87] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain
[88] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
[89] Satlantis, University Science Park, Sede Bld 48940, Leioa-Bilbao, Spain
[90] Centre for Electronic Imaging, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[91] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal

[92] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, Plaza del Hospital 1, 30202 Cartagena, Spain
[93] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[94] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
[95] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[96] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
[97] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
[98] Junia, EPA department, 41 Bd Vauban, 59800 Lille, France
[99] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste TS, Italy
[100] ICSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing, Via Magnanelli 2, Bologna, Italy

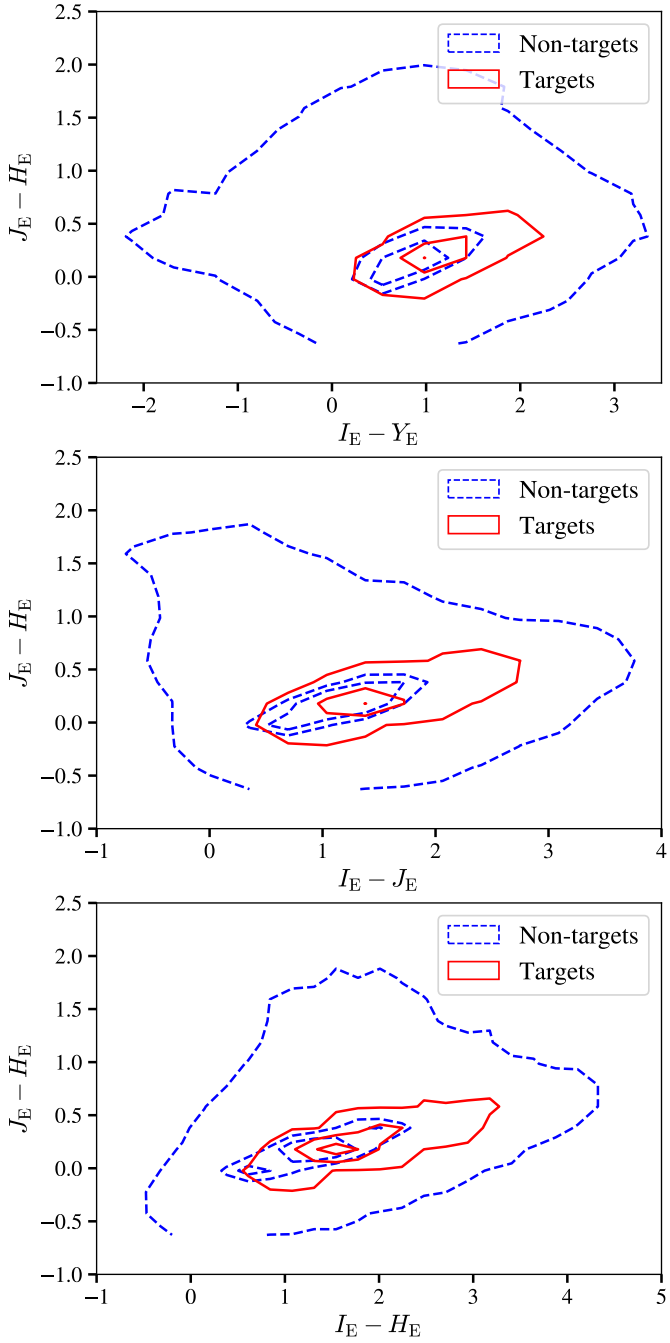


Fig. A.1. Target and non-target distributions in colour-colour and colour-magnitudes planes for the Flagship2 catalogue. The contours contain 99%, 50%, and 25% of the samples.

Appendix A: Colour-magnitude projection planes

We present in Fig. A.1 the colour-colour and colour-magnitude distributions for targets and non-targets in the Flagship2 catalogue for three different combinations. The contours contain 99%, 50%, and 25% of the samples. There is a nearly complete overlap of the targets and non-targets in the colours.

Appendix B: Photo-z as input variables

As an additional test, we trained a neural network with Flagship2 data in the *Euclid* plus ground-based configuration with the additional information of the measured photo-z. In principle, the neural network can extrapolate the redshift from the photometric information. However, by directly providing the photo-z we may facilitate the selection process as the network will be able to put more attention on the emission line flux.

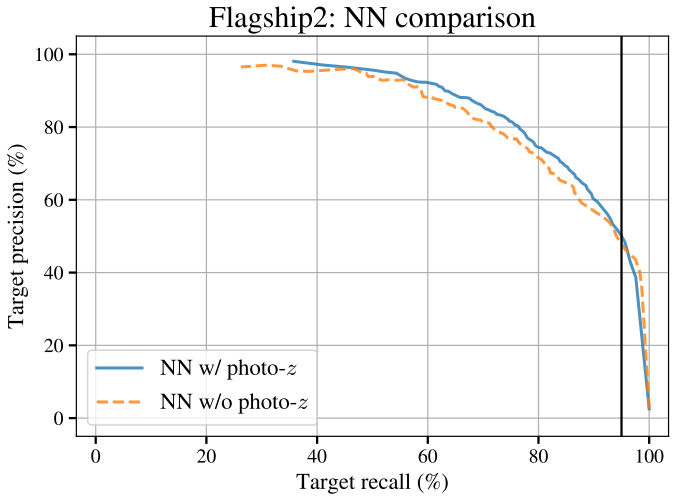


Fig. B.1. Comparison of the precision versus recall curves of two neural networks trained with and without photo-zs as an input feature. The two neural networks were trained with *Euclid* and ground-based photometry, but in the case of the solid blue line the algorithm takes the photo-z of the galaxy as an additional feature. The solid vertical line corresponds to 95% recall.

The precision at 95% recall is 47.9% in the case without photo-z, as reported in Table 3. When we add the photo-z of the galaxy as an input feature the precision rises to 50.1%. Nevertheless, the addition of the photo-z to the input information makes the classifier dependent on the complex process used to produce the photo-zs, including the training sets, spectral energy distribution models, and algorithms used. We postpone to a future work the detailed study of these dependencies and the performance on realistic data.

Appendix C: Selection probability maps

Figure C.1 gives a visualisation of the selection probability for each classifier in planes through the parameter space. We show the results from the Flagship2 catalogue for the case when classifiers are trained with *Euclid* photometry alone. Each row shows the colour-colour and colour-magnitude plots for a given algorithm. The parameter space is four-dimensional, and the two-dimensional planes are made by fixing two of the parameters to their median values.

One notices that the different classifiers identify a similar region of maximum probability for a given pair of features. The shape and the gradients of these regions, however, vary for each algorithm. This is due to the differences in the selection algorithms and possible projection effect when the boundaries are represented on the planes. In the case of the single classifiers (top three rows: self-organising map, neural network, and support vector classifier) they are compact and well defined, unlike for the cases of voting classifiers based on decision trees (bottom three rows). Also, the contours and gradients are less smooth for the self-organising map than for the neural network and the support vector classifier. The probability gradient of the support vector classifier is very steep, especially in comparison to the neural network.

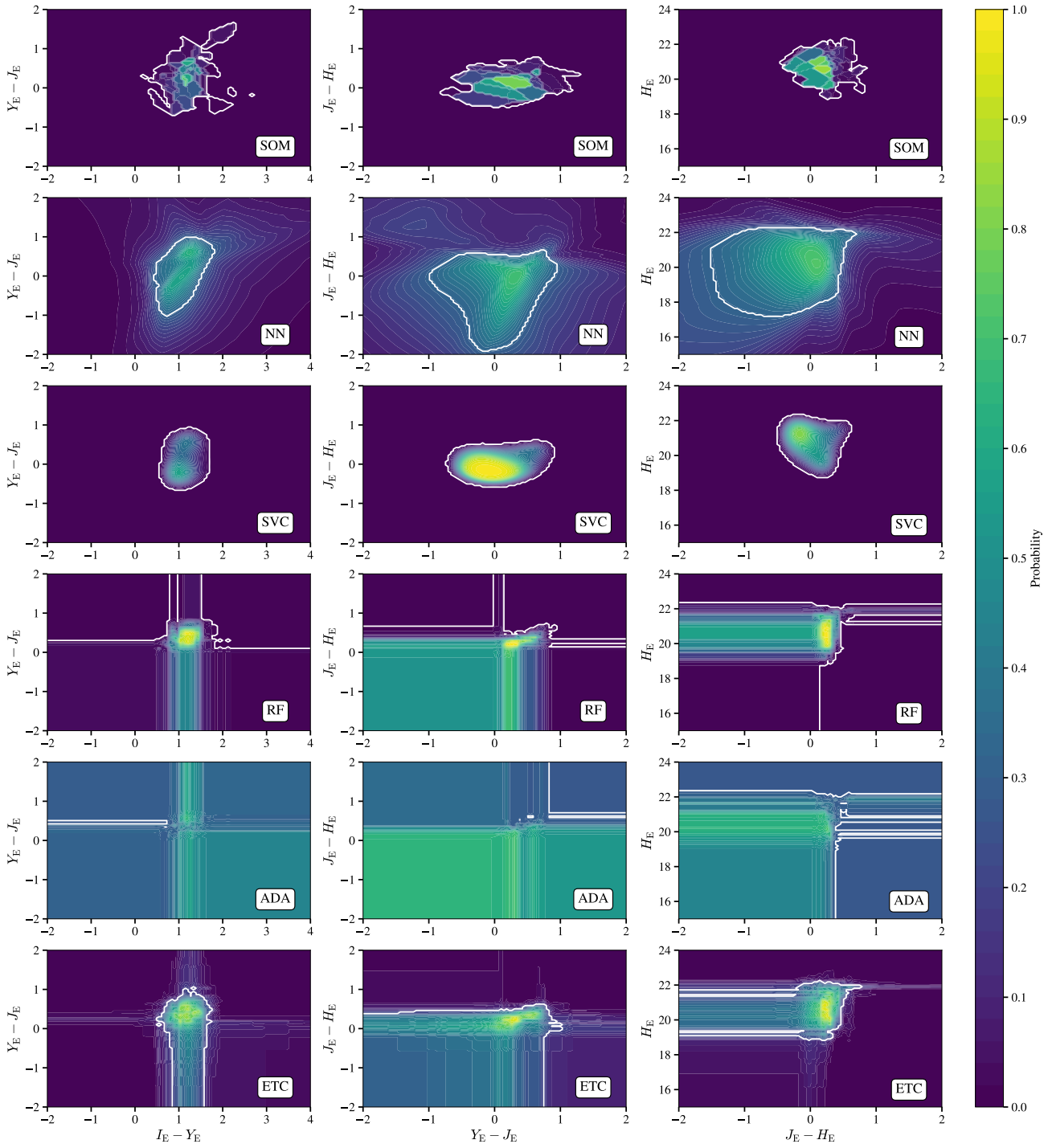


Fig. C.1. Probability maps in colour-colour and colour-magnitude planes, for the six classifiers tested in this paper, trained using Flagship2 *Euclid* photometry only. The thick white contour marks the probability threshold that gives 95% recall.

The probability maps for the voting classifiers (bottom three rows) show orthogonal contours. This is due to the common base classifier of these algorithms, the decision tree, which tends to produce decision rules orthogonal to one another. At the same time, the three algorithms have very different probability contours. These differences are related to the batch selection rule used to train the decision trees (see Sect. 2). We expect that the algorithms that give a classification model with irregular and steep contours (such as the self-organising map) or stepped contours (such as the decision trees) will be prone to over-fitting and will show poorer performance than algorithms that give smooth probability contours.