Astronomy & Astrophysics manuscript no. Euclid_preparation_XXII_Humphrey_etal
September 28, 2022

Euclid preparation: XXII. Selection of Quiescent Galaxies from Mock Photometry using Machine Learning

Euclid Collaboration: A. Humphrey[1,2⋆], L. Bisigello[3], P. A. C. Cunha[1,4], M. Bolzonella[3], S. Fotopoulou[5], K. Caputi[6], C. Tortora[7], G. Zamorani[3], P. Papaderos[8,9], D. Vergani[3], J. Brinchmann[1], M. Moresco[10,3], A. Amara[11], N. Auricchio[3], M. Baldi[10,3,12], R. Bender[13,14], D. Bonino[15], E. Branchini[16,17], M. Brescia[7], S. Camera[18,15,19], V. Capobianco[15], C. Carbone[20], J. Carretero[21,22], F.J. Castander[23,24], M. Castellano[25], S. Cavuoti[26,7,27], A. Cimatti[28,29], R. Cledassou[30,31], G. Congedo[32], C.J. Conselice[33], L. Conversi[34,35], Y. Copin[36], L. Corcione[15], F. Courbin[37], M. Cropper[38], A. Da Silva[8,39], H. Degaudenzi[40], M. Douspis[41], F. Dubath[40], C.A.J. Duncan[42], X. Dupac[35], S. Dusini[43], S. Farrens[44], S. Ferriol[36], M. Frailis[45], E. Franceschi[3], M. Fumana[20], P. Gómez-Alvarez[35,46], S. Galeotta[45], B. Garilli[20], W. Gillard[47], B. Gillis[32], C. Giocoli[48,49], A. Grazian[50], F. Grupp[13,14], L. Guzzo[51,52,53], S.V.H. Haugan[54], W. Holmes[55], F. Hormuth[56], K. Jahnke[57], M. Kümmel[14], S. Kermiche[47], A. Kiessling[55], M. Kilbinger[44], T. Kitching[38], R. Kohley[35], M. Kunz[58], H. Kurki-Suonio[59], S. Ligori[15], P. B. Lilje[54], I. Lloro[60], E. Maiorano[3], O. Mansutti[45], O. Marggraf[61], K. Markovic[55], F. Marulli[28,3,62], R. Massey[63], S. Maurogordato[64], H.J. McCracken[65,66], E. Medinaceli[48], M. Melchior[67], M. Meneghetti[3,12], E. Merlin[25], G. Meylan[37], L. Moscardini[62,10,3], E. Munari[45], R. Nakajima[61], S.M. Niemi[68], J. Nightingale[63], C. Padilla[21], S. Paltani[40], F. Pasian[45], K. Pedersen[69], V. Pettorino[70], S. Pires[44], M. Poncet[30], L. Popa[71], L. Pozzetti[3], F. Raison[13], A. Renzi[72,43], J. Rhodes[55], G. Riccio[7], E. Romelli[45], M. Roncarelli[10,3], E. Rossetti[10], R. Saglia[13,14], D. Sapone[73], B. Sartoris[74,45], R. Scaramella[25,75], P. Schneider[61], M. Scodeggio[20], A. Secroun[47], G. Seidel[57], C. Sirignano[72,43], G. Sirri[62], L. Stanco[43], P. Tallada-Crespí[76,22], D. Tavagnacco[45], A.N. Taylor[32], I. Tereno[8,9], R. Toledo-Moreo[77], F. Torradeflot[76,22], I. Tutusaus[58], L. Valenziano[3,62], T. Vassallo[14], Y. Wang[78], J. Weller[13,14], A. Zacchei[45], J. Zoubian[47], S. Andreon[52], S. Bardelli[3], A. Boucaud[79], R. Farinelli[80], J. Graciá-Carpio[13], D. Maino[51,20,53], N. Mauri[28,62], S. Mei[79], N. Morisset[40], F. Sureau[44], M. Tenti[62], A. Tramacere[40], E. Zucca[3], C. Baccigalupi[74,45,81,82], A. Balaguera-Antolínez[83,84], A. Biviano[74,45], A. Blanchard[85], S. Borgani[86,45,81,74], E. Bozzo[40], C. Burigana[87,12,88], R. Cabanac[85], A. Cappi[64,3], C.S. Carvalho[9], S. Casas[89], G. Castignani[10,3], C. Colodro-Conde[90], A.R. Cooray[91], J. Coupon[40], H.M. Courtois[92], O. Cucciati[3], S. Davini[93], G. De Lucia[45], H. Dole[41], J.A. Escartin[13], S. Escoffier[47], M. Fabricius[13], M. Farina[94], F. Finelli[3,12], K. Ganga[79], J. Garcia-Bellido[95], K. George[14], F. Giacomini[62], G. Gozaliasl[96], I. Hook[97], M. Huertas-Company[98,99,84,90], B. Joachimi[100], V. Kansal[44], A. Kashlinsky[101], E. Keihanen[96], C.C. Kirkpatrick[59], V. Lindholm[96,102], G. Mainetti[103], R. Maoli[104,25], S. Marcin[67], M. Martinelli[95], N. Martinet[105], M. Maturi[106,107], R. B. Metcalf[10,3], G. Morgante[3], A.A. Nucita[108,109], L. Patrizii[62], A. Peel[37], J.E. Pollack[79], V. Popa[71], C. Porciani[61], D. Potter[110], P. Reimberg[65], A.G. Sánchez[13], M. Schirmer[57], M. Schultheis[64], V. Scottez[65,111], E. Sefusatti[74,45,81], J. Stadel[110], R. Teyssier[112], C. Valieri[62], J. Valiviita[113,102], M. Viel[74,45,81,82], F. Calura[3], H. Hildebrandt[114]

(Affiliations can be found after the references)

Received June 20, 2022; accepted September 15, 2022

ABSTRACT

The *Euclid* Space Telescope will provide deep imaging at optical and near-infrared wavelengths, along with slitless near-infrared spectroscopy, across $\sim 15\,000\,\mathrm{deg}^2$ of the sky. *Euclid* is expected to detect ~ 12 billion astronomical sources, facilitating new insights into cosmology, galaxy evolution, and various other topics. In order to optimally exploit the expected very large data set, there is the need to develop appropriate methods and software tools. Here we present a novel machine-learning based methodology for the selection of quiescent galaxies using broad-band Euclid I_E, Y_E, J_E, H_E photometry, in combination with multiwavelength photometry from other large surveys (e.g. the Rubin LSST). The `ARIADNE` pipeline uses meta-learning to fuse decision-tree ensembles, nearest-neighbours, and deep-learning methods into a single classifier that yields significantly higher accuracy than any of the individual learning methods separately. The pipeline has been designed to have 'sparsity-awareness', such that missing photometry values are informative for the classification. In addition, our pipeline is able to derive photometric redshifts for galaxies selected as quiescent, aided by the 'pseudo-labelling' semi-supervised method, and using an outlier detection algorithm to identify and reject likely catastrophic outliers. After application of the outlier filter, our pipeline achieves a normalized mean absolute deviation of $\lesssim 0.03$ and a fraction of catastrophic outliers of $\lesssim 0.02$ when measured against the COSMOS2015 photometric redshifts. We apply our classification pipeline to mock galaxy photometry catalogues corresponding to three main scenarios: (i) Euclid Deep Survey photometry with ancillary *ugriz*, WISE, and radio data; (ii) Euclid Wide Survey photometry with ancillary *ugriz*, WISE, and radio data; (iii) Euclid Wide Survey photometry only, with no foreknowledge of galaxy redshifts. In a like-for-like comparison, our classification pipeline outperforms *UVJ* selection, in addition to the Euclid $I_\mathrm{E} - Y_\mathrm{E}$, $J_\mathrm{E} - H_\mathrm{E}$ and $u - I_\mathrm{E}$, $I_\mathrm{E} - J_\mathrm{E}$ colour-colour methods, with improvements in completeness and the F1-score (the harmonic mean of precision and recall) of up to a factor of 2.

Key words. Galaxies: photometry – Galaxies: high-redshift – Galaxies: evolution – Galaxies: general

1. Introduction

The study of galaxies plays a pivotal role in the effort to understand how the baryonic component of the Universe evolved

⋆ e-mail: Andrew.Humphrey@astro.up.pt

across cosmic time; it is on the size-scale of galaxies that key processes such as star-formation, chemical evolution, black hole growth, and feedback predominantly take place. Large, systematic imaging and spectroscopic surveys are among the most valuable resources for investigating galaxy evolution, allowing a wide range of studies ranging from identification and study of rare or elusive objects (e.g. Alexandroff et al. 2013), to statistical studies of large samples of galaxies (e.g. Kauffmann et al. 2003). A variety of ground- or space-based surveys has already provided rich databases for investigating questions surrounding the evolution of galaxies (e.g. the Sloan Digital Sky Survey: York et al. 2000; Gunn et al. 1998), with even more data to be generated by ongoing and future campaigns and facilities which will map the extragalactic sky to even fainter flux levels, and out to even higher redshifts [e.g. the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST): Ivezić et al. 2019; the Nancy Grace Roman Space Telescope (NGRST): Akeson et al. 2019; the Dark Energy Spectroscopic Instrument survey (DESI): Dey et al. 2019; the 4-metre Multi-Object Spectroscopic Telescope (4MOST); Guiglion et al. 2019; the Multi Object Optical and Near-infrared Spectrograph for the VLT (MOONS); Taylor et al. 2018; Cirasuolo et al. 2020; the Square Kilometer Array (SKA): Dewdney et al. 2009; the extended ROentgen Survey with an Imaging Telescope Array (eROSITA): Predehl et al. 2021].

In the next years, the *Euclid* Space Telescope will make a substantial contribution to our understanding of galaxy evolution. *Euclid* will observe $\sim 15\,000\,\mathrm{deg}^2$ of the extragalactic sky at visible to near-infrared (NIR) wavelengths, to a $5\,\sigma$ point-source depth of 26.2 mag[1] in the Euclid VISible Instrument (VIS: Cropper et al. 2016) I_E (R+I+Z) filter and 24.5 mag in the Near Infrared Spectrometer and Photometer (NISP: Maciaszek et al. 2016) Y_E, J_E and H_E filters (Euclid Collaboration: Scaramella et al. 2021; Euclid Collaboration: Schirmer et al. 2022). In addition, three fields with an area totalling 53 deg^2 will be the subject of a deeper survey, to a $5\,\sigma$ depth of 28.2 mag in I_E and 26.5 mag in Y_E, J_E, and H_E. *Euclid* is expected to detect ~ 12 billion astronomical sources ($3\,\sigma$), providing multicolour imaging at 0.″1–0.″4 resolution, and is also expected to obtain spectroscopic redshifts for ~ 35 million galaxies (e.g. Laureijs et al. 2011). While the primary science drivers of the *Euclid* mission are baryonic acoustic oscillations, weak lensing cosmology and redshift-space distortions, the survey is also expected to enable a multitude of high-impact extragalactic science projects, either in stand-alone form or in combination with multiwavelength data from other surveys (e.g. LSST).

Automated classification and derivation of physical properties are crucial steps towards scientific exploitation of data from any large astronomical survey, and traditionally this would be done using colour-colour methods (e.g. Haro 1956; Daddi et al. 2004; Leja, Tacchella, & Conroy 2019) or by fitting spectral models or templates (e.g. Bolzonella, Miralles, & Pelló 2000; Fotopoulou et al. 2012; Gomes & Papaderos 2017). Machine learning techniques have proven to be particularly powerful in this context, since they have the ability to detect structure, correlations, and outliers in large, multidimensional data sets, producing models that are typically stronger and more efficient than the traditional methods (e.g. Baqui et al. 2021; Ulmer-Moll et al. 2019; Logan & Fotopoulou 2020; Clarke et al. 2020; Cunha & Humphrey 2022). While machine learning techniques have been applied to extragalactic problems for several decades already (e.g. Odewahn et al. 1993), in recent years there has been a rapid growth in their application to this area. Notable examples include detailed morphological classification via application of deep learning to galaxy images (e.g. Dieleman, Willett, & Dambre 2015; Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018; Tuccillo et al. 2018; Nolte et al. 2019; Bowles et al. 2021; Bretonnière et al. 2021), high-accuracy estimation of the redshift (z) of galaxies from imaging/photometric data (e.g. Collister & Lahav 2004; Brescia et al. 2013; Cavuoti et al. 2017; Pasquet et al. 2019; Razim et al. 2021; Guarneri et al. 2021), selection and classification of galaxies into various phenomenological types (e.g. Cavuoti et al. 2014; Steinhardt et al. 2020), and derivation of their physical properties (e.g. Bonjean et al. 2019; Delli Veneri et al. 2019; Mucesh et al. 2021; Simet et al. 2021). A few studies have taken a hybrid or 'cooperative' approach, combining results from traditional methods (e.g. template fitting) with machine learning methods, resulting in improved accuracy (e.g. Fotopoulou & Paltani 2018; Cavuoti et al. 2017).

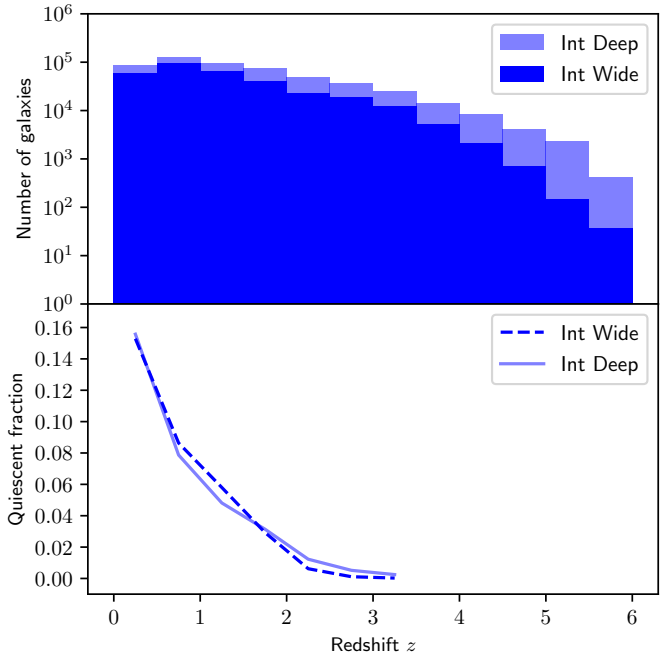

Fig. 1. The redshift distribution of galaxies in the Int Wide and Int Deep catalogues (top panel), and the quiescent galaxy fraction as a function of redshift up to $z = 3$ (lower panel).

The identification of quiescent galaxies[2] is among the most challenging classification problems in extragalactic astronomy, and represents a crucial task in our quest to understand the evolution of galaxies across cosmic time. A particular problem is the fact that there exist substantial degeneracies between stellar age, metallicity, and reddening by dust (e.g. Worthey 1994), causing the scattering of galaxies across classification boundaries when using broad-band spectral energy distributions (SEDs hereinafter). These degeneracies can be significantly exacerbated by the absence of redshift information.

A frequently applied technique for the selection of quiescent galaxy candidates is UVJ colour-colour selection (Strateva et al. 2001; Baldry et al. 2004; Wuyts et al. 2007; Williams et al. 2009; Muzzin et al. 2013; van der Wel et al. 2014; Leja, Tacchella, & Conroy 2019; Shahidi et al. 2020). This method selects objects in rest-frame $U - V$, $V - J_\mathrm{E}$ colour-colour space

[1] AB magnitudes are used herein.

[2] We adopt the specific star-formation rate (sSFR) of $10^{-10.5}\mathrm{yr}^{-1}$ as the boundary between quiescent and star-forming.

that are red because their UV to near-infrared SED is dominated by an old stellar population, as opposed to star-forming galaxies that are reddened by dust. While the UVJ selection method clearly works (e.g. Fumagalli et al. 2014), there is a substantial contamination by star-forming galaxies in the region of ∼ 10–30 per cent (e.g. Moresco et al. 2013; Schreiber et al. 2018; Fang et al. 2018). Various observer-frame colour combinations have also been proposed for the selection of quiescent galaxies, such as the BzK method (Daddi et al. 2004), or the GALEX FUV − V, $V - J_{\rm E}$, and FUV − V, $V - W3$ methods proposed by Leja, Tacchella, & Conroy (2019). Once selected via colour-colour techniques, spectral energy distribution (SED) fitting (e.g. Wiklind et al. 2008; Girelli, Bolzonella, & Cimatti 2019) and spectroscopic observations (e.g. Belli, Newman, & Ellis 2015; Glazebrook et al. 2017; Schreiber et al. 2018) may then be employed to confirm their passive nature. Alternative approaches have also been successful at selecting quiescent galaxies, such as template fitting followed by colour selection (e.g. Laigle et al. 2016; Deshmukh et al. 2018).

In the context of preparations for the Euclid survey, Bisigello et al. (2020, B20 hereinafter) recently developed $I_{\rm E} - Y_{\rm E}$, $J_{\rm E} - H_{\rm E}$ and $u - I_{\rm E}$, $I_{\rm E} - J_{\rm E}$ colour-colour criteria to separate quiescent galaxies and star-forming galaxies up to $z = 2.5$, for use in the case where spectroscopic or photometric redshifts are available. The proposed colour-colour criteria significantly outperform the traditional UVJ technique which, when derived using only the four Euclid filters, provides a completeness of only ∼ 0.2 at $z < 3$. For example, using their $u - I_{\rm E}$, $I_{\rm E} - J_{\rm E}$ criteria, B20 were able to select a sample of quiescent galaxies at $0.75 < z < 1$ with a completeness of ∼ 0.7 and a precision of > 0.85. Their $I_{\rm E}-Y_{\rm E}$, $J_{\rm E}-H_{\rm E}$ criteria also allowed the authors to select quiescent galaxies at $1 < z < 2$ with a completeness of > 0.65 and a precision of > 0.8.

An important limitation of colour-colour selection techniques is that they are, in effect, lossy dimensionality-reduction methods. As such, they are likely to discard a significant quantity of otherwise useful information that is present in a broad-band SED. In this context, machine learning methods offer a promising alternative, since they are able to perform selection from within highly multidimensional data sets, and can be tuned to make a trade-off between purity and completeness that is appropriate for a desired science application. Indeed, Steinhardt et al. (2020) recently explored the selection of quiescent galaxies using the unsupervised t-distributed stochastic neighbour embedding method (van der Maaten & Hinton 2008; van der Maaten 2014), reporting a significant improvement over the UVJ and template fitting methods.

In this paper, we present a new supervised machine-learning method for the separation of quiescent and star-forming galaxies using Euclid and ancillary photometry, which is designed to handle sparse data and which can provide photometric redshift estimates where necessary. The paper is organized as follows. In Sect. 2 we describe the mock photometry catalogues used in this study. The metrics of model quality we use are defined in Sect. 3. Full details of the `ARIADNE` pipeline are given in Sect. 4. We describe in Sect. 5 the results of applying our separation methods to the mock photometric data. Next, in Sect. 6 we compare our method with colour-colour methods previously proposed by B20 and others. In Sect. 7 we summarize results from a number of further analyses and tests, with full details given in Appendix B. Finally, in Sect. 8 we summarize our results and conclusions.

2. Mock galaxy catalogues

In this work we make use of an updated version of the mock catalogues presented in B20. All catalogues were derived from magnitudes in the COSMOS2015 multi-wavelength catalogues (Laigle et al. 2016). Objects labelled as stars or X-ray sources, and objects with inadequate[3] optical photometry, being removed. After this selection, the COSMOS2015 catalogue contains 518404 objects up to $z = 6$. We now briefly introduce the two methods used to derive Euclid-like mock catalogues.

In the first method, we interpolated over the observed COSMOS2015 photometry to produce a broken-line template running from ultraviolet to infrared wavelengths, which is then convolved with the filters of interest to derive mock photometry. As this method is affected directly by the COSMOS2015 photometric errors, which are similar or larger than those expected in the Euclid Wide Survey, we did not include any additional photometric scatter. The 20 cm VLA radio continuum flux, where measured, is also included without modification. Using this approach, we derived two separate catalogues, resembling the Euclid Wide Survey (Euclid Collaboration: Scaramella et al. 2021) and Euclid Deep Survey. We refer to these mock catalogues as 'Int Wide' and 'Int Deep', respectively.

The question of the potential impact of emission lines on the Int catalogues was discussed previously by B20. In short, the broad-band magnitudes in the Int and SED catalogues include contributions from emission lines. Because the Int catalogues were constructed using real observed photometry, this means that for some bands and some galaxies, a contribution from nebular emission is present. In the case of the SED catalogues, the LePhare code was used with emission lines included (see Ilbert et al. 2006), allowing nebular emission to contribute to the mock magnitudes.

Nevertheless, we do not expect the inclusion (or exclusion) of emission lines to have a significant effect on our results. Given the widths of the filters considered herein (Euclid Collaboration: Schirmer et al. 2022), the effect of emission lines is marginal: observed equivalent widths larger than 350 Å, 260 Å, 390 Å, and 480 Å would be required to produce a boost of ∼ 0.1 mag in the VIS, Y, J, and H bands, respectively; such high equivalent widths are rare in the sample of galaxies used for our mock catalogues (e.g., Amorín et al. 2015)

In the second approach, we used the public code `LePhare` (Arnouts et al. 2007; Ilbert et al. 2006) to fit the COSMOS2015 photometry with a large set of Bruzual & Charlot (2003) templates, with the redshift fixed at its COSMOS2015 value from Laigle et al. (2016). In particular, we considered templates with two different metallicities ($Z_\odot$ and 0.4 $Z_\odot$), exponentially declining star-formation histories with an e-folding timescale τ varying from 0.1 to 10 Gyr, and ages between 0.1 and 12 Gyr. For the dust extinction, we considered the reddening law of Calzetti et al. (2000) with 12 values of colour excess from 0 to 1. For each galaxy, we obtained the best SED template applying a χ^2 minimisation procedure and we convolved the resulting template with the filters of interest to calculate the desired mock photometry. In effect, the resulting photometric SED is a synthetic representation of the observed one. For each galaxy, we derived 10 mock galaxies by randomising the mock photometry within the expected photometric errors. As with the first approach, here we derived two different catalogues, one for the Euclid Wide Survey (SED Wide) and one for the Euclid Deep Survey (SED Deep).

[3] Sources that are "masked in optical broad-bands" in the COSMOS2015 catalogue.

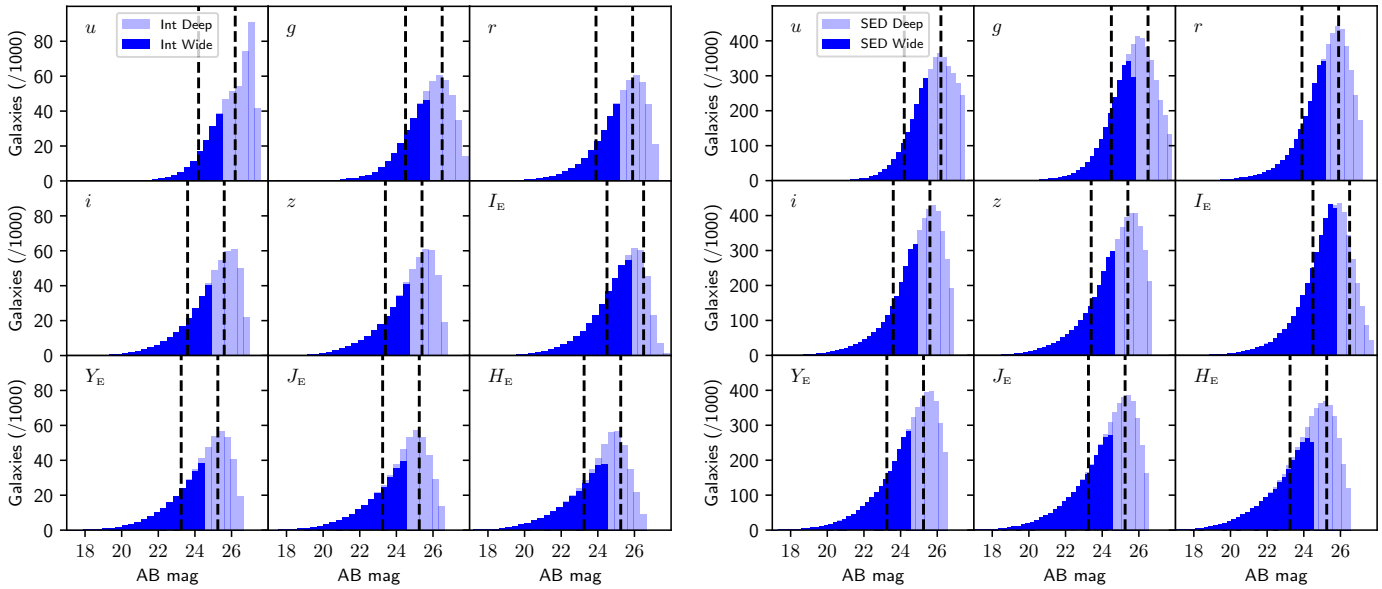

Fig. 2. Distribution of galaxy magnitudes in the Int catalogues (left) and the SED catalogues (right) for the various optical and near-IR bands used herein. Vertical dashed lines indicate the expected $10\,\sigma$ sensitivity of the Euclid Wide and Deep Surveys in the I_E, Y_E, J_E, and H_E bands. In all bands, only photometry with a signal-to-noise ratio of ≥ 3 is included.

Table 1. $10\,\sigma$ depth in AB magnitudes of the Wide Survey for the filters included in the mock catalogues. The Deep Survey is expected to be two magnitudes deeper than the Wide survey in the Euclid , CFIS and SDSS bands.

I_E	Y_E	J_E	H_E	CFIS/u	SDSS/g	SDSS/r	SDSS/i	SDSS/z	$W1$	$W2$
24.5	23.25	23.25	23.25	24.20	24.50	23.90	23.60	23.40	18.39	18.04

Because the templates do not extend into the radio regime, the SED Wide and SED Deep mock catalogues do not include the 20 cm radio band.

We have included the VIS I_E filter, the NISP Y_E, J_E, and H_E filters (Euclid Collaboration: Schirmer et al. 2022), and the CFIS/u filter, as previously presented in B20. In addition, we derived mock fluxes for the Sloan Digital Sky Survey (SDSS; Gunn et al. 1998) *griz* filters and the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) filters at 3.4 and 4.6 μm ($W1$ and $W2$). For the WISE filters, we considered the observational depth of the WISE All-Sky survey, i.e. $5\,\sigma$ 0.08 and 0.11 mJy (Wright et al. 2010). For the other filters we instead assumed the observational depths reported in the Euclid Red Book (Laureijs et al. 2011), which are expected to be reached with a variety of ground-based telescopes, for which we use the SDSS filters as proxies. The complete list of observational depths are listed in Table 1. Our choice of depths for the Euclid and ground-based photometry is motivated primarily by the need to make a direct comparison with the colour-colour results of B20. Thus, we have adopted the depths used in B20[4].

[4] The Euclid photometric depths adopted herein differ slightly from the most recent forecasts: the I_E photometry is now forecast to be 0.5 mag deeper, and the NISP photometry 0.25 mag deeper (Euclid Collaboration: Scaramella et al. 2021), compared to the values adopted herein. In the case of ground-based optical photometry overlapping the Euclid survey areas there is no single forecast, since the photometry is expected to come from several different surveys. For instance, at the time the Euclid DR3 release, the UNIONS survey (Chambers et al. 2020) is forecast to be 0.6 mag shallower in u and 0.2 mag deeper in r compared to the Wide Survey depths adopted herein, with simular depths in the g, i and z bands. In the Southern Hemisphere, LSST is expected to provide deeper optical data.

For all bands, only photometry measurements with a signal-to-noise ratio of $\geq$ 3 are considered. This threshold has been chosen to ensure that only reliable measurements are used. The use of low signal-to-noise data deserves a detailed and in-depth study, and is under investigation for a future paper. For all bands except I_E, non-detections are flagged as missing; objects for which I_E has a signal-to-noise below 3 are excluded from the catalogues. We refer to B20 for further details on the creation of the mock catalogues.

Specific star-formation rates (sSFR) were derived for each galaxy at $z \leq 3$ by fitting the 30-band photometric SED of Laigle et al. (2016) using `LePhare`. Further details of this process are given in B20. The dividing line between 'quiescent' and 'star-forming' in terms of sSFR is somewhat arbitrary, and various different thresholds are in use in the literature, although these are usually in the range $< 10^{-10}$ yr^{-1} (e.g. Wu et al. 2018) to $< 10^{-11}$ yr^{-1} (e.g. Ilbert et al. 2013). For consistency with B20, we define 'quiescent' to mean that a galaxy has sSFR $< 10^{-10.5}$ yr^{-1}, and 'star-forming' to mean that a galaxy has sSFR $\geq 10^{-10.5}$ yr^{-1}. In any case, the classification metrics we obtain herein are not significantly dependent on which value of sSFR we adopt for the threshold between quiescent and star-forming galaxies, for values of the threshold between 10^{-10} yr^{-1} and 10^{-11} yr^{-1}.

For the redshift labels we adopt the 30-band COSMOS photometric redshifts estimated by Laigle et al. (2016). For the SED catalogues, the redshift labels represent the true redshifts (i.e., with no uncertainty), because the mock photometry is derived directly from templates with known redshifts. On the other hand, in the case of the Int catalogues, the redshift labels are merely photometric redshift estimates, and thus have an uncertainty with systematic and random components.

We characterize the properties of the mock catalogues in Figs. 1–4. In Fig. 1 we show the redshift distribution of galaxies. Also shown is the distribution and fraction of quiescent galaxies as a function of redshift, up to $z = 3$. It can be seen that the quiescent fraction falls rapidly with increasing redshift, starting at ~ 0.16 at $z \sim 0$ and declining to ≤ 0.05 by $z \sim 2.5$, illustrating the highly challenging nature of the search for quiescent (or passive) galaxies at high redshift.

The distribution of magnitudes is shown in Fig. 2, with the expected 10σ detection limits in the Euclid filters marked. This Figure shows that the Wide mock catalogues are complete down to the 3σ detection limit in all optical and near-IR bands. However, the Deep catalogues are not complete, due to the limits of the COSMOS2015 photometry catalogue.

Fig. 3 shows the distribution of $I_{\rm E} - H_{\rm E}$, $u - I_{\rm E}$, and $Y_{\rm E} - H_{\rm E}$ colours in the four mock catalogues. There are slight colour differences between the Int and SED catalogues that arise due to their different construction methodologies. For instance, the SED catalogues are slightly bluer in terms of their average observer-frame $u - I_{\rm E}$ colours, compared to the Int catalogues.

In Fig. 4 we show the fraction of missing photometry measurements as a function of redshift for the optical and infrared bands. In the case of our mock catalogues, photometry is flagged as missing when (i) a galaxy or area is unobserved in that band, (ii) a galaxy or area was masked, (iii) a photometry measurement falls below the detection threshold of the catalogue, or (iv) a photometry measurement has a signal-to-noise ratio lower than 3. In general terms, the missing fraction is higher for higher redshifts, although in some cases (notably the near-IR bands) there is a turnover and decrease in the missing fraction at very high redshifts ($z \gtrsim 4$). The WISE $W1$ and $W2$ bands and the 20 cm radio band data are highly sparse, with missing fractions of 0.947, 0.976, and 0.992, respectively. Conversely, the $I_{\rm E}$ band has a missing fraction of exactly 0, since detection in this band is a requirement for inclusion in our mock catalogues.

Finally, Fig. 5 shows the impact on the number and redshift distribution of galaxies from application of the main photometric pre-selection criteria used herein: (i) $3\,\sigma$ detections in $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$; (ii) $3\,\sigma$ detections in u, $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$; (iii) $3\,\sigma$ detections in *ugriz*, $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$. As expected, the impact of requiring detection in the u-band is to induce a step in the distribution at $z \sim 3$, as the Lyman break is redshifted out of the u-band filter; this step is much stronger in the Wide catalogues than in the SED catalogues. In the Int Deep catalogue, the reduction in the number of sources at $z \gtrsim 3$ when detection in u or *ugriz* is required is surprisingly small; this is likely to be at least partly caused by the presence of sources with incorrect photometric redshifts in the COSMOS2015 catalogue. Table C.1 lists the number of galaxies that are detected in each band for each of the mock catalogues.

3. Metrics of model quality

To evaluate our classification models, we use several metrics that are useful to quantify model quality. Precision P is the fraction of assignments to a particular class that are correct, calculated as

$$P = \frac{\rm TP}{\rm TP + FP}\,, \tag{1}$$

where TP is the number of true positives and FP is the number of false positives. Precision is also sometimes known as 'purity' in astronomy, or the 'positive predictive value'.

Recall R is the fraction of galaxies of a particular class within the data set that are correctly classified as such. It is calculated as

$$R = \frac{\rm TP}{\rm TP + FN}\,, \tag{2}$$

where FN is the number of false negatives. Recall is also sometimes known as 'completeness' or 'sensitivity'.

The F1-score is the harmonic mean of the precision and the recall, and as such provides a more general metric for model quality (Dice 1945; Sørensen 1948). The F1-score is calculated as

$$\text{F1-score} = 2\,\frac{P\,R}{P + R}\,, \tag{3}$$

where we have opted to give equal weights to P and R. The metrics P, R, and F1-score have values between 0 and 1. Since there is a large imbalance between classes in the data sets used here, we compute the metrics separately for each class. Unless otherwise stated, the values of P, R, and the F1-score we quote are computed for the 'quiescent' galaxy class only.

To assess the quality of photometric redshift estimates produced by our pipeline, we use the following measures. As a measure of accuracy, we use the normalized median absolute deviation (NMAD), which we calculate as

$$\text{NMAD} = 1.48\,\text{median}\left(\frac{|z_{\rm phot} - z_{\rm ref}|}{1 + z_{\rm ref}}\right), \tag{4}$$

where $z_{\rm phot}$ is our photometric redshift and $z_{\rm ref}$ is the reference redshift used as the 'ground truth'. In this study, we adopt the 30-band photometric redshifts from COSMOS2015 (Laigle et al. 2016) for $z_{\rm ref}$. The NMAD can be loosely interpreted as the standard deviation.

As a further measure of quality, we also calculate the fraction of catastrophic outliers ($f_{\rm out}$). A photometric redshift estimate is considered to be a catastrophic outlier when

$$\frac{|z_{\rm phot} - z_{\rm ref}|}{1 + z_{\rm ref}} > 0.15. \tag{5}$$

To test whether (and to which extent) our pipeline systematically over- or underestimates galaxy redshifts, we define the bias of the photometric redshifts as

$$\text{bias} = \text{median}\left(\frac{z_{\rm phot} - z_{\rm ref}}{1 + z_{\rm ref}}\right). \tag{6}$$

It is also useful to have a metric that quantifies the degree to which a false positive classification is in error. We define the *incorrectness* of an individual false positive classification as

$$I_{\rm FP} = \log_{10}(\text{sSFR yr}) + 10.5. \tag{7}$$

We consider a false positive to be 'marginal' when $I_{\rm FP} \leq 0.5$, and 'catastrophic' when $I_{\rm FP} \geq 1.0$. We also define an average incorrectness parameter as

$$\bar{I}_{\rm FP} = \sum_{i=0}^{n} \frac{\log_{10}(\text{sSFR yr}) + 10.5}{\rm FP}\,. \tag{8}$$

We have not used the commonly used 'accuracy' metric, which gives the fraction of predictions that are correct, because it can be misleading when the test data are significantly imbalanced, as is the case here.

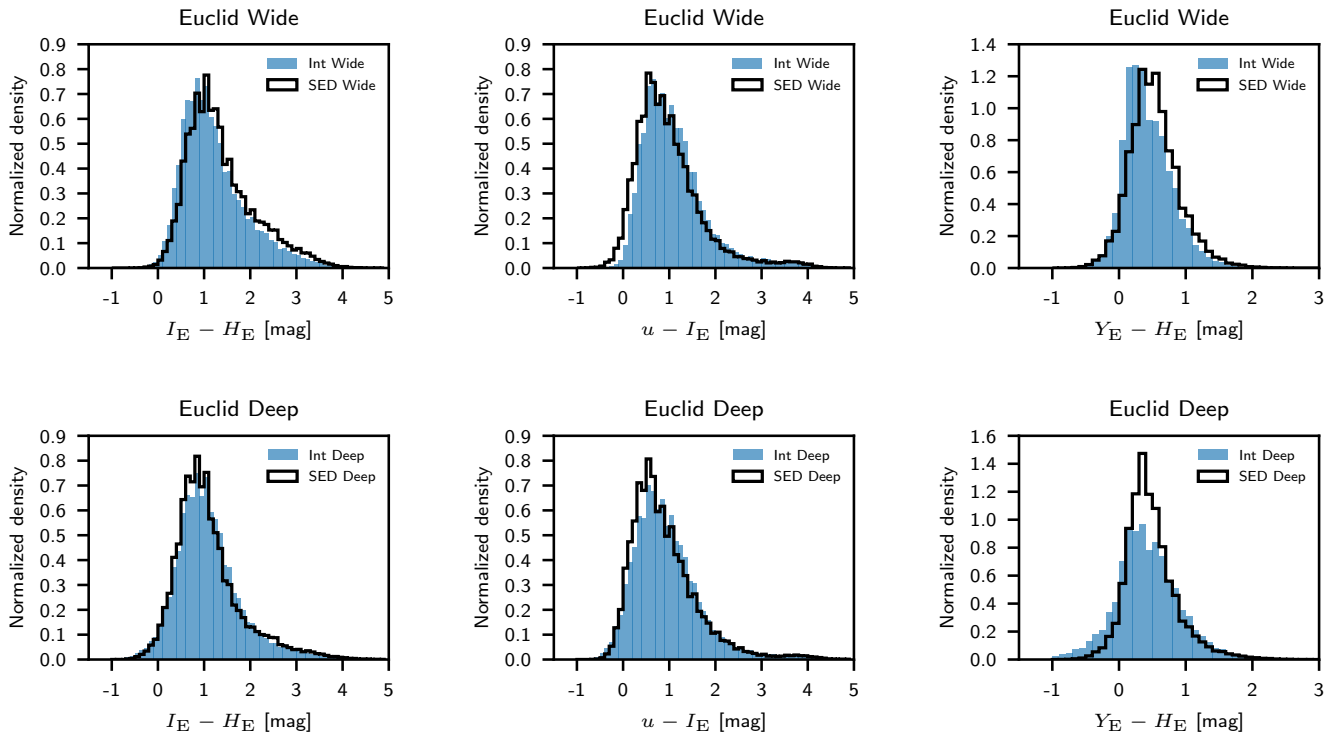

Fig. 3. Comparison between the distribution of the $I_E - H_E$, $u - I_E$, and $Y_E - H_E$ colours in the Int and SED mock catalogues. **Top row:** Euclid Wide catalogues. **Bottom row:** Euclid Deep catalogues. Significant differences between the Int and SED methods are apparent, most notably with the SED method giving rise to significantly bluer $u - I_E$ colours compared to the Int method. These differences arise when the galaxy templates are unable to closely match the observed broad-band spectral energy distribution; this may be due to the photometric redshift being incorrect, and/or due to the absence of a template that sufficiently represents the properties of the galaxy.

4. The ARIADNE pipeline

`ARIADNE` is a flexible, modular machine learning pipeline designed for the purpose of classification and derivation of physical properties of astronomical sources on the basis of their photometric SEDs. In a nutshell, `ARIADNE` takes a table of photometry as input, and uses algorithms to learn how the input data maps to labels corresponding to galaxy properties. Here we describe the functionality of `ARIADNE` in its classification mode (a flowchart is also shown in Fig. 6); its application to estimation of physical properties, such as stellar mass, star-formation rate, extinction, etc., will be presented in a future publication (Humphrey et al., in prep.).

4.1. The learning algorithm

After completion of the feature engineering and preprocessing steps described in Sect. 4.2, the data are split into a training set and test set with a 2:1 ratio between the two. This ratio was chosen to give a good balance between having a large number of examples on which to train the learners, while still having a test set that is representative of the whole data set[5]. The split is done randomly, without attempting to balance target labels, redshift, or any other observational properties, and each execution of the pipeline produces a different train-test split.

The pipeline then trains binary classification models on the training data set, using five different 'base-learners' (see Fig. 6). Three of the base-learners are tree-ensemble methods with somewhat different implementations (`CatBoostClassifier`[6] `version 0.23.2`; `LGBMClassifier`[7] `version 0.90`; `RandomForestClassifier`), one is nearest-neighbours-based (`KNeighborsClassifier`), and one is deep-learning-based (`MLPClassifier`). All are open source and are briefly summarized below.

The Python `Scikit-Learn`[8] `version 0.22.2` package (Pedregosa et al. 2011) offers various machine learning methods with excellent functionality and interoperability with a multitude of in-package preprocessing tools. From `Scikit-Learn` we make use of the nearest-neighbours-based, non-parametric `KNeighborsClassifier` method, the Multi-layer Perceptron-based `MLPClassifier`, and the randomized decision-tree classifier `RandomForestClassifier` (Breiman 2001).

Several other, advanced machine learning methods are used in our pipeline. `CatBoost` (Prokhorenkova et al. 2018) is a gradient-boosting, tree-ensemble method that offers high performance classification or regression, using 'ordered-boosting' in place of the classic boosting algorithm to significantly reduce the 'prediction shift' commonly associated with the latter.

Arguably the most advanced tree-based learning algorithm publicly available at the time of writing is `LightGBM`, a gradient-boosting, decision-tree method that deploys a number of key innovations that are especially relevant for the classification problems we approach in this work (Ke et al. 2017). For instance, `LightGBM` uses leaf-wise (best-first), rather than level-wise tree-growth, for improved classification accuracy. Also of relevance

[5] When applied to the actual Euclid survey data it is expected that the ratio used for the train test split may be of order $\sim 1:10\,000$

[6] https://catboost.ai
[7] https://lightgbm.readthedocs.io
[8] https://scikit-learn.org

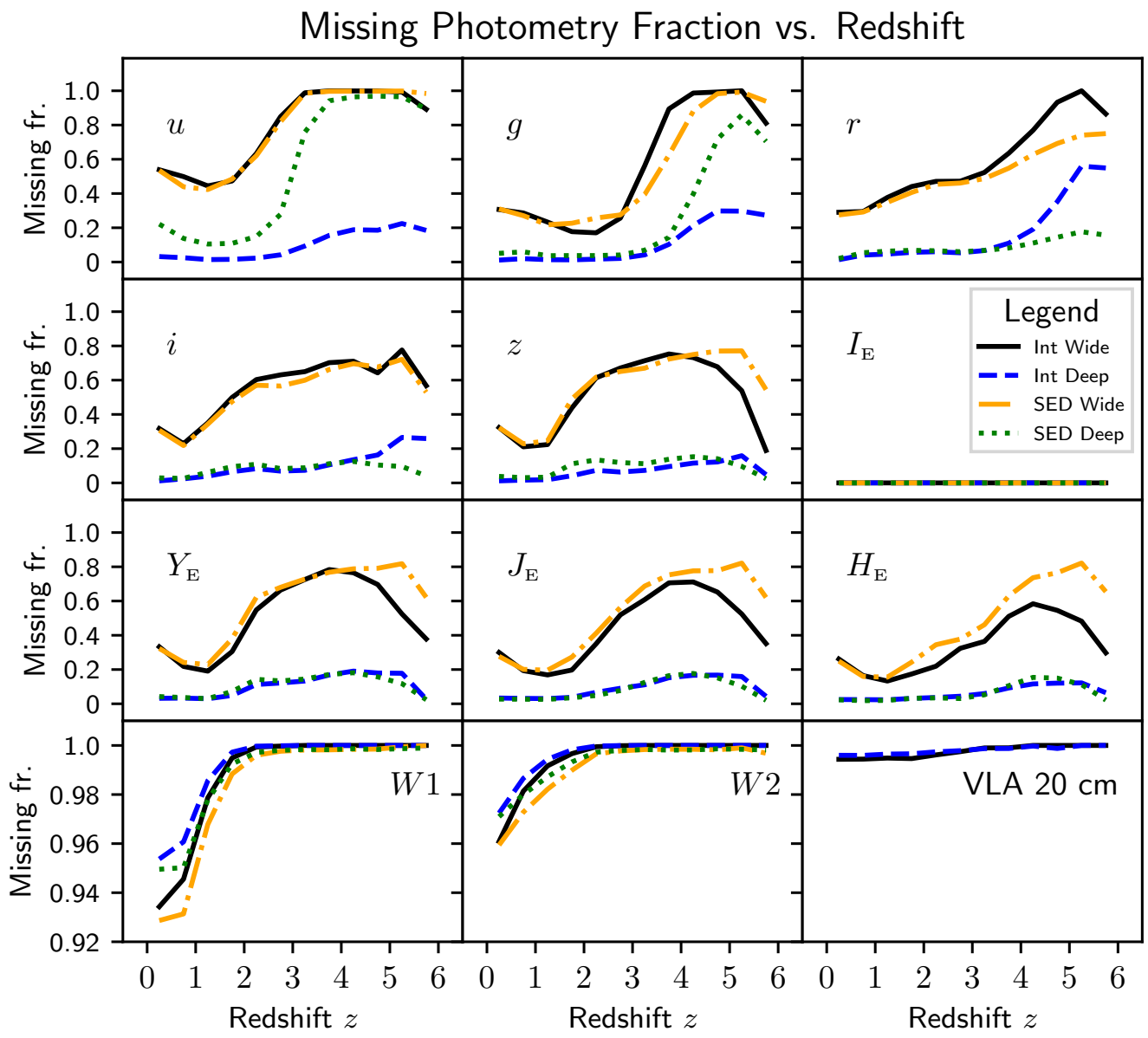

Fig. 4. Fraction of missing photometry measurements vs. redshift for each of the filters used for model training. By construction, the I_E band has no missing values. Very few of the galaxies are detected in the $W1$, $W2$, or VLA 20 cm bands.

is the use of histogram-based algorithms which place continuous feature values into discrete bins, significantly reducing training time and memory usage.

In addition to the five default base-learners described above, our pipeline also includes the option to use the tree-based method `XGBoost`[9] `version 0.25.3` (Chen & Guestrin 2016), for use in situations where this learning algorithm is able to train stronger models than those described above. The `XGBoost` algorithm uses gradient boosting and has several key innovations, including sparsity-aware split-finding, which is of particular relevance for data sets containing a significant fraction of missing values (e.g. photometric non-detections, masked or unobserved areas).

Our application of these learning algorithms to the data at hand reveals that no single algorithm consistently outperforms the others over the full range of classification problems posed herein. Thus, our pipeline employs model ensembling and interactive algorithm selection.

The base-learners are trained within a stratified k-fold procedure: the training data are shuffled and split, with replacement, into 5 similarly sized folds, ensuring that each fold contains the same proportion of each target class. Each base-learner is trained 5 times, each time leaving out a different fold, for which a set of out-of-fold (hereafter OOF) class predictions is produced. This results in an array of OOF predictions for each base-learner. For each iteration of the k-fold procedure, class predictions are also produced for the test set, averaging the 5 sets of class predictions for each base-learner. Unless otherwise stated, we adopt a class probability threshold of 0.5: predictions of < 0.5 correspond to class 0 (star-forming galaxies), and predictions of ≥ 0.5 correspond to class 1 (quiescent galaxies).

The pipeline then performs several iterations of non-linear combination of class predictions from the individual learners.

[9] https://xgboost.readthedocs.io

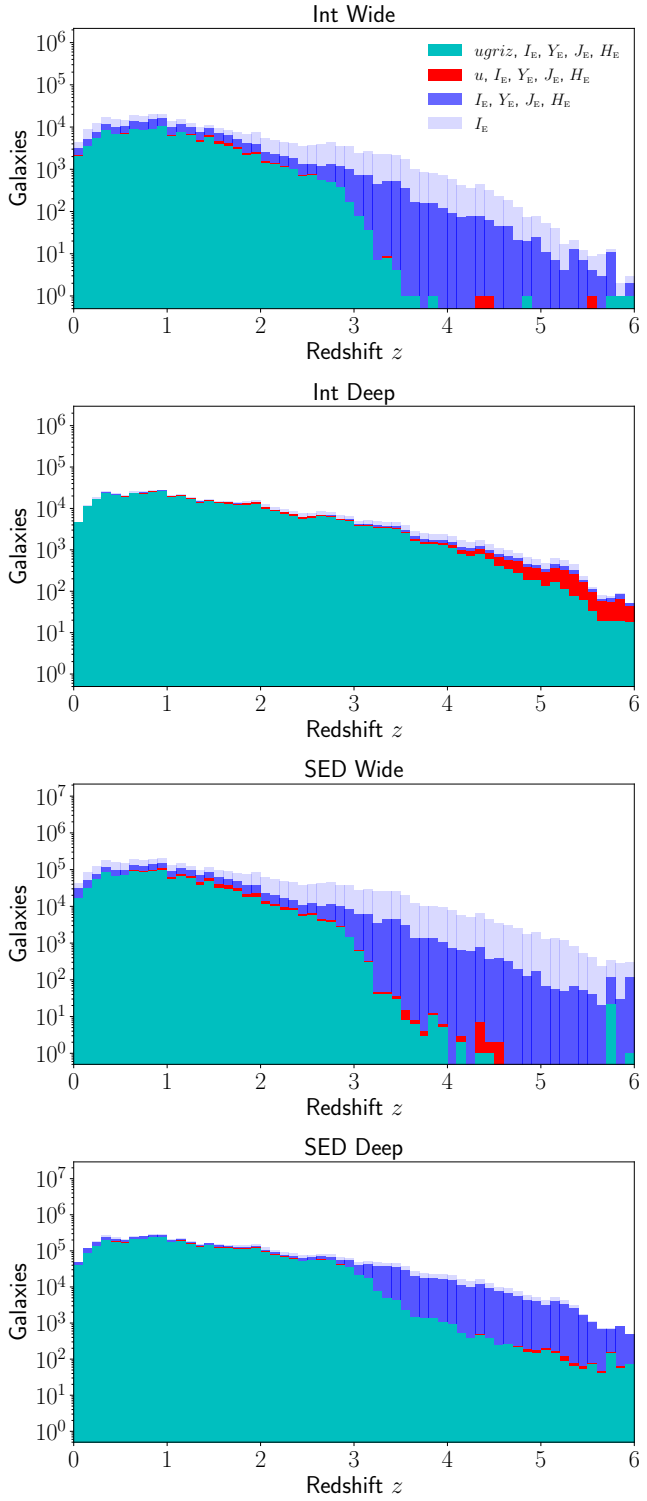

Fig. 5. The effect of various photometric pre-selection criteria on the number and redshift distribution of galaxies in the four mock catalogues. We show four cases: (i) only a detection in I_E is required; (ii) detection is required in all four Euclid bands (I_E, Y_E, J_E, H_E); (iii) detection is required in u and all four of the Euclid bands; (iv) detection is required in *ugriz* and all of the Euclid bands. As described in the main text, for all bands we adopt a signal-to-noise detection threshold of ≥ 3.

First, a hard-voting ensemble is produced for the OOF and test predictions. Each base learner contributes one vote towards the class of a galaxy, and the class with the highest number of votes is chosen. For the data and classification problem considered in this work, the hard-voting ensemble almost always results in a significantly higher F1-score than any of the individual base-learners, or a simple average of the class probabilities.

To further improve model quality, the pipeline contains our implementation of the 'generalized stacking' method proposed by Wolpert (1992). A linear discriminant analysis or `MLPClassifier` meta-learner is trained to classify galaxies using the OOF predictions from the 5 base-learners as features. This is performed within a stratified k-fold procedure as described above, and produces a new set of OOF predictions for the training data set, in addition to a new set of predictions for the test data set. A second iteration of meta-learning is then performed, this time training on just two features: (i) the results of the hard-voting ensemble, and (ii) the OOF predictions produced in the previous iteration of meta-learning. Finally, the resulting model is used to predict classes for the test data. For an alternative implementation of generalized stacking, applied to redshift estimation, see Zitlau et al. (2016).

One of the benefits of generalized stacking is that the optimization of base-learner hyperparameters, while still necessary to some extent, is not as crucial for the final performance of the pipeline as it would be when a single learning algorithm is used. This is partly because when the meta-learner distills the base-classifiers into a single classifier, it performs a process broadly analogous to optimization and model-selection. Arguably, this process can be more efficient than traditional optimization and model selection methods, since it is performed in a single step and is not restricted to selecting a single model or a single set of hyperparameters, and instead can combine the strengths of several different classifiers that are best able to model different subsets of the data. In the case where a single base-learner is used within our pipeline, the generalized stacking procedure instead serves as an 'error-correction' algorithm.

We have not attempted to perform an exhaustive optimization of the base-learners prior to applying our generalized stacking method. Instead, for each learner we have manually identified a set of default hyperparameters that gives what we judged to be near to the global maximum of the F1-score for selecting quiescent galaxies from the mock catalogues. This is done to avoid biasing the individual base-learners towards producing classifiers that all succeed (or fail) in modelling the same subset of the data, and to expand the diversity of classification models available to the meta-learner.

In the default configuration of our pipeline, all five default base-learners are used. However, the user can instead use a subset of the base learners, or a single base-learner, when appropriate. For example, we opt to use `XGBoostClassifier` for the selection of quiescent galaxies at $2.5 < z < 3.0$, where its sparsity awareness confers a significant advantage over the other classifiers. In addition, the pipeline has a '*fast mode*' which uses `LightGBM` for all classification or regression tasks, at the cost of a small but significant reduction in P, R, and F1-score (~ 0.01–0.03). Timed on a mid-range laptop with a quad-core Intel i5-8350U CPU and 16 Gigabytes of RAM, the pipeline used in *fast mode* takes at total of $\sim$ 2 minutes to train its classifier on a data set with $\sim$ 120 000 examples and 70 features, compared to $\sim$ 74 minutes when using the default (5 base-learners) pipeline configuration.

Our pipeline makes use of redshift information in one of several ways, depending on the classification problem that is posed. When redshifts are available, these can be included as an additional feature in the training and test data. In addition, when the objective is to select quiescent galaxies in a specific redshift in-

terval, pre-binning can be used to discard objects that lie outside the desired redshift interval.

In the absence of redshifts for the test sample, the pipeline first performs a global selection of quiescent galaxies (without regard to redshift), then trains a `KNeighborsRegressor` model to predict photometric redshifts for the selected quiescent galaxies. The photometric redshift point estimates are refined using our implementation of the semi-supervised 'pseudo-labelling' technique (Lee 2013), which aims to use both labelled and unlabelled data to learn the underlying structure of the data, thereby improving generalisation. Finally, analogous to Fotopoulou & Paltani (2018); Singal et al. (2022), a `KNeighborsRegressor` model is trained to predict whether a galaxy's redshift estimate is a catastrophic outlier, with a tunable probability threshold to control the strength of the outlier removal.

4.2. Feature engineering

4.2.1. Broad-band colours

Before applying our algorithm to the data, we first performed a pre-processing step known as 'feature engineering', whereby the data are enriched with information to help the algorithm learn more efficiently. We start with the broad-band magnitudes and their $1\,\sigma$ errors as base features. Because colours offer a potentially clearer description of the relative shape of the broad-band SED than magnitudes, we have also calculated all possible (unique) broad-band colour permutations, and included them as features. In cases where fluxes are given instead of magnitudes (e.g. VLA radio flux measurements), we converted the values to magnitudes before deriving the related colours.

4.2.2. Missing data imputation strategy

One of the advantages of our machine learning approach to galaxy classification is the possibility to efficiently deal with missing data. A subset of galaxies in the mock catalogues is undetected, or unobserved, in one or more filter; we consider it highly desirable to include them in our analysis, where possible, for several reasons: (i) such objects enlarge our training data set; (ii) non-detection in some bands is likely to carry useful information for our learning algorithm (e.g. u-band drop-outs); (iii) the upcoming large surveys (e.g. Euclid, LSST, etc.) that motivate this work will produce large data sets where many galaxies have missing data in one or more bands, and such objects need to be utilized to make the most effective use of the survey data.

Here, we impute values for missing data with a method that is independent of the reason for it to be missing. When using tree-based learning algorithms (`CatBoostClassifier`, `LightGBMClassifier`, `RandomForestClassifier`, or `XGBoostClassifier`), we impute the missing values with the arbitrarily chosen constant value −99.9. All broad-band colours that would have used the missing value are also set to −99.9. Because none of the measured magnitudes or colours have this value (nor do they have similar values), information about the presence of missing values is thus preserved such that the tree-based learners can use non-detections to aid their classification of sources; this information is essentially lost if the average, median, or minimum value would be used instead for the imputation. Conversely, the `MLPClassifier` and `KNeighborsClassifier` learning algorithms are generally more sensitive to the normalisation of the input features, and imputing data with the value −99.9 is likely to create inappropriate and unhelpful artefacts in feature-space; therefore, in the cases of `MLPClassifier` and `KNeighborsClassifier` we instead impute missing values with the mean of the respective feature, computed across the sample of galaxies using all the non-missing values.

We emphasize that the primary motivation behind our imputation strategy is to flag non-detections such that the learning algorithms can deduce how to use them most effectively; an added benefit of this strategy is that it allows the use of objects with photometric SEDs that are missing one or more bands, without necessarily having to discard them.

Table 2 illustrates the impact on the F1-score from using one of several different imputation strategies, using a fixed random seed for the train/test split and base learners. Results for the following strategies are shown: imputation with the mean, the median, the minimum, a constant value of −99.9, or dynamically switching between −99.9 for tree based learners and the mean for `MLPClassifier` and `KNeighborsClassifier`. In each case, the F1-scores are shown for the base learners and for the final stacked ensemble classifier. While the results vary significantly depending on the choice of random seed, the general outcome is that, as expected, the tree-based learning algorithms (`CatBoostClassifier`, `LightGBMClassifier`, `RandomForestClassifier`, or `XGBoostClassifier`) generally give higher F1-scores when using missing values that are imputed with −99.9, whereas `MLPClassifier` and `KNeighborsClassifier` generally produce higher F1-scores when using the mean, median, or minimum of a feature for imputation.

To provide the learning algorithms with additional help to treat missing data, we created a feature (`n_missing`) which counts the number of missing magnitude values for each galaxy. Although not needed in the present study, the `ARIADNE` pipeline has the capability to make use of categorical flags that specify the reason for each missing data point in the input data (i.e., non-detection vs. not observed or masked).

Standardization was performed using the `Standard Scaler` from `Scikit-Learn`, which removes the mean and scales to unit variance. Missing values flagged with the value −99.9 are ignored during the standardization procedure.

4.2.3. Target variable

We generate a target feature representing the binary classification of the galaxies. This feature is dynamically filled, depending on the specific subset of galaxies to be selected. At its simplest, the target variable is set to 0 for star-forming galaxies and 1 for quiescent galaxies. To select quiescent galaxies in a specific redshift range, the target feature is set to 1 for all quiescent galaxies in that redshift band, and 0 for all other galaxies.

It is important to note that in the case of the Int catalogues, the sSFR label, and consequently the binary target variable, has an intrinsic uncertainty. Depending on the nature of the uncertainties, it is entirely possible that our classification methodology is outperforming the initial sSFR evaluation. However, a detailed analysis of this potential effect is beyond the scope of this Paper.

4.2.4. Feature importance and selection

It is important to ensure that our models are trained using only features that provide useful information for the prediction of the target variable. First, we examined the feature importance information provided by three of the tree-based learning algorithms that we use here (`RandomForestClassifier`,

Table 2. The impact on the F1-score from using one of several different imputation strategies. In this example, we have trained models to select quiescent galaxies from the Int Wide mock catalogue in the redshift bin $2 < z < 2.5$, using *ugriz*, Euclid, *W*1, *W*2, and 20 cm photometry and colours, and with pre-binning by redshift, as described in Sect. 5.1.2. A single random seed is used for the train/test split and base learners to allow a relatively controlled comparison between methods. The columns are as follows: (1) The learning algorithm; also shown are the final F1-scores after the models produced by the 5 default base-learners have been ensembled using meta-learners; (2) F1-score when missing values are imputed with the constant value −99.9; (3) F1-score when imputing with the average value of a feature; (4) F1-score when missing values are dynamically imputed with either the constant value −99.9 (tree-based learners) or the mean of a feature (`MLPClassifier` and `KNeighborsClassifier`); (5) F1-score when imputing with the median value of a feature; (6) F1-score when imputing with the minimum value of a feature. Some F1-scores differ significantly to those presented in Sect. 5, since here we use only a single random seed instead of averaging results over multiple pipeline runs that use different random seeds. Note that in this test, `MLPClassifier` was unable to correctly identify any quiescent galaxies when missing values were imputed with −99.9; nonetheless, this failure did not appear to be detrimental to the final meta-learner ensemble.

Learning algorithm	Imputed value:	−99.9	mean	−99.9 or mean	median	minimum
(1)		(2)	(3)	(4)	(5)	(6)
`CatBoostClassifier`		0.633	0.632	0.633	0.644	0.621
`LightGBMClassifier`		0.678	0.621	0.678	0.596	0.655
`RandomForestClassifier`		0.607	0.561	0.607	0.576	0.607
`MLPClassifier`		0.000	0.610	0.610	0.621	0.633
`KNeighborsClassifier`		0.519	0.526	0.526	0.526	0.519
Meta-learner ensemble of the above		0.667	0.600	0.656	0.623	0.610
`XGBoostClassifier`		0.610	0.576	0.610	0.586	0.621

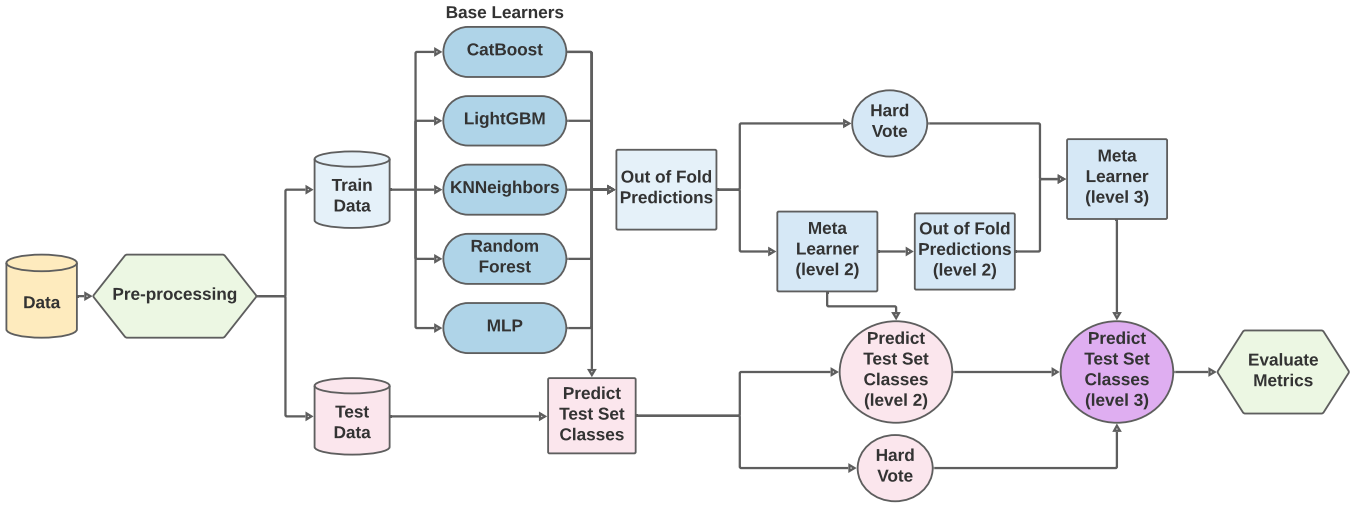

Fig. 6. Flow chart illustrating the overall learning algorithm used for the `ARIADNE` classification pipeline.

`CatBoostClassifier`, and `XGBoostClassifier`). Feature importance provides a general picture of the relative usefulness of each feature in the construction of the resulting classification models. Each of the aforementioned learning algorithms uses a slightly different method to calculate feature importance values. `RandomForestClassifier` calculates the mean decrease in impurity when a feature is used in a split. `CatBoostClassifier` provides several options, from which we select `PredictionValuesChange`, which indicates the average change in the predicted values that result from a change in the feature value. In the case of `XGBoostClassifier`, we opt to use the gain, defined as the improvement in accuracy resulting from the use of a feature in the branches it is on.

In Fig. 7 we show examples of the feature importances resulting from training each of the three aforementioned tree-based learners to select quiescent galaxies in the range $0 < z < 3$, without foreknowledge[10] of galaxy redshifts. Significant differences are evident among the results in terms of the importance values themselves and feature importance rank, reflecting differences among the learning algorithms, the fact that many of the features are strongly correlated, and the somewhat different methods used to compute feature importances.

In general, the broad-band colours typically show some of the highest feature importance values, indicating they are among the most informative, as one would expect given the strong correlation between the shape of a galaxy's SED and its activity type. In addition, the broad-band magnitudes are also clearly useful to some degree. The feature that counts missing values (`n_missing`) also appears to be useful, at least in some circumstances. However, the magnitude errors (not shown) show very low importance values, implying they provide little or no useful information. An important caveat is that feature importance val-

[10] In other words, the input features for the classification models do not contain redshift values, nor are they binned or sorted by redshift. In cases where redshift information is included among the input features, this will be stated.

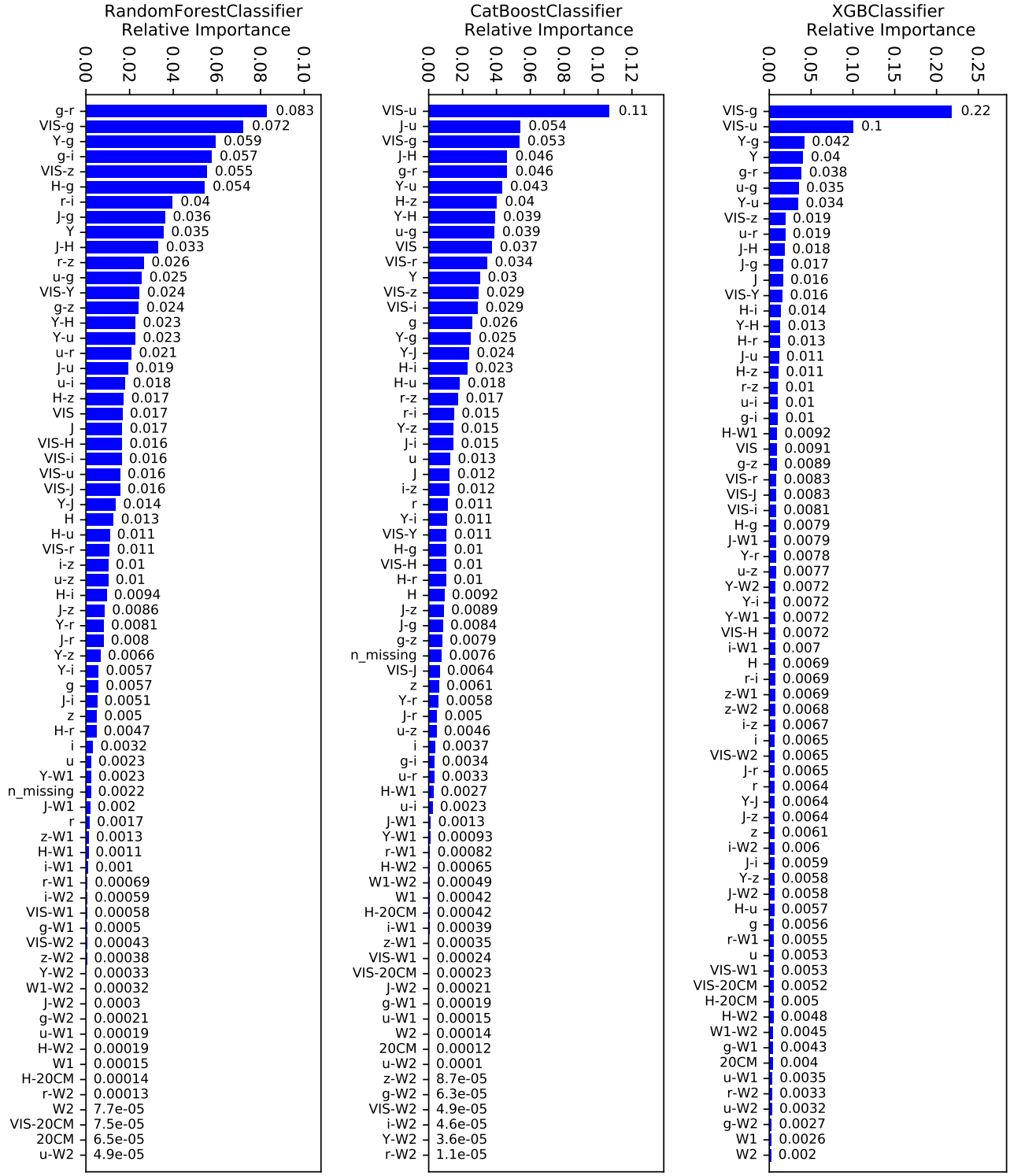

Fig. 7. Examples of feature importance derived from the `RandomForest`, `CatBoost`, and `LightGBM` classifiers when selecting quiescent galaxies from within the $0 < z < 3$ interval, without foreknowledge of galaxy redshifts. For each learner, the feature importance values are normalized such that their sum is 1.0. The y-axis labels correspond to feature names used by the pipeline after the pre-processing steps outlined in Sect. 4.2 have been applied, and should be self-explanatory. For example, the feature named 'u' is derived from the u-band magnitudes, i.e,, after our pre-processing steps; the feature 'VIS-20CM' is derived from the $I_E - 20$ cm colours, etc.

ues do not necessarily give an accurate picture of how the inclusion (or removal) of a particular feature affects the metric used to quantify model performance. Moreover, our feature analysis applies only to the tree-based learning algorithms we have used, and there is no guarantee that it applies to any of the other learning algorithms present in our pipeline.

Thus, to better understand the general usefulness of each of our features, we tested how removing a feature affects the F1-score metric that we calculate after execution of the complete

pipeline. We have conducted A/B tests in which our algorithm is trained and cross-validated twice, once with a particular feature removed, and once more with this feature reinstated. All other parameters were kept constant between the two training runs, including the training/test data split and all random seeds; this A/B test process was repeated multiple times (> 10) for each feature, each time using a new random seed to ensure the A/B test results are not dependent on which random seed is used. The resulting difference in the F1-scores between the two A/B test runs then indicates whether the inclusion of a feature is useful (higher F1-score) or not (lower or unchanged F1-score).

We find that removal of any of the broad-band colours or magnitude values results in a noticeable decrease in the F1-score, indicating they are all useful for training our algorithm. While the broad-band magnitudes technically provide the learning algorithms with a full description of the broad-band SED enclosed by the respective wavelength range, it is clear that explicitly providing spectral slopes in the form of broad-band colours allows significantly more accurate model training.

We also find that inclusion of `n_missing` results in a significant improvement in the F1-score, although the size of the improvement appears to depend somewhat on the classification objective, such as the redshift range of galaxies under selection, and the base learner used. Conversely, removing any (or all) of the magnitude error features results in a small but significant improvement in the F1-score, indicating they are uninformative and merely add noise to the training data.

Therefore, our machine learning pipeline trains on the following features: (i) the magnitudes; (ii) the broad-band colours; (iii) the `n_missing` feature.

5. Results

5.1. Selection in redshift bins

We now apply our classification pipeline to the problem of selecting quiescent galaxies from the mock Euclid photometry catalogues. The objective is to examine the suitability of our method for the separation of quiescent and star-forming galaxies, to facilitate expected Euclid legacy science related to quiescent galaxies. We assume that prior to application of our pipeline, the following steps have been performed: (i) correction for Galactic extinction; (ii) pre-classification into star, active galaxy, and non-active galaxy classes; (iii) photometric or spectroscopic redshifts have been determined, where applicable.

To allow a direct comparison with the colour-colour methods of B20, we select quiescent galaxies in redshift bins delimited by the values $z = 0$, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0, 2.5. We include an additional bin covering the range $2.5 \leq z < 3$. The binning is performed using the 30-band photometric redshifts from Laigle et al. (2016), and assumes that high quality photometric (or spectroscopic) redshifts will be available for all the galaxies. Hereinafter, all redshift values correspond to the 30-band photometric redshifts from Laigle et al. (2016), with the obvious exception of those derived using our pipeline in Sect. 5.2.

A significant fraction of the Euclid survey area is expected to have deep, overlapping ground-based imaging observations from optical surveys (e.g. LSST), but in some areas these observations may be sparse or non-existent. Therefore, we test the performance of our pipeline under the three main expected cases in terms of photometric depth and coverage: (i) Euclid Deep Survey photometry with supporting *ugriz*, *W*1, *W*2, and 20 cm photometry; (ii) Euclid Wide Survey photometry with supporting *ugriz*, *W*1, *W*2, and 20 cm photometry; (iii) Euclid Wide Survey photometry only. Results for these, and additional cases, are shown in Fig. 8 and 9.

5.1.1. Deep survey: *Euclid*, *ugriz*, *W*1, *W*2, 20 cm

When selecting from the Int Deep catalogue using features derived from the Euclid, *ugriz*, WISE, and VLA photometry (blue points, second row of Fig. 8), the F1-score shows a general rise from values of ~ 0.7 at low-z, before peaking at a value of 0.86 in the $1.0 < z < 1.25$ bin and declining towards higher redshifts, reaching a value of 0.48 in the $2.5 < z < 3.0$ bin.

Selecting from the SED Deep catalogue results in a broadly similar F1-curve (blue points, fourth row of Fig. 8), albeit with a broad plateau over the range $0.75 \lesssim z \lesssim 2.0$, with maximum and minimum values of 0.97 and 0.75, respectively. The F1-scores are systematically higher by $\sim$ 0.1–0.3 compared to values obtained in the same redshift bins using the Int Deep catalogue.

Interestingly, at $z \gtrsim 1$ there is only a marginal reduction in F1-score when the selection is performed without the *ugriz*, WISE, and 20 cm photometry. In other words, provided the redshifts are known beforehand, these bands are largely superfluous for the selection of quiescent galaxies, presumably due to the fortuitous positioning of the 4000-Å break within the Euclid broad-band SED.

5.1.2. Wide survey: *Euclid*, *ugriz*, *W*1, *W*2, 20 cm

The situation is broadly similar when selecting from the Wide catalogue using features derived from the Euclid, *ugriz*, WISE and 20 cm photometry (blue points in the first and third rows of Fig. 8). The F1-curve shows a gradual increase from $z = 0$ to a broad peak or plateau at $0.75 \lesssim z \lesssim 2.0$, after which there is a gradual decline towards higher redshifts. In the case of the Int Wide selection, the maximum and minimum F1-scores are 0.87 and 0.42, respectively. For the SED Wide selection, these values are 0.89 and 0.69.

As before, the F1-scores are usually higher when selecting from the SED Wide catalogue compared to the Int Wide catalogue, with values that are up to $\sim$ 0.3 higher. Again, at $z \gtrsim 1$ there is only a marginal reduction in F1-score when the selection is performed without the *ugriz*, WISE, and VLA photometry.

As discussed in Sect. 4.1, when selecting from the $2.5 < z < 3.0$ bin we have used a single base-learner, `XGBoost`, together with the generalised stacking algorithm (see also Fig. B.1, right panel). When selecting quiescent galaxies from the Wide catalogues in this redshift bin, this pipeline setup provides significantly higher F1-scores compared to the default pipeline configuration where five base-learners are employed. Although it is not immediately clear why this is the case, we suggest that its ability to understand which values are missing, and its subsequent use of missing values when performing splits, allows the `XGBoost` algorithm to build stronger classifiers than other learners when there is a high fraction of (informative) missing values in the data set, as is the case here (see Fig. 4).

It is also interesting to note that the Int Wide catalogue contains only 16 quiescent galaxies in the range $2.5 \leq z \leq 3$ with detections in all of the Euclid bands. As such, the training set contains on average 10.7 quiescent galaxies, and the test set 5.3, making this a 'few-shot learning' problem. Remarkably, despite the small number of examples in this redshift band, our pipeline is able to obtain P, R, and, F1-score of ~ 0.43.

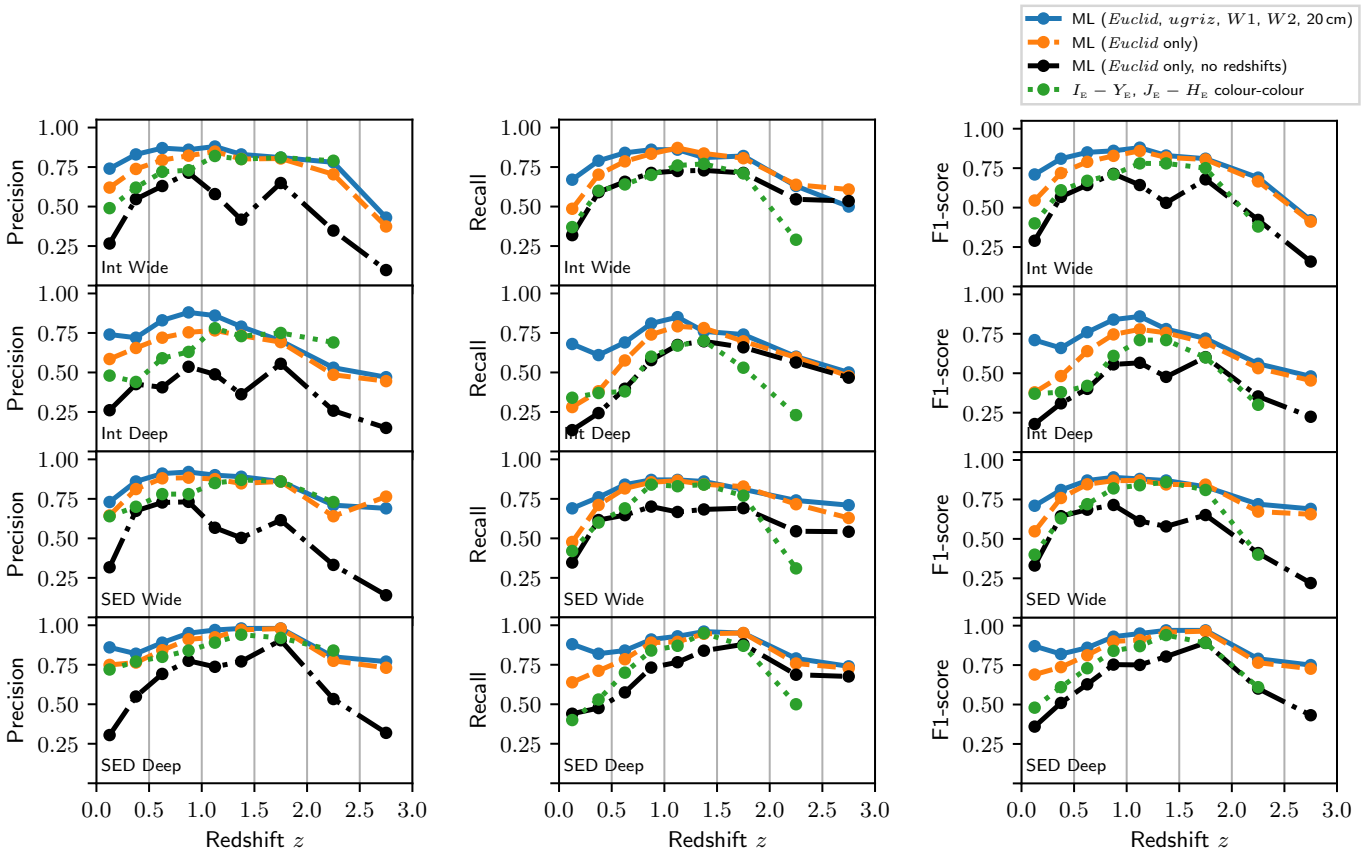

Fig. 8. Precision, recall, and F1-score for various methods of identifying quiescent galaxies. We show results for the three cases discussed in Sect. 5.1: (i) Euclid Deep Survey photometry with supporting *ugriz*, *W*1, *W*2, and 20 cm photometry; (ii) Euclid Wide Survey photometry with supporting *ugriz*, *W*1, *W*2, and 20 cm photometry; (iii) Euclid Wide Survey photometry only. Two curves are shown for the Euclid-only case, corresponding to results obtained either with, or without, foreknowledge of the galaxy redshifts. All other results shown in this figure were obtained assuming foreknowledge of redshifts. In addition, we show the result of applying the $I_E - Y_E$, $J_E - H_E$ colour-colour selection method developed by B20, assuming foreknowledge of (photometric) redshifts. In this and subsequent plots showing metrics vs. redshift, the x-axis represents the 'ground truth' photometric redshifts from COSMOS2015 (Laigle et al. 2016).

5.1.3. Wide survey: *Euclid* only

We also examine the performance of our pipeline when only Euclid observations are available and for which a reliable redshift is not available (black points in Fig. 8). These conditions are likely to pertain to a small, but potentially significant number of galaxies in the Wide survey. In this case, our classification pipeline must learn to place galaxies simultaneously into the correct activity class and into the correct redshift bin.

The quality of the classification varies substantially across the redshift range, with a peak F1-score of 0.71 in the $z = 0.75$–1 bin, and with minima of ~ 0.2–0.3 at either end of the range. The results are largely independent of whether the Int Wide or SED Wide catalogue is used.

Without *ugriz* photometry or redshifts, the classification problem becomes much more challenging, and unsurprisingly the resulting selection is of reduced quality compared to the cases discussed in Sect. 5.1.1 and Sect. 5.1.2. In particular, the F1-scores are consistently lower than those obtained when also using *ugriz* photometry and redshifts, with differences of a factor of two or more occurring near endpoints of the considered redshift range. Thus, a key result is that redshift information allows for a significantly more accurate selection of quiescent galaxies.

5.2. Global selection and redshift estimation

An alternative to selection in bins (Sect. 5.1) is first to perform a global selection of quiescent galaxies, ignoring redshift information, and subsequently derive photometric redshifts for the selected galaxies. In this approach, we set the Target variable to 1 for all quiescent galaxies in the range $0 \leq z \leq 3$. The Target is set to 0 for star-forming galaxies at $z \leq 3$, and also for all galaxies at $z > 3$, regardless of their sSFR. This analysis is performed on two different subsets of the mock data. Casting a relatively wide net, we use all galaxies from the subset of each mock catalogue for which there is a detection in all of the Euclid bands. In addition, the analysis is performed for the subset of galaxies for which there is a detection in each of the *ugriz* and Euclid bands.

The results are shown in Table 3. The F1-score, P, and R metrics can vary significantly depending on which mock catalogue is used and whether galaxies with a non-detection in an optical band are included or rejected. When detections are required in Euclid I_E, Y_E, J_E, and H_E bands only, we obtain $P = 0.85$, $R = 0.75$, and an F1-score of 0.80 for the Int Wide catalogue, or $P = 0.88$, $R = 0.76$, and an F1-score of 0.82 for the SED Wide catalogue. Using the Int Deep catalogue, we obtain the metric values $P = 0.77$, $R = 0.60$, and an F1-score of 0.67, which are significantly lower than those obtained with Int Wide. On the

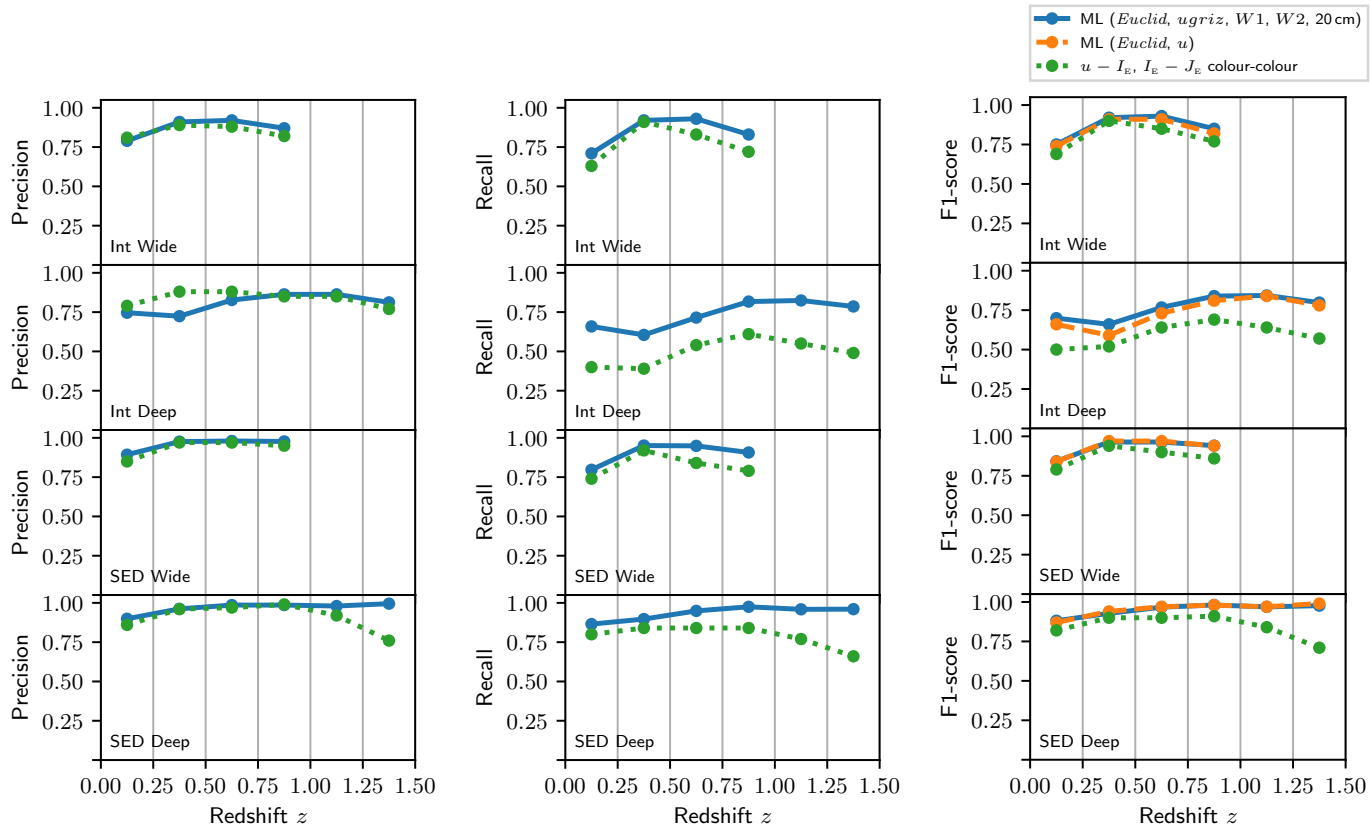

Fig. 9. Precision, recall and F1-score for quiescent galaxy identification methods when applied to the subset of galaxies detected in u, I_E, and J_E, using features derived from the *ugriz*, Euclid, $W1$, $W2$, and (for the Int catalogues) the 20 cm bands. For comparison, we show the F1-scores when only features derived from the u and Euclid bands are used. We also show curves representing the $u - I_E$, $I_E - J_E$ colour-colour selection method proposed by B20.

Table 3. Results from global selection of quiescent galaxies at $0 \leq z \leq 3$, and photometric redshift estimation. The columns are as follows: (1) Mock catalogue; (2) bands in which galaxies are required to be detected; (3) precision P for quiescent galaxy selection; (4) recall R for quiescent galaxy selection; (5) the F1-score for quiescent galaxy selection; (6) normalized median absolute deviation (NMAD) for the photometric redshifts; (7) catastrophic outlier fraction (f_{out}) for the photometric redshifts; (8) bias of the photometric redshifts; (9) fraction of quiescent galaxy redshifts rejected as potential catastrophic outliers (f_{rej}). The typical uncertainties on the P, R, and F1-score values herein are ≤ 0.01.

		classification statistics			photometric redshift statistics			
Catalogue	Detections Required	P	R	F1-score	NMAD	f_{out}	Bias	f_{rej}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Int Wide	I_E, Y_E, J_E, H_E	0.85	0.75	0.80	0.027	0.022	0.0031	0.14
SED Wide	I_E, Y_E, J_E, H_E	0.88	0.76	0.82	0.033	0.017	0.0000	0.22
Wide (averaged)	I_E, Y_E, J_E, H_E	0.87	0.76	0.81	0.030	0.019	0.0016	0.18
Int Deep	I_E, Y_E, J_E, H_E	0.77	0.60	0.67	0.030	0.022	0.0002	0.44
SED Deep	I_E, Y_E, J_E, H_E	0.88	0.79	0.83	0.022	0.015	−0.0008	0.29
Deep (averaged)	I_E, Y_E, J_E, H_E	0.83	0.70	0.75	0.026	0.018	−0.0003	0.37
Int Wide	*ugriz*, I_E, Y_E, J_E, H_E	0.88	0.85	0.86	0.023	0.020	0.0024	0.20
SED Wide	*ugriz*, I_E, Y_E, J_E, H_E	0.96	0.91	0.94	0.024	0.009	0.0008	0.01
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E	0.92	0.88	0.90	0.024	0.014	0.0016	0.10
Int Deep	*ugriz*, I_E, Y_E, J_E, H_E	0.80	0.63	0.70	0.030	0.020	0.0013	0.39
SED Deep	*ugriz*, I_E, Y_E, J_E, H_E	0.96	0.92	0.94	0.020	0.010	0.0000	0.01
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E	0.88	0.78	0.82	0.025	0.015	0.0007	0.20

other hand, when using the SED Deep catalogue, the metrics are $P = 0.88$, $R = 0.79$, and an F1-score of 0.83.

When detection in Euclid I_E, Y_E, J_E, H_E, and all of *ugriz* is required, the metrics are substantially improved, with $P = 0.88$, $R = 0.85$, and an F1-score of 0.86 for the Int Wide catalogue, or $P = 0.96$, $R = 0.91$, and an F1-score of 0.94 for the SED Wide catalogue. These improvements are predominantly due to the fact that requiring a detection in each of the optical and near-IR bands reduces the input data to a substantially smaller subset with relatively well constrained broad-band spectral energy distributions (see Fig. 5 and Table C.1). Under these conditions, for the Int Deep catalogue, we obtain $P = 0.80$, $R = 0.63$, and

an F1-score of 0.70; for SED Deep, $P = 0.96$, $R = 0.92$, and an F1-score of 0.94 are obtained.

To evaluate the quality of the photometric redshift estimates, the 30-band photometric redshifts derived by Laigle et al. (2016) were used as the 'ground truth'. In Table. 3, we also give the values of NMAD, bias, and the catastrophic outlier fraction ($f_{\rm out}$) for the quiescent subset of galaxies whose photometric redshifts were not flagged as outliers by our pipeline.

Removal of likely catastrophic outliers was performed as described in Sect. 4.1, using a relatively stringent probability threshold of 0.15. In other words, we rejected photometric redshift estimates which were assigned a probability of ≥ 0.15 of being a catastrophic outlier. As with most outlier removal methods, the removal of genuine catastrophic outliers usually comes at the cost of also removing cases that are not catastrophic outliers. The fraction of quiescent galaxies whose photometric redshifts were rejected as catastrophic outliers ($f_{\rm rej}$), and thus were not used to calculate NMAD, bias, or $f_{\rm out}$, is also shown.

Like the classification metrics, the metrics of photometric redshift quality show variation depending on which mock catalogue (or subset thereof) is used. We obtain values for NMAD between 0.020 and 0.033, catastrophic outlier fractions between 0.009 and 0.022, and values of bias in the range −0.0008 to 0.0031. While these values appear to improve on the results of the recent Euclid Photometric Redshift Challenge (Euclid Collaboration: Desprez et al. 2020), it is important to recognise that the mock Euclid photometry catalogue used therein has a significantly different construction to those we have used herein, making the intercomparison of photometric redshift metrics potentially unreliable.

6. Comparison with other methods

We test the performance of our quiescent galaxy selection pipeline against four different colour-colour selection methods. We make like-for-like comparisons, such that our pipeline and the colour-colour method under consideration are applied to photometrically identical subsets of the mock catalogues. In summary, our machine learning method outperforms each of the colour-colour methods we tested; full details are given in the following subsections.

6.1. $I_{\rm E} - Y_{\rm E}$, $J_{\rm E} - H_{\rm E}$ selection

To perform a like-for-like comparison with the $I_{\rm E}-Y_{\rm E}$, $J_{\rm E}-H_{\rm E}$ selection method, we select from the mock catalogues those galaxies that have a detection in all of the Euclid bands (i.e., $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$). We then bin the galaxies by redshift as described in Sect. 5.1, and for all bins (except $z = 2.5$–3.0 where there are too few galaxies; see B20), we select quiescent galaxies using the $I_{\rm E} - Y_{\rm E}$, $J_{\rm E} - H_{\rm E}$ criteria given by B20. The results are shown by the green points in Fig. 8.

Our selection pipeline outperforms the $I_{\rm E} - Y_{\rm E}$, $J_{\rm E} - H_{\rm E}$ for almost every combination of redshift and mock catalogue, with the greatest improvements in the F1-score occurring near each endpoint of the considered redshift range. When our pipeline uses all the available photometry (blue line in Fig. 8), the improvement in the F1-score ranges from negligibly small at $z \sim 1.5$, to a factor of ~ 2 in the $z = 0$–0.25 and $z = 2$–2.5 bins.

While the inclusion of additional photometry bands is clearly part of the reason for the improvements we have obtained, it is not the whole story. We find that, even when our pipeline has access to exactly the same photometry bands as the $I_{\rm E}-Y_{\rm E}$, $J_{\rm E}-H_{\rm E}$ method (i.e., all four Euclid bands; orange points in Fig. 8), it still significantly outperforms the colour-colour method. This is because the other Euclid colours ($I_{\rm E}-J_{\rm E}$, $I_{\rm E}-H_{\rm E}$, $Y_{\rm E}-J_{\rm E}$, $Y_{\rm E}-H_{\rm E}$) carry information regarding the sSFR that is not present in $I_{\rm E}-Y_{\rm E}$ or $J_{\rm E} - H_{\rm E}$.

When our pipeline is configured to select quiescent galaxies in the same set of redshift bins, using Euclid photometry only, and in the absence of redshift information (black points in Fig. 8,

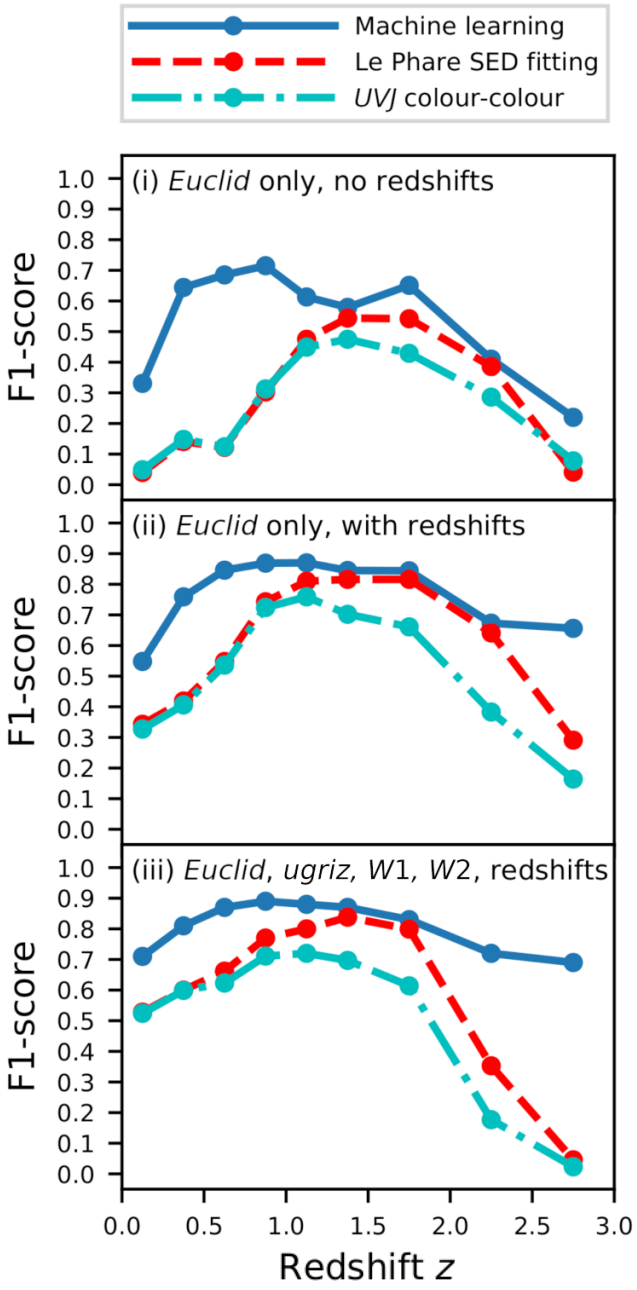

Fig. 10. Comparison between our machine learning method, `LePhare` SED fitting, and UVJ colour-colour method, for the selection of quiescent galaxies from the SED Wide mock Euclid catalogue. The following three configurations of input data were used: (i) Euclid photometry only, with no redshift information; (ii) Euclid photometry only, this time with photometric redshifts from Laigle et al. (2016); (iii) Euclid, *ugriz*, $W1$ and $W2$ photometry, again using photometric redshifts from Laigle et al. (2016).

we find that, in roughly half of the redshift bins, the resulting F1-scores are similar to those obtained using the $I_E - Y_E$, $J_E - H_E$ method (which has the benefit of foreknowledge of galaxy redshifts). This is generally the case for $z \lesssim 1$, while in the range $1 \lesssim z \lesssim 1.75$ our pipeline yields significantly lower F1-scores in this configuration. Interestingly, the F1-scores obtained using our pipeline to select quiescent galaxies in the $z = 2$–2.5 bin are practically identical to, or slightly above, those obtained using the $I_E - Y_E$, $J_E - H_E$ method.

6.2. $u - I_E$, $I_E - J_E$ selection

For our comparison with the $u - I_E$, $I_E - J_E$ method, we use only galaxies that are detected in each of the u, I_E, and J_E bands. Due to the increasing sparsity of the u-band data at high redshifts, B20 were only able to derive $u - I_E$, $I_E - J_E$ selection criteria in redshift bins delimited by the values $z = 0$, 0.25, 0.5, 0.75, 1.0 in the case of the Wide survey mock catalogues, or $z = 0$, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5 for the Deep catalogues. To these bins we have applied our quiescent galaxy selection pipeline. The results are shown in Fig. 9, together with the result of applying the $u - I_E$, $I_E - J_E$ method of B20.

Once again, our pipeline (blue or orange points in Fig. 9) outperforms the colour-colour selection method (green points). The improvement is minimal when using the Wide survey mock catalogues, but is more substantial in the case of the Deep catalogues (up to a factor of ~ 1.4), mainly due to improved recall. The largest improvements occur at the upper endpoint of the considered redshift ranges. There is little to no reduction in the performance of our pipeline when only the u, I_E, Y_E, J_E, and H_E bands are used (orange points in Fig. 9), compared to when the full suite of photometry is used (blue points in Fig. 9). Once again, the improvement obtained from using our machine learning pipeline is largely due to its ability to make use of a larger colour- and magnitude-space.

6.3. *BzK* selection

In addition, we compare the quality of our quiescent galaxy selection method against the *BzK* selection method, which is designed to select quiescent galaxies at $1.4 < z < 2.5$ using the $z - K$ and $B - z$ broad band colours (Daddi et al. 2004). For this comparison, we have used the observed B, z and K magnitudes from the COSMOS2015 catalogue, for galaxies in the Int Wide catalogue, adopting the criteria of Daddi et al. (2004) of $(z - K) - (B - z) < -0.2$ and $z - K > 2.5$ to select quiescent galaxies. To evaluate the success of the *BzK* method, we use the sSFR values from the Int Wide catalogue, adopting the same $\mathrm{sSFR} < 10^{-10.5}\mathrm{yr}^{-1}$ threshold to define quiescence.

First, we take the subset of the Int Wide catalogue that lies in the redshift range $1.4 < z < 2.5$, and apply the *BzK* selection method. We obtain P, R, and F1-score values of 0.51, 0.50 and 0.50, respectively. Applying the `ARIADNE` pipeline to the same subset of galaxies, using features derived from the Euclid, *ugriz*, $W1$, $W2$, and 20 cm bands, we obtain P, R, and F1-score values of 0.79, 0.80, and 0.79, respectively. Repeating the same tests over the full redshift range of the Int Wide catalogue, the *BzK* method results in P, R, and F1-score of 0.44, 0.07, and 0.13, compared to 0.85, 0.75, 0.80 from `ARIADNE`. In summary, our machine learning method strongly outperforms the *BzK* selection method.

6.4. SED fitting with `LePhare` and UVJ selection

We also tested, though not exhaustively, our selection pipeline against an SED fitting method wherein the galaxies are separated into quiescent and star-forming on the basis of the sSFR estimated from the `LePhare` best fit to the mock photometry. The fitting procedure is identical to that described in Sect. 2, except that the fitting is performed on SED Wide mock photometry, instead of on observational data. Three slightly different approaches are taken, as follows: (i) Only Euclid photometry is used, with redshift being a free parameter; (ii) only Euclid photometry is used, but redshift is fixed to the value given in Laigle et al. (2016); (iii) Euclid, *ugriz*, $W1$, and $W2$ photometry is used, including upper limits, with redshift fixed. We derive rest-frame UVJ colours from the best-fitting spectral templates, and use the UVJ selection criteria of Whitaker et al. (2011). The approach used here may overestimate how well `LePhare` and the UVJ method can recover the galaxy class, since the creation of the SED Wide mock catalogue and our subsequent fitting were in both cases performed using `LePhare`, with an identical set of base templates.

The F1-scores as a function of redshift are shown in Fig. 10 for the `LePhare` fitting and UVJ methods. Also shown is the F1-score for our machine learning method when applied to the same subset of data, with identical redshift information.

Our machine learning method significantly outperforms both approaches, most noticeably at $z < 1$ and $z \gtrsim 2.5$. While the `LePhare` SED fitting method sometimes comes close to reaching the F1-scores obtained by our machine learning method within the $z \sim 1$–2.5 range, in most redshift bins its F1-scores are dramatically lower; in the case of the UVJ method, the F1-scores are always dramatically lower, typically by ~ 0.2, than those obtained with our machine learning method. The superior performance of our method is at least partly due its ability to learn how to optimally weight the different bands and colours in different regions of feature-space, unlike the `LePhare` fitting and UVJ methods.

7. Further analysis and tests

Here we sumarize some additional analysis and tests we have conducted. Full details are provided in Appendix B.

7.1. Stacking vs. individual learners

Our implementation of the generalized stacking method demonstrably improved classification performance:

- With few exceptions, the stacking method consistently outperforms each individual base-learner, as well as outperforming model averaging and hard-voting (Fig. B.1);
- The method is robust against pollution by multiple low-quality classifier models, and can be used as a form of model selection;
- When applied to a single classifier model, the meta-learner often makes a substantial improvement over the original model.

7.2. The nature of the false positives

We have also examined the distribution of false positives within sSFR space, with the following main conclusions:

- As expected, the incorrectly classified objects cluster around the class threshold value ($10^{-10.5}\mathrm{yr}^{-1}$), with a density that is

highest in the bins immediately adjacent to the class boundary (Figs. B.2 and B.3);
- The precise distribution of the incorrect classifications differs between the different mock catalogues;
- Our pipeline offers a significant improvement in $\bar{I}_{\rm FP}$ over the $I_{\rm E} - Y_{\rm E}$, $J_{\rm E} - H_{\rm E}$ and $u - I_{\rm E}$, $I_{\rm E} - J_{\rm E}$ colour-colour methods, reducing the degeneracy between quiescent galaxies and dusty, star-forming galaxies as shown in Fig. B.4 (in addition to improving on the P, R, and the F1-score metrics as described above).

7.3. Reconciling the Int and SED results

Generally speaking, our classification results differ depending on whether the mock catalogue was constructed using the Int or SED method. Firstly, we argue that results obtained using the Int catalogues are likely to be pessimistic with regard to the performance of our pipeline when applied to real Euclid data; this is because the Euclid photometry is expected to have significantly higher signal-to-noise ratios, and because the method of interpolating photometry to simulate the Euclid bands is also likely to introduce errors. Conversely, we expect that results obtained with the SED catalogues are likely to be somewhat optimistic, because the construction of the SED catalogues involves forcing the photometry to conform to one of a limited range of galaxy templates.

Thus, we argue that results obtained with the Int and SED catalogues will bracket the real-Universe performance of our pipeline. As a result, we show performance metrics averaged over the Int catalogue and the corresponding SED catalogue (i.e., Int Wide and SED Wide; Int Deep and SED Deep) in Tables 3 and 4, where appropriate.

7.4. Tuning the probability threshold

We investigated the impact on the precision and recall of tuning the value of the class probability threshold, instead of adopting the default threshold value of 0.5 (see Fig. B.5). Our main findings are:

- There exists a trade-off between precision and recall such that one may be increased, but at the cost of reducing the other; tuning the probability threshold allows a balance to be struck between P and R that is suitable for different scientific needs;
- Using the case shown in Fig. B.5 (bottom panel) as an example, adopting a probability threshold of 0.85 yields a very pure sample of quiescent galaxies (P = 0.98) but with moderate incompletness (R = 0.56); conversely, a probability threshold of 0.05 results in high completeness (R = 0.97) but moderate purity (P = 0.61).

7.5. Impact of including redshift as a feature

We have examined the usefulness of including the Laigle et al. (2016) COSMOS2015 photometric redshifts as a feature in the data used for model training (e.g. Simet et al. 2021). Selected results are included in Table 4, and are shown in Fig. B.6. Our main findings are:

- Inclusion of redshifts as a feature in the data significantly improves the classification metrics by reducing the degeneracy between redshift and sSFR, particularly for galaxies at $z \leq 0.5$ or $z \geq 2.5$;
- Even when only a subset of the redshifts are included, the classification metrics are still improved compared to the case where no redshift are included; in other words, it is beneficial to include any available redshift information, even if it is somewhat sparse, when training classification models.

7.6. The impact of noise

The impact of different types of noise has been explored. The main results are summarized below:

- As expected, adding extra noise to the photometric data results in a reduction in the F1-score when selecting quiescent galaxies.
- However, even when the data becomes extremely noisy (S/N $\sim$ 3), our classification pipeline remains nominally functional with an F1-score of $\sim$ 0.67 (see Fig. B.7).
- Our classification pipeline is robustly resistant to labelling errors when these are random; however, systematic labelling error tend to propagate into the final classification results.
- When photometric redshifts are included as a feature, adding Gaussian noise to their values generally results in little or no reduction in the classification metrics; we find that including noisy redshift values still gives generally better classification results compared to the case where no redshifts are included.

7.7. Transfer learning experiments

7.7.1. Training on templates and predicting on real SEDs

We have also experimented with using classification models trained on spectral templates to select quiescent galaxies from catalogues of observed photometry (see Fig. B.8). We find that classifiers trained on the SED Wide catalogue are indeed able to select quiescent galaxies from the Int Wide catalogue, albeit with marginally lower F1-scores compared to models trained on the Int Wide catalogue itself. Thus, machine learning models trained on synthetic galaxy SEDs are a potential alternative to traditional methods used selecting quiescent galaxies at redshifts where there are few (or no) known examples (e.g. Girelli, Bolzonella, & Cimatti 2019; Cecchi et al. 2019).

7.8. Which observables are useful to select quiescent galaxies?

We have investigated the impact of adding one of u,g,r,i,z, $W1$, $W2$, and 20 cm to the Euclid bands, when selecting quiescent galaxies. The main results are summarized as follows:

- Generally speaking, this results in a significant improvement in the F1-score when selecting quiescent galaxies. The improvement varies depending on which band is added, and the redshift interval in which the selection is performed (Figs. B.9, B.10, B.11, B.12).
- As expected, the addition of a longer-wavelength optical band is typically more useful for selecting quiescent galaxies at higher redshift; conversely, shorter wavelength optical bands are more useful for low-redshift selection.
- Interestingly, the 20 cm radio band provides a significant improvement in F1-score at $1.75 \lesssim z \lesssim 2.5$, despite this data being very sparse.

Table 4. Global selection of quiescent galaxies at $0 \leq z \leq 3$ for different pipeline and data configurations. No pre-binning of the data by redshift was performed. The columns are as follows: (1) Mock catalogue; where the metrics from using the SED and Int catalogues have been averaged, this is indicated in parentheses; where the pipeline has been used in its *fast mode*, this is also indicated; in the case of transfer-learning, we specify separately the catalogues from which the training and test data are drawn; (2) bands used for galaxy selection, which includes missing values; (3) bands in which galaxies are required to be detected; (4) redshift information included in the input data; (none, or a percentage of redshifts included in a feature); (5) precision P for quiescent galaxy selection; (6) recall R for quiescent galaxy selection; (7) the F1-score for quiescent galaxy selection. The typical uncertainties on the P, R, and F1-score values herein are ≤ 0.01.

Catalogue (1)	Bands Used (2)	Detections Required (3)	Redshifts (4)	P (5)	R (6)	F1-score (7)
Int Wide	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.85	0.75	0.80
Int Wide (*fast mode*)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.84	0.74	0.79
SED Wide	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	none	0.88	0.76	0.82
SED Wide (*fast mode*)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	none	0.88	0.75	0.81
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	*ugriz*, I_E, Y_E, J_E, H_E	none	0.92	0.88	0.90
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	*ugriz*, I_E, Y_E, J_E, H_E	100%	0.92	0.91	0.92
Wide (averaged)	I_E, Y_E, J_E, H_E	I_E, Y_E, J_E, H_E	none	0.81	0.68	0.74
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.87	0.76	0.81
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	50%	0.86	0.79	0.82
Wide (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	100%	0.86	0.83	0.85
Wide (averaged, $\sigma_z = 0.025$)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	100%	0.86	0.83	0.84
Wide (averaged, $\sigma_z = 0.05$)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	100%	0.86	0.82	0.84
Wide (averaged, $\sigma_z = 0.075$)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	100%	0.86	0.81	0.83
Int Deep	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.77	0.60	0.67
Int Deep (*fast mode*)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.77	0.59	0.67
SED Deep	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	none	0.88	0.79	0.83
SED Deep (*fast mode*)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	none	0.88	0.78	0.83
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	*ugriz*, I_E, Y_E, J_E, H_E	none	0.88	0.78	0.82
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	*ugriz*, I_E, Y_E, J_E, H_E	100%	0.89	0.85	0.87
Deep (averaged)	I_E, Y_E, J_E, H_E	I_E, Y_E, J_E, H_E	none	0.73	0.54	0.62
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.83	0.70	0.75
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	50%	0.83	0.74	0.78
Deep (averaged)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	100%	0.85	0.81	0.83
SED Wide (train), Int Wide (test)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	100%	0.76	0.86	0.81
Int Wide (train), SED Wide (test)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	100%	0.89	0.76	0.82
Int Deep (train), Int Wide (test)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$, 20 cm	I_E, Y_E, J_E, H_E	none	0.76	0.72	0.74
SED Deep (train), SED Wide (test)	*ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$	I_E, Y_E, J_E, H_E	none	0.71	0.83	0.76

8. Summary and concluding remarks

We have introduced the `ARIADNE` machine learning pipeline for the classification of galaxies, which uses a novel architecture with meta-learning to combine the strengths of tree, nearest-neighbours, and deep-learning methods, resulting in significantly higher classification accuracy compared to the individual learning algorithms. The most relevant technical conclusions from this study are as follows:

- We have applied the tree-ensemble methods `CatBoostClassifier` and `LightGBM` to the selection of quiescent galaxies, and find that both offer significant performance improvements over the commonly-used `Scikit-Learn RandomForrestClassifier` method, in terms of model quality and training efficiency.
- Providing our pipeline with sparsity-awareness, by quantifying the sparsity of the photometry for each galaxy, has been found to improve classification performance. In addition, the sparsity-aware method `XGBoostClassifier` was found to be well suited for selecting quiescent galaxies at high redshifts ($z > 2.5$), which tend to have many missing photometry values.
- We have shown that our implementation of the 'generalised stacking' method can be used to perform error-correction on individual machine learning based galaxy classification models, sometimes turning a mediocre model into a significantly better one.
- We have used the pseudo-labelling technique Lee (2013) to improve the quality of our photometric redshift estimates, with improvements in NMAD and the catastrophic outlier fraction of $\sim$ 1–3 per cent. When applied to the high-volume of data that will come from future very large surveys such as Euclid or LSST, we expect pseudo-labelling to have a much greater effect. Further exploration of the application of pseudo-labelling to the estimation of galaxy physical properties is presented in Humphrey et al. (2022).

We have applied our pipeline to the selection of quiescent galaxies from mock Euclid photometric catalogues, using simulated Euclid I_E, Y_E, J_E, H_E photometry, optionally using ancillary optical, infrared or radio measurements. The main results are as follows:

- We have shown that our classification pipeline is able to efficiently select quiescent galaxies from within the redshift range $0 < z < 3$, using mock Euclid I_E, Y_E, J_E, and H_E photometry and somewhat sparse supporting data at other wavelengths. The precision (purity), recall (completeness) and F1-scores vary substantially with redshift, and between the various mock catalogues and subsets thereof.
- We find that including ancillary *ugriz*, mid-infrared (WISE) and radio (20 cm) photometry yields substantial improvement in the selection of quiescent galaxies at $z \lesssim 1$. Smaller, but nonetheless significant, improvements were found at $z \gtrsim 1$.

- In like-for-like comparisons, our machine learning pipeline strongly outperforms the UVJ method (Whitaker et al. 2011) when derived from Euclid (and ancillary) survey mock photometry, and usually outperforms the Euclid-specific $I_E - Y_E$, $J_E - H_E$, and $u - I_E$, $I_E - J_E$ colour-colour selection methods (B20). The improvement we obtain over the colour-colour methods can exceed a factor of 2 in terms of completeness and F1-score, with the greatest improvements occurring at $z \lesssim 1$ and $z \gtrsim 2$.
- In addition to being fewer in number, the false-positives resulting from our classification pipeline are less extreme than those resulting from the $I_E - Y_E$, $J_E - H_E$, and $u - I_E$, $I_E - J_E$ methods, in that their actual sSFR values are typically closer to the boundary between quiescent and star-forming galaxies.
- The significantly improved classification compared to the colour-colour, UVJ, and template-fitting methods is likely the result of the more efficient use of the available data by our machine learning methodology. Compared to the colour-colour methods, the ability to perform the classification in a higher-dimensional colour-magnitude space clearly helps. More generally, machine learning methodologies have the ability to automatically weight the different colours and filters according to their usefulness for the classification task at hand, whereas traditional methods often take a somewhat 'blind' approach to data weighting.
- Our pipeline is able to derive photometric redshifts for galaxies selected as quiescent, aided by the 'pseudo-labelling' semi-supervised method, also using an outlier detection algorithm to identify and reject likely catastrophic outliers. Our pipeline achieves a normalized mean absolute deviation of $\lesssim 0.03$ and a fraction of catastrophic outliers of $\lesssim 0.02$ when measured against the COSMOS2015 photometric redshifts of Laigle et al. (2016). These values appear to improve on the results of the Euclid Photometric Redshift Challenge (Euclid Collaboration: Desprez et al. 2020), but we emphasize that the mock photometry catalogue used therein is of a significantly different construction to those we have used herein, making any intercomparison of photometric redshift metrics potentially unreliable.
- The inclusion of galaxy redshifts among the train and test datasets offers significant improvement in the quality of our classification models, even when the redshifts are relatively noisy or incomplete.
- We have investigated the potential impact of various systematics. Most notably, we find that our pipeline results are robust against the presence of random errors in the class labels of the training data, for label error rates of up to $\sim$ 33 per cent.

This work has added to the growing body of evidence supporting the importance of machine learning techniques (or artificial intelligence) in astronomy and astrophysics. In particular, we have demonstrated that machine learning usually outperforms colour-colour methods for the selection of quiescent galaxies; while part of this improvement is due to the ability to make use of a larger number of bands and colours, we have also shown that machine learning methods still perform a superior selection even when the training data contains only the bands used by the respective colour-colour method. In future publications the methods presented herein will be further developed and applied to other related problems in extragalactic astrophysics.

Acknowledgments

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through grants UID/FIS/04434/2019, UIDB/04434/2020, UIDP/04434/2020 and PTDC/FIS-AST/29245/2017, and an FCT-CAPES Transnational Cooperation Project. LB acknowledges financial support by the Agenzia Spaziale Italiana (ASI) under the research contract 2018-31-HH.0. KIC acknowledges funding from the European Research Council through the award of the Consolidator Grant ID 681627-BUILDUP. AH acknowledges support from the NVIDIA Academic Hardware Grant Program. AH also thanks colleagues Jean Gomes, João Pedroso, Catarina Lobo, Tom Scott, Ana Afonso, Patricio Lagos, Israel Matute, Stergios Amarantidis, Jose Afonso, Rodrigo Carvajal, and Ciro Pappalardo for useful discussions or comments. 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 Academy of Finland, the Agenzia Spaziale Italiana, the Belgian Science Policy, the Canadian Euclid Consortium, the 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 Ministerio de Economia y Competitividad, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, 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`). In the development of our pipeline, we have made use of the Pandas (McKinney 2010), Numpy (Harris et al. 2020), Scipy (Virtanen et al. 2020) and Dask (Rocklin 2015) packages for Python.

References

Akeson, R., Armus, L., Bachelet, E., et al. 2019, ArXiv e-prints [arXiv:1902.05569]
Alexandroff, R., Strauss, M. A., Greene, J. E., et al. 2013, MNRAS, 435, 3306
Amorín, R., Pérez-Montero, E., Contini, T., et al. 2015, A&A, 578, A105
Arnouts, S., Walcher, C. J., Le Fèvre, O., et al. 2007, A&A, 476, 137
Arnouts, S., Le Floc'h, E., Chevallard, J., et al. 2013, A&A, 558, A67
Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. 2004, ApJ, 600, 681
Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87
Belli, S., Newman, A. B., & Ellis, R. S. 2015, ApJ, 799, 206
Bisigello, L., Kuchner, U., Conselice, C. J., et al. 2020, MNRAS, 494, 2337 (B20)
Bisigello, L., Gruppioni, C., Feltre, A., et al. 2021, A&A, 651, A52
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Bonjean, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137
Bowles, M., Scaife, A. M. M., Porter, F., Tang, H., & Bastien, D. J. 2021, MNRAS, 501, 4579
Breiman, L. 2001, Mach. Learn., 45, 1
Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140
Bretonnière, H., Boucaud, A., & Huertas-Company, M. 2021, ArXiv e-prints [arXiv:2111.15455]
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Cavuoti, S., Tortora, C., Brescia, M., et al. 2017, MNRAS, 466, 2039
Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2014, MNRAS, 437, 968
Cecchi, R., Bolzonella, M., Cimatti, A., & Girelli, G. 2019, ApJ, 880, L14
Chambers, K., Unions Team Including Pan-Starrs Team, & Cfis Team 2020, American Astronomical Society meeting 235, id. 154.04. Bulletin of the American Astronomical Society, Vol. 52, No. 1
Chen, T., Guestrin, C., 2016, ArXiv e-prints [arXiv:1603.02754v3]
Cirasuolo, M., Fairley, A., Rees, P., et al. 2020, The Messenger, 180, 10

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, A&A, 639, A84
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Cropper, M., Pottinger, S., Niemi, S., et al. 2016, Proc. SPIE, 9904, 99040Q
Cunha, P. A. C. & Humphrey, A. 2022, A&A, accepted, [arXiv:2204.02080]. doi:10.1051/0004-6361/202243135
Daddi, E., Cimatti, A., Renzini, A., et al. 2004, ApJ, 617, 746
Delli Veneri, M., Cavuoti, S., Brescia, M., Longo, G., & Riccio, G. 2019, MNRAS, 486, 1377
Deshmukh, S., Caputi, K. I., Ashby, M. L. N., et al. 2018, ApJ, 864, 166
Dewdney, P. E., Hall, P. J., Schilizzi, R. T., & Lazio, T. J. L. W. 2009, IEEE Proceedings, 97, 1482
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Dice, L. R., 1945, Ecology, 26(3), 297
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2018, MNRAS, 476, 3661
Euclid Collaboration: Desprez, G., Paltani, S., et al. 2020, A&A, 644, A31
Euclid Collaboration: Moneti, A., McCracken, H. J., et al. 2022, A&A, 658, A126
Euclid Collaboration, Scaramella, R., Amiaux, J., et al. 2022, A&A, 662, A112
Euclid Collaboration, Schirmer, M., Jahnke, K., et al. 2022, A&A, 662, A92
Fang, J. J., Faber, S. M., Koo, D. C., et al. 2018, ApJ, 858, 100
Fotopoulou, S., Salvato, M., Hasinger, G., et al. 2012, ApJS, 198, 1
Fotopoulou, S., & Paltani, S. 2018, A&A, 619, A14
Fumagalli, M., Labbé, I., Patel, S. G., et al. 2014, ApJ, 796, 35
Girelli, G., Bolzonella, M., & Cimatti, A. 2019, A&A, 632, A80
Glazebrook, K., Schreiber, C., Labbé, I., et al. 2017, Nature, 544, 71
Gomes, J. M., & Papaderos, P. 2017, A&A, 603, A63
Guarneri, F., Calderone, G., Cristiani, S., et al. 2021, MNRAS, 506, 2471
Guiglion, G., Battistini, C., Bell, C. P. M., et al. 2019, The Messenger, 175, 17
Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040
Haro, G. 1956, Boletin de los Observatorios Tonantzintla y Tacubaya, 2, 8
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Humphrey, A., Cunha, P. A. C., Paulino-Afonso, A., Amarantidis, S., Carvajal, R., Gomes, J. M., Matute, M., Papaderos, P. 2022, MNRAS, submitted
Ilbert, O., Salvato, M., Le Floc'h, E., et al. 2010, ApJ, 709, 644
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Advances in Neural Information Processing Systems, vol. 30, 3146
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Lee, D., "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks." ICML 2013 Workshop: Challenges in Representation Learning (WREPL), Atlanta, Georgia, USA, 2013 [eprint]
Leja, J., Tacchella, S., & Conroy, C. 2019, ApJ, 880, L9
Logan, C. H. A., & Fotopoulou, S. 2020, A&A, 633, A154
Maciaszek, T., Ealet, A., Jahnke, K., et al. 2016, Proc. SPIE, 9904, 99040T
McKinney, w., 2010, Data Structures for Statistical Computing in Python, in Proceedings of the 9th Python in Science Conference, pp 51
Moresco, M., Pozzetti, L., Cimatti, A., et al. 2013, A&A, 558, A61
Mucesh, S., Hartley, W. G., Palmese, A., et al. 2021, MNRAS, 502, 2770
Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJS, 206, 8
Nolte, A., Wang, L., Bilicki, M., Holwerda, B., & Biehl, M. 2019, ESANN 2018 Special Issue of Neurocomputing; Neurocomputing 342: 172-190 [arXiv:1903.07749]
Odewahn, S. C., Humphreys, R. M., Aldering, G., & Thurmes, P. 1993, PASP, 105, 1354
Pasquet, J., Bertin, E., Treyer, M., et al. 2019, A&A, 621, A26
Pedregosa, F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Predehl, P., Andritschke, R., Arefiev, V., et al. 2021, A&A, 647, A1
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018, Advances in Neural Information Processing Systems, 31, 6638
Razim, O., Cavuoti, S., Brescia, M., Riccio, G., Salvato, M., Longo, G., 2021, MNRAS, 507, 5034
Rocklin, M., 2015, Dask: Parallel Computation with Blocked Algorithms and Task Scheduling, in Proceedings of the 14th Python in Science Conference, pp 130
Schreiber, C., Glazebrook, K., Nanayakkara, T., et al. 2018, A&A, 618, A85
Shahidi, A., Mobasher, B., Nayyeri, H., et al. 2020, ApJ, 897, 44
Simet, M., Chartab, N., Lu, Y., & Mobasher, B. 2021, ApJ, 908, 47
Singal, J., Silverman, G., Jones, E., et al. 2022, ApJ, 928, 6
Sørensen T. 1948, Kongelige Danske Videnskabernes Selskab, 5(4): 1
Steinhardt, C. L., Weaver, J. R., Maxfield, J., et al. 2020, ApJ, 891, 136
Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
Taylor, W., Cirasuolo, M., Afonso, J., et al. 2018, Proc. SPIE, 10702, 107021G
Tuccillo, D., Huertas-Company, M., Decencière, E., et al. 2018, MNRAS, 475, 894
Ulmer-Moll, S., Santos, N. C., Figueira, P., Brinchmann, J., & Faria, J. P. 2019, A&A, 630, A135
van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9(86), 2579
van der Maaten L. 2014, Journal of Machine Learning Research, 15(93), 3221
van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, ApJ, 788, 28
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261
Whitaker, K. E., Labbé, I., van Dokkum, P. G., et al. 2011, ApJ, 735, 86
Wiklind, T., Dickinson, M., Ferguson, H. C., et al. 2008, ApJ, 676, 781
Williams, R. J., Quadri, R. F., Franx, M., van Dokkum, P., & Labbé, I. 2009, ApJ, 691, 1879
Wolpert, D.H, 1992, Neural Networks, 5(2), 241-259
Worthey, G. 1994, ApJS, 95, 107
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Wu, P.-F., van der Wel, A., Gallazzi, A., et al. 2018, ApJ, 855, 85
Wuyts, S., Labbé, I., Franx, M., et al. 2007, ApJ, 655, 51
York, D. G., Adelman, J., Anderson, J. E., et al. 2000, AJ, 120, 1579
Zitlau, R., Hoyle, B., Paech, K., et al. 2016, MNRAS, 460, 3152

[1] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[2] DTx – Digital Transformation CoLAB, Building 1, Azurém Campus, University of Minho, 4800-058 Guimarães, Portugal
[3] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, I-40129 Bologna, Italy
[4] Faculdade de Ciências da Universidade do Porto, Rua do Campo de Alegre, 4150-007 Porto, Portugal
[5] School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, UK
[6] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[7] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy
[8] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisboa, Portugal
[9] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, PT-1349-018 Lisboa, Portugal
[10] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, I-40129 Bologna, Italy
[11] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK
[12] INFN-Bologna, Via Irnerio 46, I-40126 Bologna, Italy
[13] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, D-85748 Garching, Germany
[14] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany
[15] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, I-10025 Pino Torinese (TO), Italy
[16] Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, I-00146 Rome, Italy
[17] INFN-Sezione di Roma Tre, Via della Vasca Navale 84, I-00146, Roma, Italy
[18] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, I-10125 Torino, Italy
[19] INFN-Sezione di Torino, Via P. Giuria 1, I-10125 Torino, Italy
[20] INAF-IASF Milano, Via Alfonso Corti 12, I-20133 Milano, Italy
[21] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[22] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
[23] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain
[24] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain

[25] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, I-00078 Monteporzio Catone, Italy
[26] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, I-80126, Napoli, Italy
[27] INFN section of Naples, Via Cinthia 6, I-80126, Napoli, Italy
[28] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Universitá di Bologna, Viale Berti Pichat 6/2, I-40127 Bologna, Italy
[29] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50125, Firenze, Italy
[30] Centre National d'Etudes Spatiales, Toulouse, France
[31] Institut national de physique nucléaire et de physique des particules, 3 rue Michel-Ange, 75794 Paris Cédex 16, France
[32] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[33] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[34] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
[35] ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[36] Univ Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, F-69622, Villeurbanne, France
[37] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[38] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[39] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, PT-1749-016 Lisboa, Portugal
[40] Department of Astronomy, University of Geneva, ch. dÉcogia 16, CH-1290 Versoix, Switzerland
[41] Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405, Orsay, France
[42] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[43] INFN-Padova, Via Marzolo 8, I-35131 Padova, Italy
[44] AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, F-91191 Gif-sur-Yvette, France
[45] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy
[46] FRACTAL S.L.N.E., calle Tulipán 2, Portal 13 1A, 28231, Las Rozas de Madrid, Spain
[47] Aix-Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France
[48] Istituto Nazionale di Astrofisica (INAF) - Osservatorio di Astrofisica e Scienza dello Spazio (OAS), Via Gobetti 93/3, I-40127 Bologna, Italy
[49] Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, I-40126 Bologna, Italy
[50] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, I-35122 Padova, Italy
[51] Dipartimento di Fisica "Aldo Pontremoli", Universitá degli Studi di Milano, Via Celoria 16, I-20133 Milano, Italy
[52] INAF-Osservatorio Astronomico di Brera, Via Brera 28, I-20122 Milano, Italy
[53] INFN-Sezione di Milano, Via Celoria 16, I-20133 Milano, Italy
[54] Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, N-0315 Oslo, Norway
[55] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA
[56] von Hoerner & Sulger GmbH, SchloßPlatz 8, D-68723 Schwetzingen, Germany
[57] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[58] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland
[59] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, 00014 University of Helsinki, Finland
[60] NOVA optical infrared instrumentation group at ASTRON, Oude Hoogeveensedijk 4, 7991PD, Dwingeloo, The Netherlands
[61] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[62] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, I-40127 Bologna, Italy
[63] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, DH1 3LE, UK
[64] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice cedex 4, France
[65] Institut d'Astrophysique de Paris, 98bis Boulevard Arago, F-75014, Paris, France
[66] Sorbonne Universités, UPMC Univ Paris 6 et CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France
[67] University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland
[68] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
[69] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, DK-8000 Aarhus C, Denmark
[70] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France
[71] Institute of Space Science, Bucharest, Ro-077125, Romania
[72] Dipartimento di Fisica e Astronomia "G.Galilei", Universitá di Padova, Via Marzolo 8, I-35131 Padova, Italy
[73] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
[74] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy
[75] INFN-Sezione di Roma, Piazzale Aldo Moro, 2 - c/o Dipartimento di Fisica, Edificio G. Marconi, I-00185 Roma, Italy
[76] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[77] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, 30202 Cartagena, Spain
[78] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[79] Université Paris Cité, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France
[80] INAF-IASF Bologna, Via Piero Gobetti 101, I-40129 Bologna, Italy
[81] INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste TS, Italy
[82] SISSA, International School for Advanced Studies, Via Bonomea 265, I-34136 Trieste TS, Italy
[83] Departamento de Astrofísica, Universidad de La Laguna, E-38206, La Laguna, Tenerife, Spain
[84] Instituto de Astrofísica de Canarias (IAC); Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38200, La Laguna, Tenerife, Spain
[85] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, F-31400 Toulouse, France
[86] Dipartimento di Fisica - Sezione di Astronomia, Universitá di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy
[87] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, I-40129 Bologna, Italy
[88] Dipartimento di Fisica e Scienze della Terra, Universitá degli Studi di Ferrara, Via Giuseppe Saragat 1, I-44122 Ferrara, Italy
[89] Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, D-52056 Aachen, Germany
[90] Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, E-38204, San Cristóbal de La Laguna, Tenerife, Spain
[91] Department of Physics & Astronomy, University of California Irvine, Irvine CA 92697, USA
[92] University of Lyon, UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, France
[93] INFN-Sezione di Genova, Via Dodecaneso 33, I-16146, Genova, Italy
[94] INAF-Istituto di Astrofisica e Planetologia Spaziali, via del Fosso del Cavaliere, 100, I-00100 Roma, Italy
[95] Instituto de Física Teórica UAM-CSIC, Campus de Cantoblanco, E-

28049 Madrid, Spain
[96] Department of Physics, P.O. Box 64, 00014 University of Helsinki, Finland
[97] Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK
[98] Observatoire de Paris, PSL Research University 61, avenue de l'Observatoire, F-75014 Paris, France
[99] Université de Paris, F-75013, Paris, France, LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, F-75014 Paris, France
[100] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[101] Code 665, NASA Goddard Space Flight Center, Greenbelt, MD 20771 and SSAI, Lanham, MD 20770, USA
[102] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[103] Centre de Calcul de l'IN2P3, 21 avenue Pierre de Coubertin F-69627 Villeurbanne Cedex, France
[104] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, I-00185 Roma, Italy
[105] Aix-Marseille Univ, CNRS, CNES, LAM, Marseille, France
[106] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
[107] Zentrum für Astronomie, Universität Heidelberg, Philosophenweg 12, D- 69120 Heidelberg, Germany
[108] Department of Mathematics and Physics E. De Giorgi, University of Salento, Via per Arnesano, CP-I93, I-73100, Lecce, Italy
[109] INFN, Sezione di Lecce, Via per Arnesano, CP-193, I-73100, Lecce, Italy
[110] Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[111] Junia, EPA department, F 59000 Lille, France
[112] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA
[113] Department of Physics, P.O.Box 35 (YFL), 40014 University of Jyväskylä, Finland
[114] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing (GCCL), 44780 Bochum, Germany

Appendix A: Impact of redundant features on feature importance

Machine learning methods that build models using decision-tree ensembles have the potential to provide insights into the structure of the training data via analyses of the feature-importances. When all features in the dataset are fully independent of each other, the feature-importance can provide a relatively straightforward indication of how useful each feature is for the model to predict the target labels. However, when there is significant colinearity between features, the importance may be shared between colinear features, resulting in potentially misleading feature-importance information. Indeed, it is likely that significant colinearity exists among the various broad-band magnitudes and colours used in this work. Therefore, here we examine how the feature importance calculations used by the `RandomForestClassifier`, `CatBoostClassifier`, and `XGBoostClassifier` tree-based algorithms are affected by the presence of multiple colinear features.

We first take the Int Wide catalogue, select features derived from u, $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$, or $H_{\rm E}$ photometry, and add five identical copies of the $u - I_{\rm E}$ feature to the dataset. We then train `RandomForestClassifier`, `CatBoostClassifier`, and `XGBoostClassifier` models to select quiescent galaxies in the $0 < z < 0.25$ range. We selected the $u - I_{\rm E}$ colour for duplication because `RandomForestClassifier`, `CatBoostClassifier`, and `XGBoostClassifier` all find this feature to be the most important for this particular classification problem. For these model runs, no information was provided regarding the redshifts of the galaxies. The results are shown in Figs. A.1– A.3.

The impact on the feature importances of duplicating the most important single feature ($I_{\rm E} - u$ in this example) is somewhat different for each of the learning algorithms, presumably due to: (i) differences in the way the algorithms select features for constructing individual decision trees; (ii) how they deal with colinearity among the input features; (iii) the different methods used to calculate feature importance values.

In the case of `RandomForestClassifier`, the impact of including duplicates of $I_{\rm E} - u$ is for this feature and its duplicates to be demoted to a significantly lower rank of importance (Fig. A.1), with all six features occupying a similar position within the feature importance ranking. Clearly, it is risky to rely on the feature importances produced by `RandomForestClassifier`, since they are a function of how informative a features is and the uniqueness of the information it provides. When using `CatBoostClassifier`, the inclusion of duplicates of $I_{\rm E} - u$ also results in the demotion of $I_{\rm E} - u$ and its duplicates to significantly lower ranks of importance (Fig. A.2), with several being placed at the very end of the importance ranking.

In contrast, the feature importance values provided by `XGBoostClassifier` are much more robust against the presence of colinearity among features. As illustrated in Fig. A.3, $I_{\rm E} - u$, and/or several copies thereof are consistently assigned the highest values of feature importance, or else are simply ignored (feature importance = 0). As such, the feature importances provided by `XGBoostClassifier` are likely to offer a relatively robust method to determine the most relevant observables for the selection of particular galaxy types, even when there is significant colinearity between the observables.

Appendix B: Further analysis and tests

In this appendix we provide full details of the analyses and tests that were summarized in Sect. 7.

B.1. Stacking vs. individual learners

We discuss the benefits of our implementation of the generalized stacking method, in which meta-learners are trained to fuse the output from several base-learners into a single classifier. We find that, with very few exceptions, our stacking method consistently outperforms each of the individual base-learners, in addition to outperforming the traditional ensembling methods of model averaging and hard-voting. This is illustrated in Fig. B.1 (left panel), where we show results from a single run of our pipeline, in this case applied to the selection of quiescent galaxies at $z = 2$–2.5 from the Int Wide catalogue using Euclid photometry, and without foreknowledge of galaxy redshifts. In this example, averaging the predictions across the base-learners results in an 'averaging-down', while the hard-vote ensemble method results in an F1-score that matches that of the best individual base-learner (in this case `LightGBMClassifier`). In contrast, the meta-learner yields an F1-score that is substantially higher ($\sim$ 46 per cent in this case) than any of the individual base-learners or other ensemble methods.

In Fig. B.1 (centre panel) we illustrate the robustness of the generalized stacking method against pollution by multiple low-quality classifier models. We ensembled a LightGBM model with hyperparameters that are well tuned for this problem (LightGBM 1), with four other LightGBM models that have purposefully poorly tuned hyperparameters (LightGBM 2–5). Whereas the average and hard-voting ensembles give poor results, the meta-learner is able to discard low-quality class predictions and also make a significant improvement over the single high-quality model (LightGBM 1).

In addition, we illustrate the usefulness of generalized stacking when applied to a single classifier model. In Fig. B.1 (right panel), we show the result of selecting quiescent galaxies in the $z = 2.5$–3 redshift range, using *ugriz*, *Euclid*, $W1$, and $W2$ photometry from the Int Wide catalogue, with foreknowledge of redshifts. The meta-learner is able to substantially improve on the F1-score of the `XGBoostClassifier`, increasing the score by 61 per cent, effectively turning a rather poor classifier into a potentially much more useful one.

B.2. The nature of the false positives

Although the P, R, and F1-score classification metrics are informative about whether galaxies are correctly or incorrectly classified, they do not provide information about the *incorrectness* of the incorrect classifications. For instance, the F1-score is insensitive to whether false positives are marginally non-quiescent (e.g. sSFR $\sim 10^{-10.4}\,{\rm yr}^{-1}$), or are in fact powerful starburst galaxies (e.g. $10^{-8}{\rm yr}^{-1}$); when selecting samples of quiescent galaxies, the former case is clearly less baneful than the latter. Therefore, we now examine the nature of the non-quiescent contaminants in samples selected as quiescent by our classification pipeline, utilizing our metrics of incorrectness $I_{\rm FP}$ and $\bar{I}_{\rm FP}$ (see Eqs. 7 and 8).

In Figs. B.2 and B.3 we show the distribution of incorrect classifications with respect to $I_{\rm E} - H_{\rm E}$, sSFR, and stellar mass, for the Int and SED catalogues. As is often the case when performing a binary classification on a continuously distributed sample, the incorrectly classified objects cluster around the class thresh-

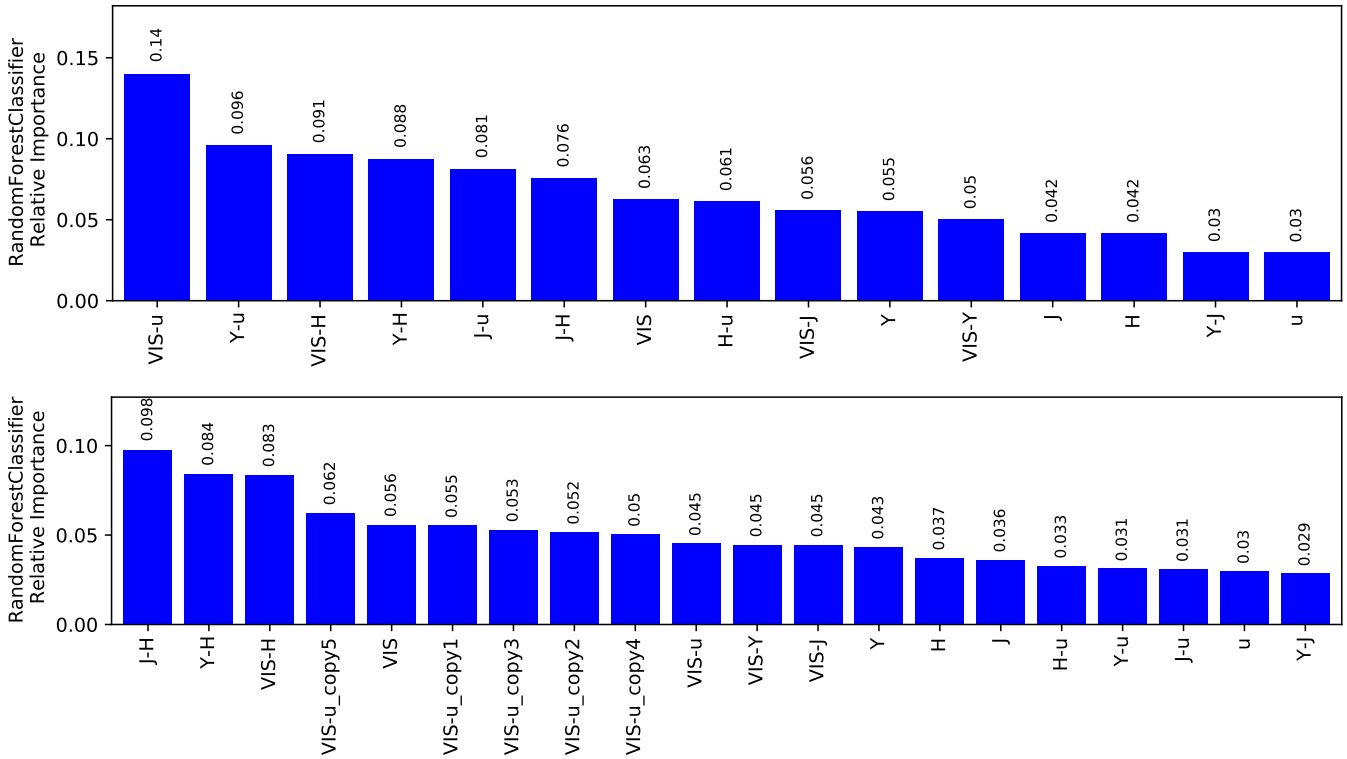

Fig. A.1. Feature importance values when using `RandomForestClassifier` to select quiescent galaxies in the redshift range $0 < z < 0.25$. The results in the upper panel are for the case where none of the features have been duplicated. The lower panel shows the feature importance values for the case where five copies of $I_E - u$ have been injected into the dataset prior to model training. The x-axis labels correspond to feature names used by the pipeline after the pre-processing steps outlined in Sect. 4.2 have been applied, and should be self-explanatory (see the caption of Fig. 7. For example, the feature named 'VIS-u_copy1' is a copy of the feature named 'VIS-u', etc.

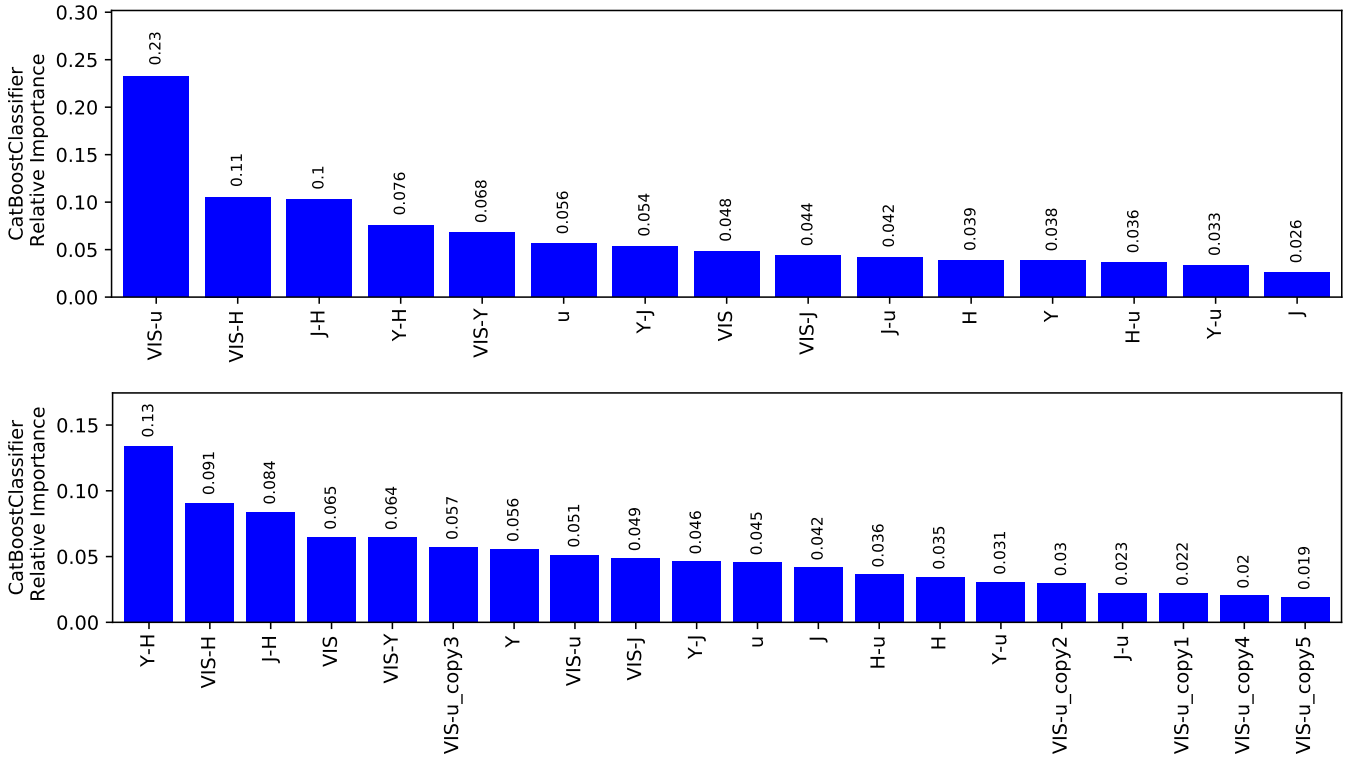

Fig. A.2. Similar to Fig. A.1, but for models trained using `CatBoostClassifier`.

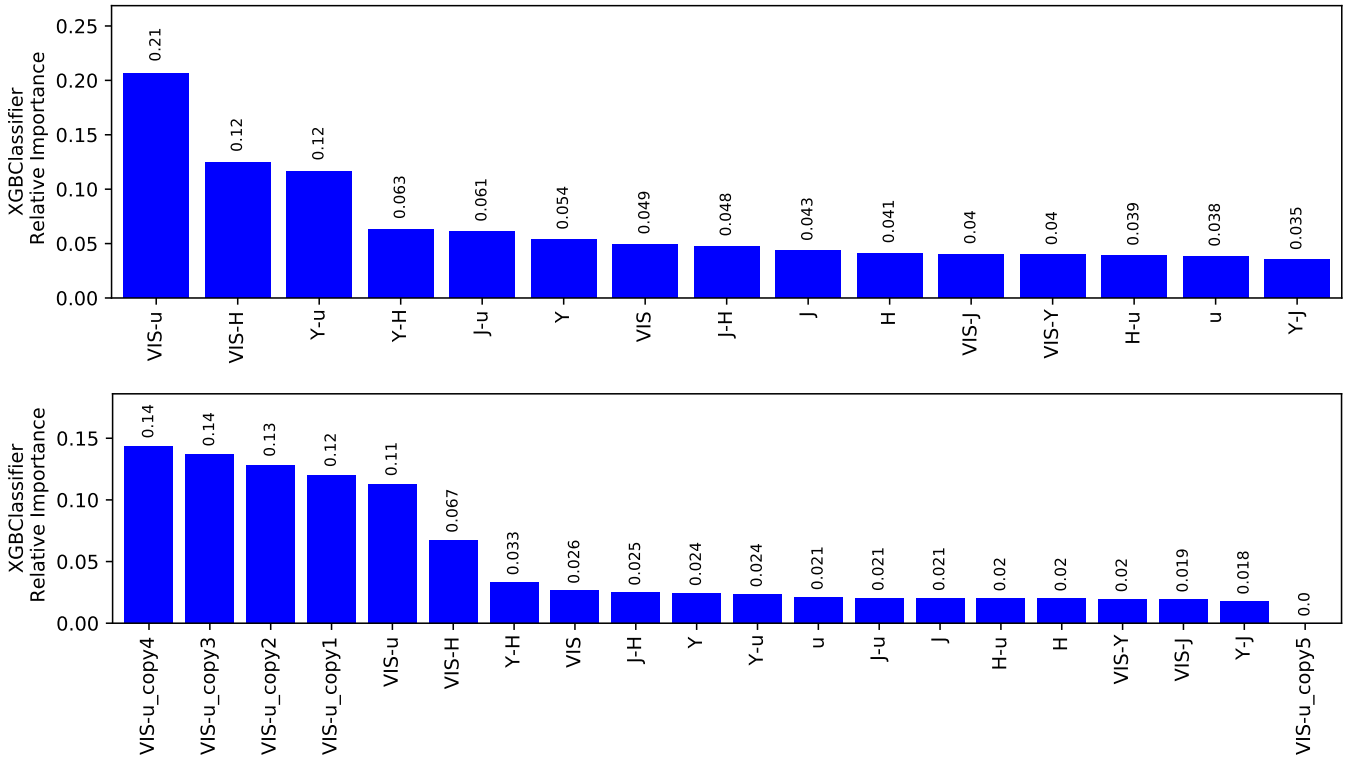

Fig. A.3. Similar to Fig. A.1, but for models trained using `XGBoostClassifier`.

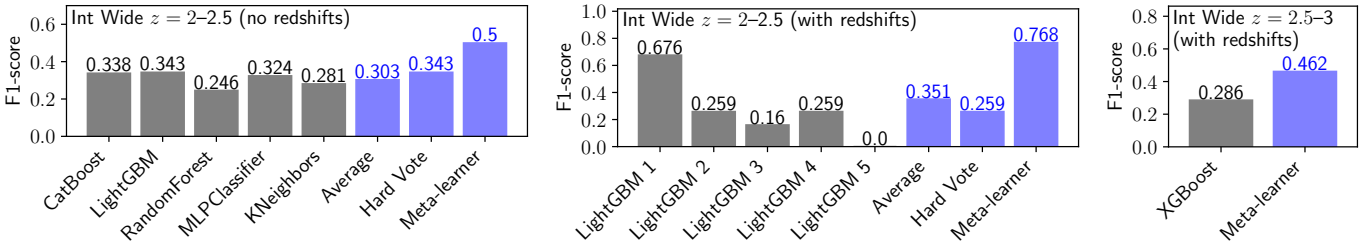

Fig. B.1. Examples of the F1-scores from individual base-learners and the model ensembling methods. **Left:** Selection of quiescent galaxies at $z = 2$–2.5 from the Int Wide catalogue using Euclid photometry, without foreknowledge of galaxy redshifts. As described in the text, the meta-learner performs a non-linear fusion of the individual classifiers, resulting in a significantly higher F1-score than obtained by any of the individual base learners or the two other ensemble methods (averaging and hard-voting). **Centre:** The impact of ensembling a `LightGBMClassifier` model, the hyperparameters of which are well-tuned for this problem (LightGBM 1), with four other `LightGBM` models that have poorly tuned hyperparameters (LightGBM 2,3,4,5). In this case, averaging the model predictions and hard-voting both produce poor results, but the meta-learner is able to identify and weight accordingly the low quality class predictions. **Right:** Application of a meta-learner to a classification model produced by a single base-learner, in this case `XGBoostClassifier`. In this circumstance, the meta-learner performs 'error-correction', resulting in a significant improvement in the quality of the classifier.

old value ($10^{-10.5}\mathrm{yr}^{-1}$), with a density that is highest in the bins immediately adjacent to the class boundary. However, the precise distribution of classification errors varies between the different catalogues (and subsets thereof).

When using the Wide survey mock catalogues, $\gtrsim 50$ per cent of false positives are what we consider to be marginal classification errors, i.e., their sSFR is within 0.5 dex of the class boundary. Furthermore, there are very few false positives with high values of sSFR: less than 25 per cent of the false positives are at sSFR $\geq 10^{-9.5}\mathrm{yr}^{-1}$, while false positives with sSFR $\geq 10^{-9}\mathrm{yr}^{-1}$ are negligible ($\lesssim 5$ per cent). On the other hand, when the Deep survey mock data are used, the fraction of false positives at relatively high values of sSFR ($\geq 10^{-9}\mathrm{yr}^{-1}$) is non-negligible ($\sim 25$ per cent). This is true regardless of whether the Int Deep or SED Deep catalogue is used. The distribution of the incorrect classifications with respect to the stellar mass shows a significant diversity and depends strongly on which mock catalogue is used. In general, the false positives are biased towards relatively high stellar mass (i.e., $\geq 10^9 M_\odot$).

For comparison, in Fig. B.4 we also show the distribution of errors with respect to sSFR for the $I_\mathrm{E} - Y_\mathrm{E}$, $J_\mathrm{E} - H_\mathrm{E}$ (left column) and $u - I_\mathrm{E}$, $I_\mathrm{E} - J_\mathrm{E}$ (middle column) colour-colour methods developed by B20. To allow a relevant comparison between the $u - I_\mathrm{E}$, $I_\mathrm{E} - J_\mathrm{E}$ colour-colour method and our machine learning pipeline, Fig. B.4 (right column) also shows results from applying our pipeline with conditions equivalent to those used for the $u - I_\mathrm{E}$, $I_\mathrm{E} - J_\mathrm{E}$ method: (i) Galaxy redshifts are included as a feature to be trained on; (ii) only galaxies detected in u, I_E, and J_E are used; (iii) only galaxies in the redshift ranges $0 < z < 1$

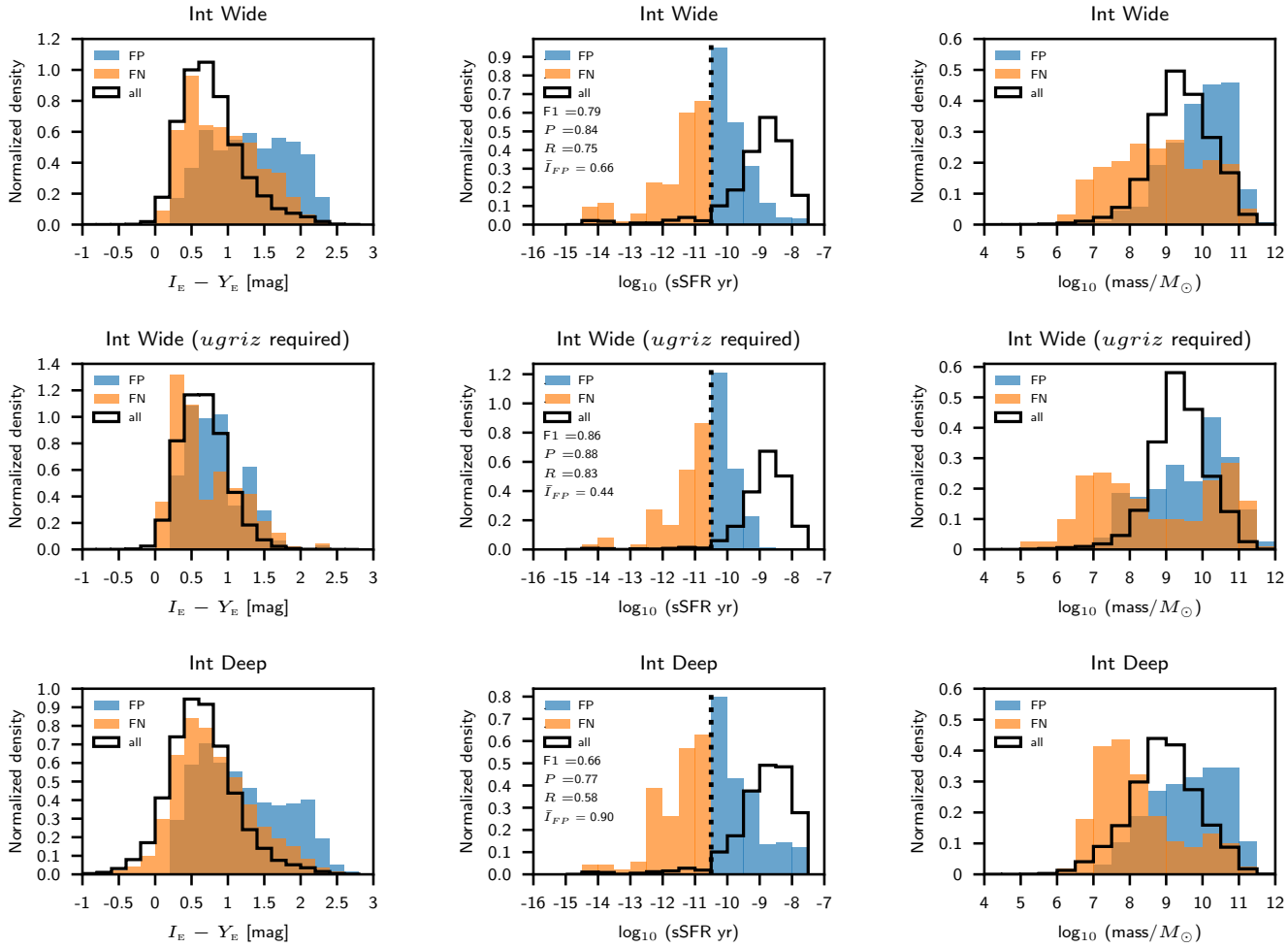

Fig. B.2. The distribution of incorrect classifications (FP and FN) with respect to $I_E - Y_E$, sSFR, and stellar mass, when selecting quiescent galaxies from the Int catalogues. Also shown is the overall distribution of galaxies in the catalogue (or catalogue subset) on which the selection was performed. **Top row:** Using the *ugriz*, Euclid, *W*1, *W*2, and 20 cm photometry from the Int Wide catalogue, excluding only those galaxies with a non-detection in any Euclid band. **Centre row:** As above, but excluding galaxies without a non-detection in any of the *ugriz* or Euclid bands. **Bottom row:** Using the *ugriz*, Euclid, *W*1, *W*2, and 20 cm photometry from the Int Deep catalogue, excluding only those galaxies with a non-detection in any Euclid band. In each case, we include the values of the F1-score, precision, recall, and $\bar{I}_{FP}$ metrics

and $0 < z < 1.5$ are used for the Wide and Deep catalogues, respectively.

Our classification pipeline results in significantly lower $\bar{I}_{FP}$ than the colour-colour methods. This is due to a reduction in the fraction of false positives that are located at high sSFR (e.g. $\gtrsim 10^{-9}$yr^{-1}). In other words, our pipeline not only offers a significant improvement over colour-colour methods, in terms of *P*, *R*, and the F1-score, but also significantly reduces the degeneracy between quiescent galaxies and dusty, star-forming galaxies.

There is another potential source of classification errors that should be mentioned. Throughout this work we have tacitly assumed that the target variable accurately represents the ground truth. This assumption is clearly valid for the SED catalogues, since they are derived from templates corresponding to known physical properties. However, for the Int catalogues there is the possibility that some instances flagged as classification errors are, in actual fact, instances where our pipeline provides the correct classification and the target variable is incorrect.

B.3. Reconciling the Int and SED results

As discussed above, our machine learning models were trained and evaluated using one of the four mock catalogues, but there are often significant differences between the results obtained using the Int or SED catalogues for a given *Euclid* survey (e.g. Fig.. 8). In particular, the precision, recall, and F1-scores for quiescent galaxy selection tend to be higher when using an SED mock catalogue, compared to its corresponding Int catalogue (i.e., SED Wide vs. Int Wide; SED Deep vs. Int Deep).

This is due to the different methods used in the construction of the catalogues (see Sect. 2 and B20). The Int catalogues have an observationally more realistic starting point since they are constructed with real photometry, albeit twice convolved with a filter, but the signal-to-noise ratio of the data is significantly lower than will be the case for the actual *Euclid* photometry, likely increasing the difficulty of the classification problem compared to when the real *Euclid* (and ancillary) survey data are used. Conversely, the mock photometry in the SED catalogues has noise properties that match those expected for *Euclid* observations, but the SEDs themselves are forced to conform to

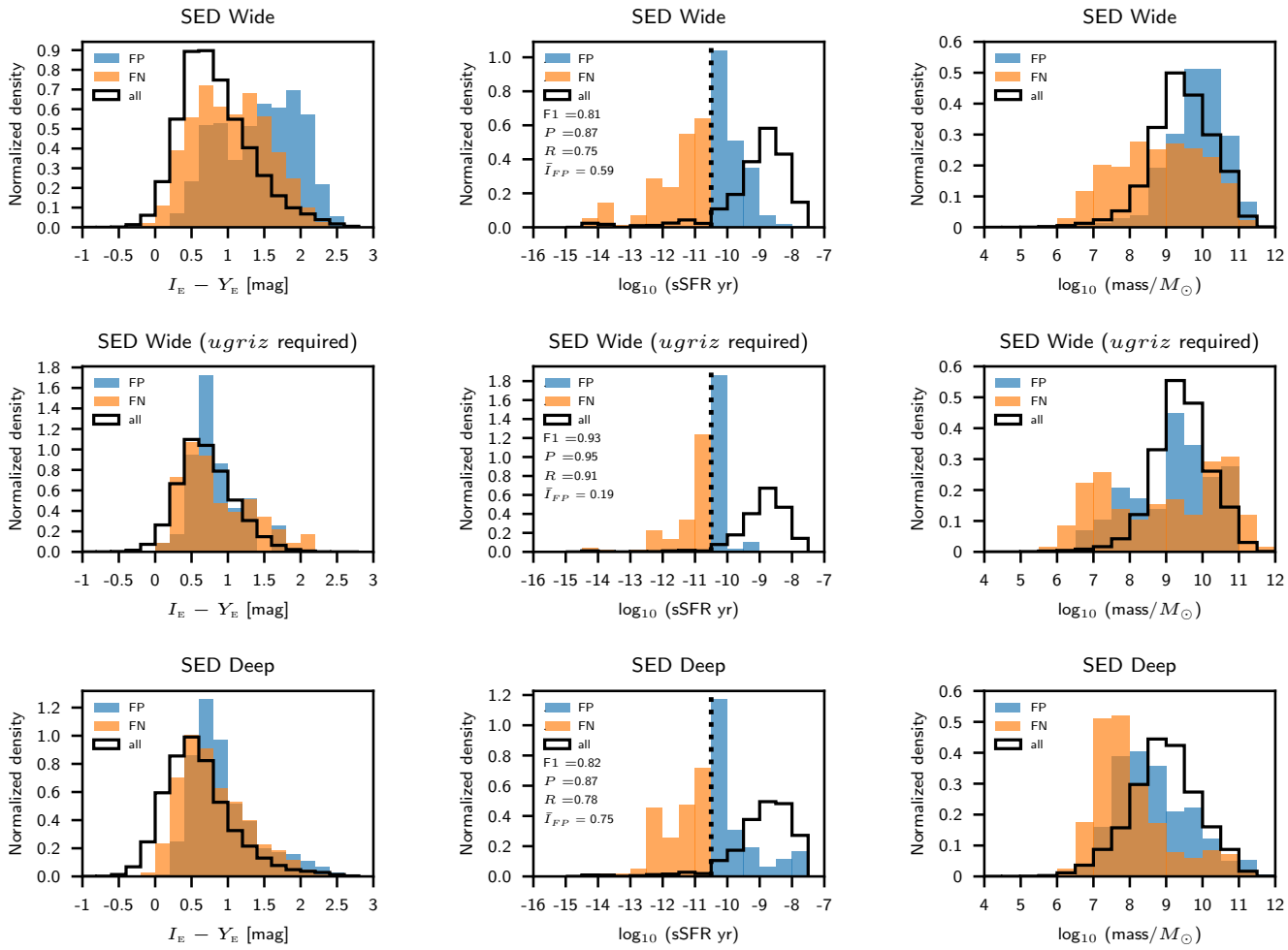

Fig. B.3. Similar to Fig. B.2, but instead using the SED catalogues.

a restricted set of simplified templates, which could potentially simplify the classification problem.

Therefore, results obtained using the Int catalogues are likely to be pessimistic, and results obtained with the SED catalogues are likely to be optimistic with regard to the performance of our pipeline when applied to real *Euclid* (and ancillary) survey data. Thus, to estimate the performance of our pipeline when selecting quiescent galaxies from *Euclid* (and ancillary) survey data, we use performance metrics averaged over the Int catalogue and the corresponding SED catalogue (i.e., Int Wide and SED Wide; Int Deep and SED Deep). Therefore, Tables 3 and 4 include averaged metrics, where appropriate. We consider results obtained with the Int and SED catalogues to bracket the likely range of performance of our pipeline when it is applied to real *Euclid* (plus LSST, etc.) photometry.

B.4. Tuning the probability threshold

In some circumstances, it is desirable to maximise either the precision (purity) or the recall (completeness) of the selected quiescent galaxy samples, according to the nature of one's scientific objectives. Thus, we investigate and illustrate the impact on the precision and recall of tuning the value of the class probability threshold, instead of adopting the default threshold value of 0.5.

In Fig. B.5 we show how precision (P), recall (R), and the $F1$-score vary with the class probability threshold when selecting quiescent galaxies from the Int Wide mock catalogue (top) or the SED Wide catalogue (middle). Also shown are the scores when averaged between the two catalogues (bottom). The redshifts are included as a feature in the input data. There exists a trade-off between precision and recall such that one may be increased, but at the cost of reducing the other. For example, from the averaged scores (bottom panel), we find that adopting a probability threshold of 0.85 yields a sample of quiescent galaxies that is very pure (P = 0.98) but somewhat incomplete (R = 0.56). Conversely, using a probability threshold of 0.05 results in a sample with moderate purity (P = 0.61) but high completeness (R = 0.97).

Tuning the probability threshold allows a balance to be struck between P and R that is suitable for different scientific needs. For instance, using a probability threshold of 0.3 gives a sample of quiescent galaxies that is both reasonably pure (P = 0.8) and reasonably complete (R = 0.9). While the examples given here pertain to selection in the redshift range $0 \leq z \leq 3$, this exercise can, of course, also be performed for the selection of quiescent galaxies in narrower redshift bands. As an example of this, when selecting quiescent galaxies in the redshift range $1 \leq z \leq 2$ using a probability threshold of 0.9, we obtain P = 0.98 and R = 0.51, while using a threshold of 0.1 yields P = 0.70 and R = 0.95.

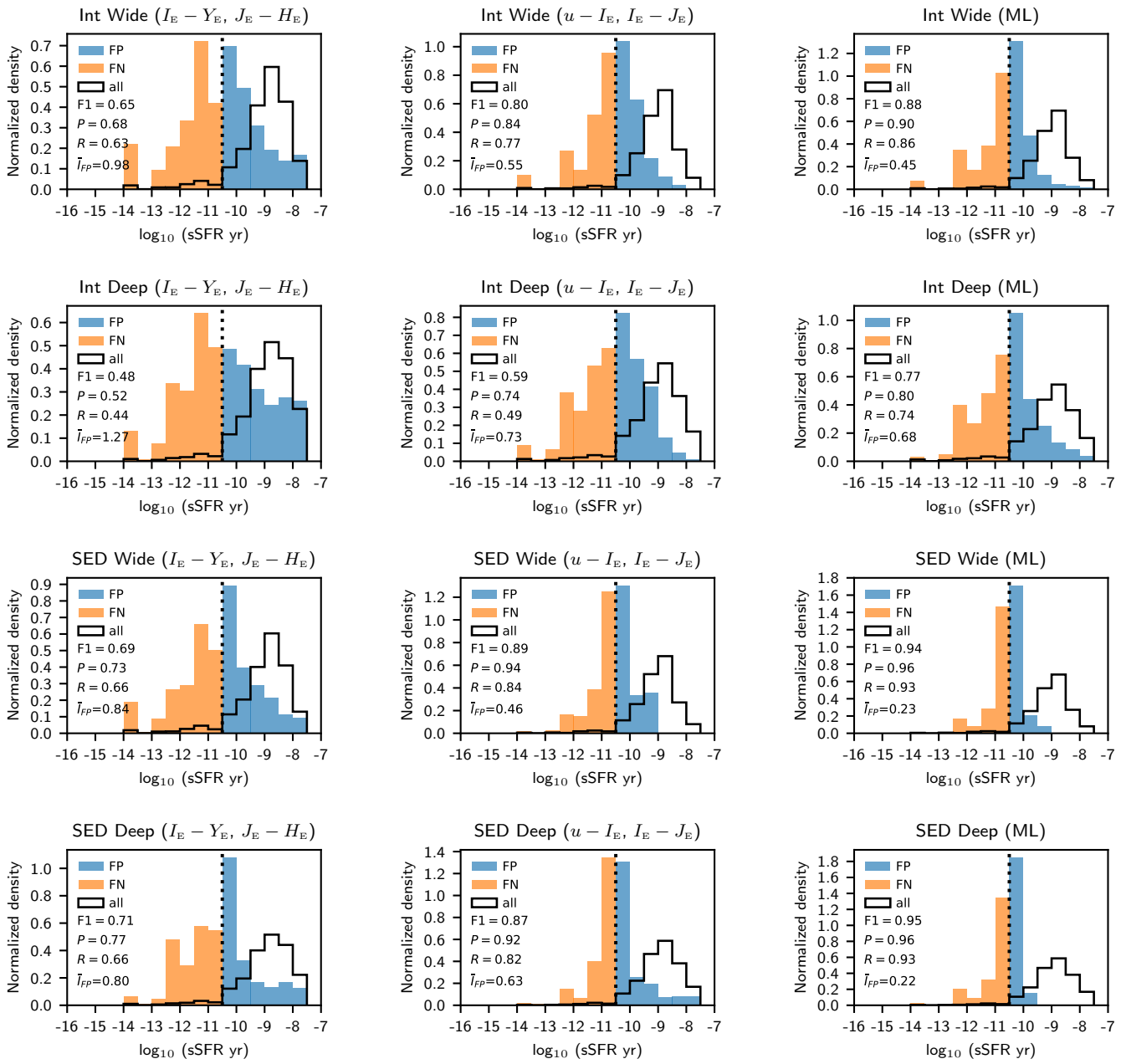

Fig. B.4. The distribution of incorrect classifications for the $I_E - Y_E$, $J_E - H_E$ method of B20 ((left column) and the $u - I_E$, $I_E - J_E$ method of B20 (centre column) To allow a direct comparison between the $u - I_E$, $I_E - J_E$ and our machine learning selection method, we show results from our pipeline under conditions equivalent to those used for the B20 $u - I_E$, $I_E - J_E$ method: Galaxy photometric redshifts are included as a feature to be trained on; only galaxies detected in u, I_E, and J_E are used; only galaxies in the redshift ranges $0 < z < 1$ and $0 < z < 1.5$ are used for the Wide and Deep catalogues, respectively. In each panel, we include the values of the F1-score, precision, recall, and $\bar{I}_{FP}$ metrics.

B.5. Impact of including redshift as a feature

Here we investigate the impact of several different methods for the treatment of redshift information in our pipeline. Thus far, we have included redshift information by pre-binning the mock catalogues using the Laigle et al. (2016) COSMOS2015 photometric redshifts (Sect. 5.1). Alternatively, we have ignored redshift information and have instead performed a global selection of quiescent galaxies, deriving photometric redshifts subsequently (Sect. 5.2). A further possibility that we now also examine is the inclusion of redshifts as a feature in the input data for model training (e.g. Simet et al. 2021).

In Fig. B.6 we show F1-score vs. redshift when selecting quiescent galaxies from the Int Wide catalogue, considering several different configurations for the included redshift information. As before, we exclude galaxies which have a non-detection in one or more *Euclid* band. The left panel of Fig. B.6 shows results from our pipeline when the mock catalogue data are pre-binned by redshift as described in Sect. 5.1, and all available photometry is used (i.e., *ugriz*, *Euclid*, *W*1, *W*2, 20 cm). We find that including the Laigle et al. (2016) photometric redshifts in the input data significantly increases the F1-scores in bins at $z \leq 0.5$ or $z \geq 2.5$, by reducing the degeneracy between redshift and sSFR. However, the F1-score is not significantly changed for bins in the range $0.5 < z < 2.5$.

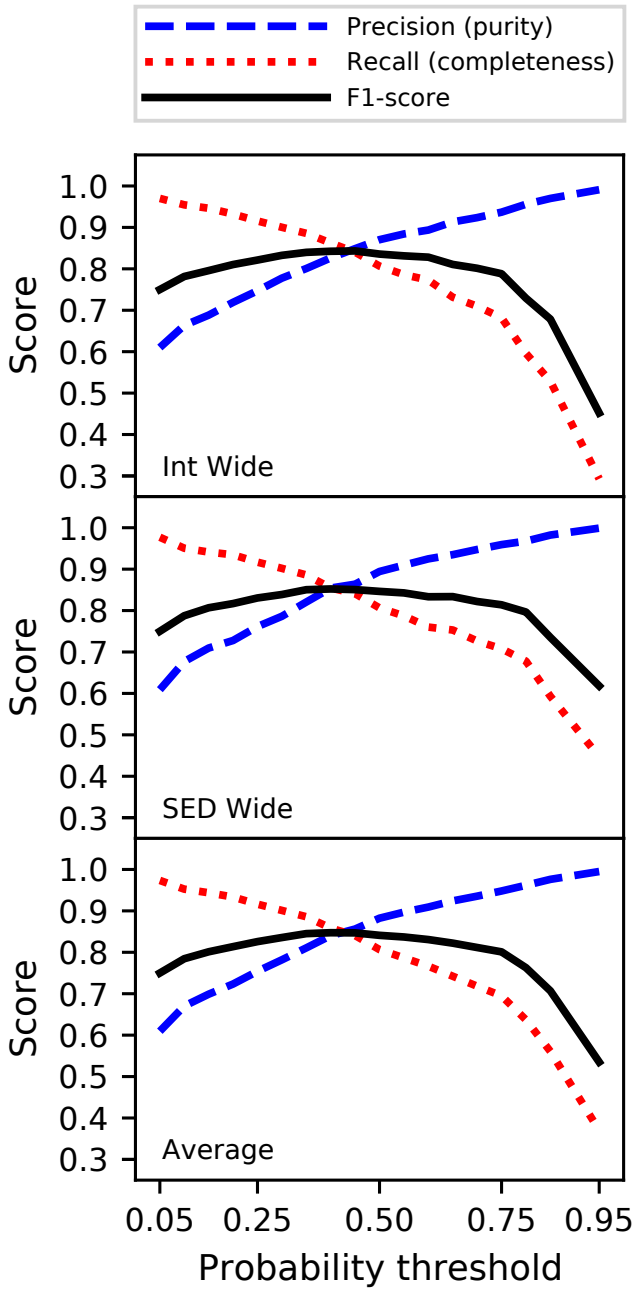

Fig. B.5. Precision (purity), recall (completeness), and the F1-score as a function of the probability threshold used to separate quiescent and star-forming galaxies, for Int Wide (top), SED Wide (centre), and the average result (bottom).

We have also experimented with the inclusion of redshifts for a fraction of galaxies only, by randomly replacing redshift values with −99.9. The left panel of Fig. B.6 also shows the result of including only 50 per cent of the redshifts. At the low end of the redshift range, we find that the inclusion of 50 per cent of the redshifts provides a small but significant improvement in the F1-score. However, at the high endpoint of the redshift range, there is no noticeable improvement compared to the F1-scores obtained when no redshifts are included.

We repeat the same experiment, but without pre-binning the galaxies by redshift (see Sect. 5.1.3). As before, we exclude objects with a non-detection in any of the *Euclid* bands, and use all available photometry from the Int Wide catalogue. The results are shown in the right panel of Fig. B.6. Within the $0 < z < 2$ range, there is essentially no penalty in terms of F1-score associated with not pre-binning by redshift, provided all redshifts are included as a feature in the input data (nevertheless, pre-binning does allow considerably faster training of models, since the training set is now much smaller). However, at $z = 2$ and above, significantly lower F1-scores are obtained compared to the case where the data are pre-binned by redshift. When only 50 per cent of redshift values are included, we find a significant decrease in F1-scores compared to the case where 100 per cent of redshift values are included, but the F1-score are still usually significantly above those obtained when none of the redshift values are included. Classification metrics for the global selection of quiescent galaxies, including 100 per cent, 50 per cent, or none of the galaxy redshifts are included in Table 4.

B.6. The impact of noise

In this subsection, we explore the impact of different types of noise that are expected to be present within the data. The experiments presented here are intended to be informative and illustrative, but not necessarily exhaustive.

B.6.1. Adding noise to the photometry

In Fig. B.7 we show results from modelling the impact of the addition or the reduction of noise in the data. For this experiment, we select quiescent galaxies at $0 \leq z \leq 3$ from the Int Wide mock catalogue, using our pipeline in *fast mode*. The upper panel of Fig. B.7 shows how the F1-score decreases when each magnitude measurement has a random offset, drawn from a Gaussian distribution of σ, applied. Interestingly, even when the data are extremely noisy, the pipeline remains nominally functional. For instance, even when the photometry has been degraded to a signal-to-noise ratio of ~ 3, it is nonetheless still able to perform a global selection of quiescent galaxies, albeit with somewhat reduced precision, recall, and F1-scores of ~ 0.67.

To simulate a reduction in noise, we perform a cut to remove galaxies fainter than an arbitrary $I_{\rm E}$ threshold, where lower values of this threshold result in a higher average signal-to-noise ratio for the mock catalogue (Fig. B.7, lower panel). Despite the crudeness of these tests, it is clear that reducing (increasing) the signal-to-noise ratio of the data results in lower (higher) quality classification models, as evaluated by the F1-scores. While the different noise characteristics probably play a significant rôle in the differences in pipeline performance between the Int and SED catalogues, additional effects may also be important.

B.6.2. Label noise

Heretofore, we have tacitly assumed that the 'quiescent' and 'star-forming' labels used in the training and evaluation of our classification models give an accurate representation of the ground truth. However, there is the possibility that some labels are incorrect, i.e., quiescent galaxies labelled as star-forming, or vice versa. We examine the potential impact of incorrect labels using two slightly different approaches.

In the first approach, we assess the impact on model quality from introducing incorrect labels at random. To do this, we select a subset of galaxies in the mock photometry catalogue, and for these galaxies we replace all occurrences of the value 0 with the value 1, and all occurrences of the value 1 with 0. We then

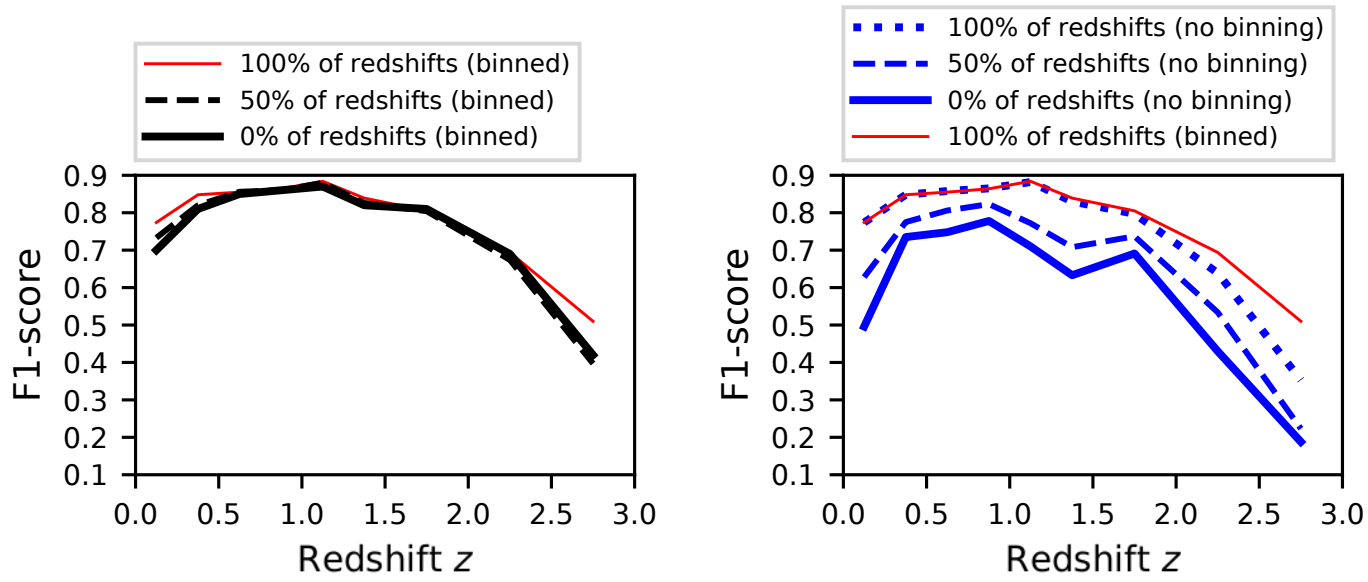

Fig. B.6. The impact of including source redshifts as an additional feature in the input data. **Left:** The case where the data set is binned by photometric redshift prior to model training, with 100 per cent, 50 percent, or none of the redshifts included as a feature in the data. **Right:** Illustrating the case where no redshift binning is performed, with classifiers being trained to identify quiescent galaxies at specific redshift intervals. For this test, the Int Wide mock catalogue was used.

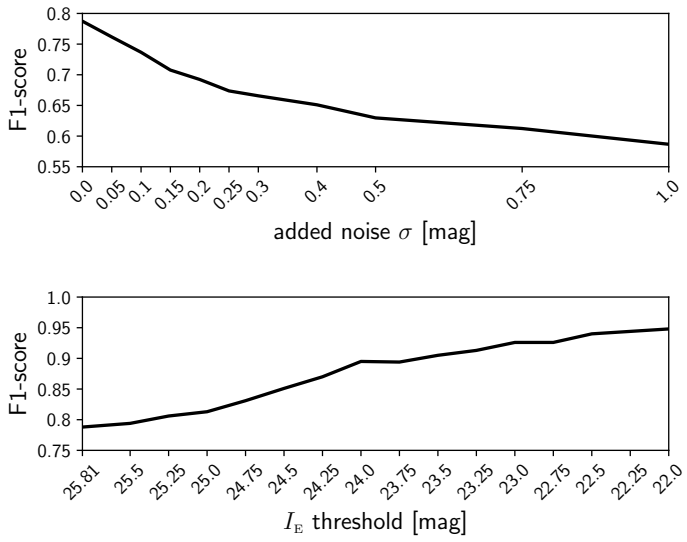

Fig. B.7. Testing the impact of photometry measurement uncertainties on the results from our classification pipeline. **Top:** An example of how the F1-score is reduced when Gaussian noise is added to the optical and near-IR photometry in the Int Wide catalogue. **Bottom:** An example of how the F1-score is increased when galaxies fainter than an arbitrary I_E threshold are excluded from the Int Wide catalogue.

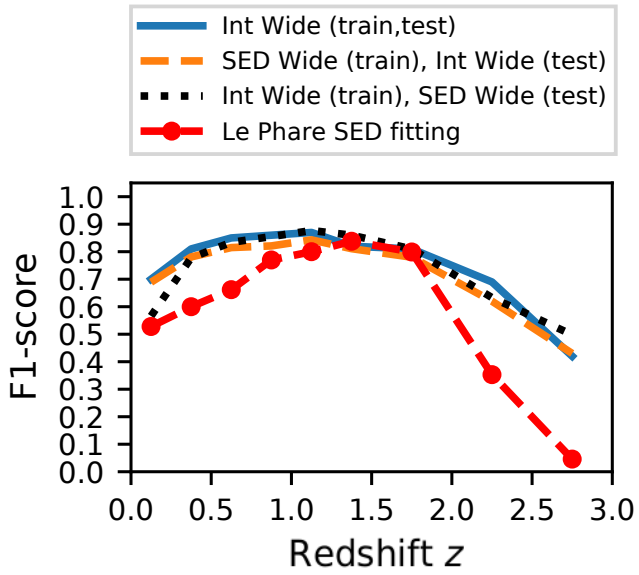

Fig. B.8. Results from our transfer learning method. When we train models on SED Wide data, and use them to select quiescent galaxies from Int Wide data (dashed orange curve), the F1-scores are only slightly lower than those we obtain from models trained on Int Wide data (solid blue curve). A generally similar result is obtained when we train models on Int Wide data and use them to select quiescent galaxies from SED Wide data (dotted black curve). For comparison we also show the results from the `LePhare` template fitting method applied to the SED Wide mock catalogue (see Sect. 6.4), using *Euclid*, *ugriz*, *W*1, *W*2, and 20 cm photometry, and with redshifts fixed at the Laigle et al. (2016) values.

perform our standard preprocessing and model training steps as described in Sect. 4. For these tests, the `ARIADNE` pipeline is used in *fast mode*. The classification metrics are evaluated for two different cases, where (i) the test set class predictions are compared with the original unaltered test set labels, or (ii) the test set class predictions are compared with the altered (noisy) test set labels.

In the context of this experiment we point out that the 'random errors' discussed here can either be truly random errors arising from the methodology used to generate the labels (e.g. `LePhare` template fitting), or else can be systematic errors that the machine learning algorithms are unable to model and reproduce. An example of the latter type of error might be a systematic error driven by photometric bands that are present in the multi-

wavelength dataset used to generate the labels, but which do not appear in the data that are seen by our pipeline.

Under these conditions, the presence of noisy labels varies depending on the particular classification problem that is being addressed. For instance, when selecting quiescent galaxies at $0 \leq z \leq 3$ from the Int Wide catalogue, without foreknowledge of redshifts, and with 33 per cent of the labels having been flipped, the metrics obtained when the test set labels contain errors are relatively poor, at $P = 0.62$, $R = 0.12$, and F1-score = 0.21. This is to be expected, because the classification results are being evaluated against a 'ground truth' that contains many incorrect labels, leading to artificially poor metrics.

Conversely, the metrics we obtain are substantially better if we instead evaluate the same classification results against the original 'ground truth', yielding $P = 0.86$, $R = 0.68$, and F1-score = 0.76; these values are only slightly different compared to the case where label errors are not introduced at all ($P = 0.84$, $R = 74$, F1-score = 0.79). In other words, for this case the presence of randomly incorrect labels in the training data did not have a substantial impact on the classification of sources in the test set.

In another example, selecting quiescent galaxies at $1 \leq z \leq 1.25$ with no foreknowledge of redshifts, and with 10 per cent of labels in error, results in an F1-score of 0.13 when evaluated using the test set 'ground truth' with label errors injected. Remarkably, when evaluating the same classification results using the test set 'ground truth' without label errors, the F1-score is 0.69, effectively unchanged from the case where no label errors are injected at all (F1-score = 0.69). In this case, the presence of random errors in the training set labels had no detrimental effect on the final classifications of sources in the test set.

One of the interesting implications of these results is that our pipeline is generally able to ignore random label errors. Furthermore, when random label errors are present, our pipeline may even produce predictions that are of higher quality than the labels in the input data, in terms of how close the predictions are to the real ground truth. Further research (beyond the scope of this paper) is needed to test whether machine learning methods such as those discussed herein may be able to improve on traditional SED fitting methods, rather than simply emulating them.

The second approach is similar to the first, but aims to simulate systematic errors in the class labels. For this, we use the `KMeans` clustering algorithm from `Scikit-Learn` to separate galaxies from the Int Wide catalogue into an arbitrary number of clusters, with an arbitrary number of the clusters being selected to have an arbitrary fraction of their class labels inverted. For illustrative purposes, we have separated the data into 100 clusters, and inverted 99 per cent of the labels in 3 randomly selected clusters; the inverted labels represent 11 per cent of all labels. We then selected quiescent galaxies at $0 \leq z \leq 3$, again using features derived from the *ugriz*, *Euclid*, *W*1, *W*2, and 20 cm bands, and without foreknowledge of galaxy redshifts. Evaluating the metrics when using the modified labels for the test set, we obtain $P = 0.89$, $R = 85$, and an F1-score of 0.87. In contrast, evaluating the metrics using the original test set labels, we obtain the substantially lower values $P = 0.36$, $R = 0.69$, and an F1-score of 0.47. In this case, our pipeline is able to emulate the systematic label errors present in the training set, resulting in test set class predictions that contain similar systematic errors.

B.6.3. Redshift noise

As we have demonstrated heretofore, the inclusion of redshift information in the training data often results in classification models that are better able to correctly identify quiescent galaxies (e.g. Appendix B.5). A key point to be considered is whether our results are significantly affected by the accuracy of the photometric redshifts used, especially since the 30-band COSMOS2015 redshifts (Laigle et al. 2016) we have used could potentially be more accurate than the redshifts that will be estimated from *Euclid* photometry and the anticipated ancillary data (see e.g. Euclid Collaboration: Desprez et al. 2020). Therefore, we explore the impact of adding Gaussian noise to the COSMOS2015 redshifts prior to model training. While a full treatment of this issue is beyond the scope of the present work, we consider several different cases in order to obtain indicative results.

Random samples were drawn from a Gaussian distribution with standard deviation σ_z; these values were multiplied by $1 + z$ and then added to the photometric redshift values after the Target variable was set, simulating the addition of Gaussian noise. The global selection of quiescent galaxies in the range $0 < z < 3$ was then repeated with the (now noisier) redshifts included as a feature, along with the *ugriz*, *Euclid*, and Wise photometry, and colours derived therefrom. This test was performed for the Int Wide and SED Wide catalogues, for the values $\sigma_z = 0.025, 0.05, 0.075$.

The results from this test, averaged over an equal number of pipeline runs on the Int Wide and SED Wide catalogues, are shown in Table 4. While P is essentially unchanged, R, and the F1-score are slightly reduced by ~ 0.01–0.02.

Next, we consider the impact of redshift noise on the selection of quiescent galaxies in the narrower redshift bins used in Sect. 5.1. As one might expect, it is more difficult to select quiescent galaxies inside these narrower redshift bins when the photometric redshifts are noisier. This is because the models, in addition to separating quiescent and star-forming galaxies, must now also separate quiescent galaxies by their redshift. For example, in the case where $\sigma_z = 0.05$, the F1-score is typically reduced by ~ 0.07 compared to the case where no noise is added. Nevertheless, the F1-scores are still significantly higher than when the redshifts are not included at all, with the improvement ranging from ~ 0.02 at $1 < z < 1.5$, to ~ 0.2 at $z < 0.5$ and $z > 2$. Interestingly, even in the case where $\sigma_z = 0.1$, representing rather noisy redshifts (NMAD ~ 0.1), the F1-scores at $z < 0.5$ and $z > 2$ are still ~ 0.1 higher than when redshifts are not included. In other words, even when photometric redshifts are somewhat noisy, their inclusion in the data can nevertheless result in significantly stronger classification models, compared to when photometric redshifts are not used.

B.7. Transfer learning experiments

B.7.1. Training on templates and predicting on real SEDs

We have also experimented with the possibility of using classification models trained on spectral templates to select quiescent galaxies from catalogues of observed photometry. To explore this, our pipeline trained its classification models on the SED Wide mock catalogue, and selected quiescent galaxies from the Int Wide catalogue using the resulting classifier. Essentially, we gave our pipeline the task of identifying and weighting the defining characteristics of quiescent and star-forming galaxies from a set of simplifying abstractions (galaxy SED templates), rather than from observed SEDs. The train-test split was performed as described in Sect. 4.1, ensuring that each galaxy is present in either the training set or the test set, but not both. For this experiment, the data are pre-binned by redshift as described in

Sect. 5.1, but we did not include the redshift values as a feature in the input data for the classification pipeline.

The results are shown in Fig. B.8, where it can be seen that classifiers trained on the SED Wide catalogue are indeed able to select quiescent galaxies from the Int Wide catalogue, albeit with marginally lower F1-scores compared to models trained on the Int Wide catalogue itself. This opens up the interesting possibility of using machine learning models trained on synthetic galaxy SEDs as a potential alternative to traditional colour-colour or template fitting methods (e.g. Girelli, Bolzonella, & Cimatti 2019; Cecchi et al. 2019) of selecting quiescent galaxies that have redshifts (or other properties) for which there are no (or few) known examples.

Fig. B.8 also reveals that the transfer learning method can also function in reverse. When the training and test data sets are swapped with each other, such that classification models are trained on the Int Wide catalogue and are then used to select quiescent galaxies from the SED Wide catalogue, very similar results are obtained. The exception is in the lowest redshift bin ($0 < z < 0.25$), where the F1-score is now reduced by ~ 0.12 compared to the previous case. Classification metrics for the global selection of quiescent galaxies using transfer learning are also included in Table 4.

We also show in Fig. B.8 (red dashed line) the F1-scores obtained when using `LePhare` to fit templates to SED Wide mock data using the same photometry, and with redshifts fixed to the values from Laigle et al. (2016, see also Sect. 6.4). This template fitting method clearly provides results of similar quality to our pipeline in the $1.0 \lesssim z \lesssim 2.0$ redshift range, but at $z \lesssim 1.0$ or at $z \gtrsim 2$ the method significantly under-performs our transfer learning method. The under-performance of our `LePhare` template fitting method is likely caused, at least partly, by the absence of priors concerning (i) the known (or suspected) distribution of the galaxy classes within the redshift and colour-spaces, and (ii) the relative importance (or weighting) that should be given to each data point in the broad-band spectral energy distribution, aside from their signal-to-noise ratio.

B.7.2. Training on Deep and predicting on Wide survey data

The Euclid Deep Survey will provide photometry (and spectra) in several fields for which there are pre-existing multiwavelength observations, allowing the construction of source catalogues with high-quality labels. In turn, this is expected to facilitate the training of classifiers which can then be used to predict labels for the enormous number of sources that will be detected in the Euclid Wide Survey. An important question, however, is whether models trained using the deep field photometry are suitable for use in selecting quiescent galaxies in the wide field survey. While the Deep and Wide Surveys are expected to be somewhat similar (but clearly not identical) within the magnitude-space covered by the Wide Survey, at least 60 per cent of the galaxies in the Deep Survey are below (or close to) the I_E 3σ detection threshold of the Wide Survey. It is not clear *a priori* how the presence of these faint galaxies will affect the quality of the classification models.

To test this, we train our pipeline using the Int Deep catalogue, and use the resulting classification model to predict classes for the galaxies in the Int Wide catalogue. No galaxy was permitted to be present in both the training and the test set. In addition, no foreknowledge of redshifts was assumed, all available photometry bands were used (*ugriz*, *Euclid*, $W1$, $W2$, and 20 cm), and only galaxies detected in all four *Euclid* bands were included. Under these conditions, the resulting F1-score for selection of quiescent galaxies from the Int Wide catalogue is 0.74, moderately lower than the F1-score of 0.80 obtained using models trained using Int Wide (see Table 4). Conducting this test instead using the SED catalogues resulted in a similarly reduced F1-score of 0.76, compared to 0.82 when using only SED Wide catalogue.

Thus, although it is possible to separate quiescent and star-forming galaxies in the Wide Survey mocks, the F1-score suffers a significant penalty and is reduced by at least ~ 0.06. Clearly, the presence of a large number of additional, faint galaxies in the training set exacerbates at least some of the degeneracies described in Sect. 1, and induces the learning algorithms to place undue weight on galaxies near the faint end of the magnitude distribution.

In this study we are, of course, only able to make use of mock *Euclid* photometry catalogues, but we must also consider how the training set for this task will be constructed from real *Euclid* data. Based on the analyses presented heretofore, we propose constructing the training set(s) from Deep Survey photometry from fields which have high-quality multi-wavelength observations (and spectroscopy), using the deepest available data in order to generate high-quality 'ground truth' labels, but then training classification models on Wide Survey-depth observations.

B.8. Alternative targets

In this work, we have used target labels that were generated using an sSFR threshold to differentiate between quiescent and star forming galaxies. However, it is also interesting to consider other methods for generating the labels. Thus, here we explore the impact of using labels derived from the NUV $- r$ vs. $r - J_E$ (Ilbert et al. 2010, 2013) or NUV $- r$ vs. $r - K$ (Arnouts et al. 2013) colour-colour methods, instead of the sSFR-based labels used throughout this work. The mock catalogue used for these tests is Int Wide, and features derived from the *ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$ and 20 cm bands were used, with detections required in all of the *Euclid* bands. Our pipeline was used in *fast mode*. The results are summarized in Table B.1.

In the global selection of quiescent galaxies ($0 \leq z \leq 3$), the impact of using labels the alternative labels is limited. In the case of the labels derived from the NUV$-r$ vs. $r-J_E$ method, the F1-score is improved by ~0.03 compared to the original (sSFR) labels. On the other hand, the F1-score when using the labels derived from the NUV$-r$ vs. $r-K$ method is ~0.02 lower compared to when the original labels are used.

When selected quiescent galaxies in narrower redshift ranges, having first removed sources whose photometric redshifts place them outside the range of interest, we usually obtain slightly higher F1-scores when using labels from the NUV $- r$ vs. $r - J_E$ method, compared to when the original labels are used. in the $z = 2.5 - 3$ bin, there is a particularly large improvement in trhe F1-score (from ~0.4 to ~0.6). Conversely, in the $z = 0 - 0.25$ bin we obtain a significantly lower F1-score when using when using labels from the NUV$-r$ vs. $r-J_E$ method (0.53 compared to 0.69). When labels derived from the NUV $- r$ vs. $r - K$ method are used, the resulting F1-scores are consistently lower compared to the case where the original labels (sSFR) are used.

Table B.1. Selection of quiescent galaxies in the redshift range $0 \geq z \geq 3$, using the Int Wide catalogue with its labels replaced by labels derived from the NUV $-$ r vs. $r - J_E$ or NUV $-$ r vs. $r - K$ colour-colour methods. For this test the *ugriz*, I_E, Y_E, J_E, H_E, $W1$, $W2$ and 20 cm bands were used, with detections required in all of the *Euclid* bands. Here, the `ARIADNE` pipeline was used in its *fast mode*. No results are shown for the NUV $-$ r vs. $r - K$ case in the redshift ranges $2 - 2.5$ or $2.5 - 3$, due to the very low number of sources labelled as quiescent by this method.
The columns are as follows:
(1) Redshift range in which the test was conducted;
(2) information on whether photometric redshifts were used to restrict the dataset to source in the specified redshift range ('Yes' or 'No');
(3) precision P for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - J_E$ method;
(4) recall R for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - J_E$ method;
(5) the F1-score for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - J_E$ method;
(6) precision P for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - K$ method;
(7) recall R for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - K$ method;
(8) the F1-score for quiescent galaxy selection when using labels derived from the NUV $-$ r vs. $r - K$ method;
(9) precision P for quiescent galaxy selection when using labels derived from the sSFR;
(10) recall R for quiescent galaxy selection when using labels derived from the sSFR;
(11) the F1-score for quiescent galaxy selection when using labels derived from the sSFR.

		NUV $-$ r vs. $r - J_E$ labels			NUV $-$ r vs. $r - K$ labels			sSFR labels		
Redshift range	Redshift binning	P	R	F1-score	P	R	F1-score	P	R	F1-score
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
0 – 3	No	0.86	0.79	0.82	0.79	0.74	0.77	0.84	0.74	0.79
0 – 0.25	Yes	0.51	0.56	0.53	0.41	0.46	0.43	0.72	0.67	0.69
0.25 – 0.5	Yes	0.82	0.81	0.82	0.78	0.74	0.76	0.81	0.80	0.80
0.5 – 0.75	Yes	0.86	0.85	0.86	0.84	0.81	0.83	0.86	0.85	0.85
0.75 – 1	Yes	0.90	0.90	0.90	0.80	0.85	0.82	0.85	0.87	0.86
1 – 1.25	Yes	0.90	0.90	0.90	0.79	0.86	0.82	0.88	0.88	0.88
1.25 – 1.5	Yes	0.81	0.85	0.83	0.69	0.80	0.74	0.82	0.84	0.83
1.5 – 2	Yes	0.77	0.79	0.78	0.61	0.72	0.66	0.80	0.82	0.81
2 – 2.5	Yes	0.70	0.69	0.69	–	–	–	0.62	0.69	0.65
2.5 – 3	Yes	0.62	0.65	0.62	–	–	–	0.40	0.57	0.37

B.9. Which observables are useful to select quiescent galaxies?

Heretofore, we have approached the question of the usefulness of different features (colours, magnitudes, etc.) in terms of whether they are informative and whether their inclusion significantly improves the quality of our classification models (see Sect. 4.2.4). However, it is also desirable to reach a deeper understanding of the usefulness of each feature, and how their usefulness depends on the circumstances under which quiescent galaxies are to be selected. For instance, the colours that are most useful for selecting quiescent galaxies in one redshift range may be less useful in another.

Moreover, while it may be self-evident that the inclusion of photometry in various optical bands allows for better characterisation of galaxy SEDs, and thus more accurate classification of galaxies, the degree of improvement may not always justify the cost of acquiring additional observations. Fig. 8 illustrates such an example, where the inclusion of *ugriz* photometry provides little or no significant improvement in the selection of quiescent galaxies at $z \gtrsim 1.25$, compared to when only the *Euclid* photometry is used (see Sect. 5.1.3 for further details).

Our objective for this section, therefore, is to provide analyses that can inform the planning of future surveys regarding which filters or frequencies would be important to include, and to guide the construction of new selection methods, be they colour-colour, template- fitting, or machine learning-based. For this we employ feature importance analysis (see also Sect. 4.2.4). The machine learning models we use for these analyses are trained using the `XGBoostClassifier` learning algorithm and the Int Wide mock catalogue, because its feature importance values are more robust, as discussed in Appendix A. We also perform A/B tests using our pipeline in *fast mode* to examine the impact on the F1-score from the addition of ancillary bands to the *Euclid* I_E, Y_E, J_E, $H_E$photometry.

It is worth noting that these analysis methods provide information about machine learning models and how they make use of the input data, rather than the data itself. As such, the information provided is limited by what the learning algorithms were able to learn from the data, and it is quite possible that additional relationships exist between features and the target that the learning algorithm was unable to find. Thus, while features found to be important are highly likely to be useful for the selection of quiescent galaxies, features found to be unimportant (or less important) may nevertheless hold hidden information that other selection methods (e.g. template fitting) may find useful when selecting quiescent galaxies.

B.9.1. Which single optical band is most useful?

In Fig. B.9 we show the effect of adding one optical band to the *Euclid* photometry set, when selecting quiescent galaxies in specific redshift ranges from the Int Wide catalogue using `XGBoostClassifier`, with no foreknowledge of redshifts. In addition, Fig. B.10 shows the improvement in F1-score due to the inclusion of each optical band.

First, we examine feature importance as a function of redshift when only the *Euclid* data are available (Fig. B.9, top left). In this circumstance, there is no single, decisive broad-band colour for the selection of quiescent galaxies at $z \lesssim 0.5$. The lack of sensitivity to the 4000 Å break, or to emission blueward thereof, results in models with relatively low F1-scores (see Fig. 8). In this

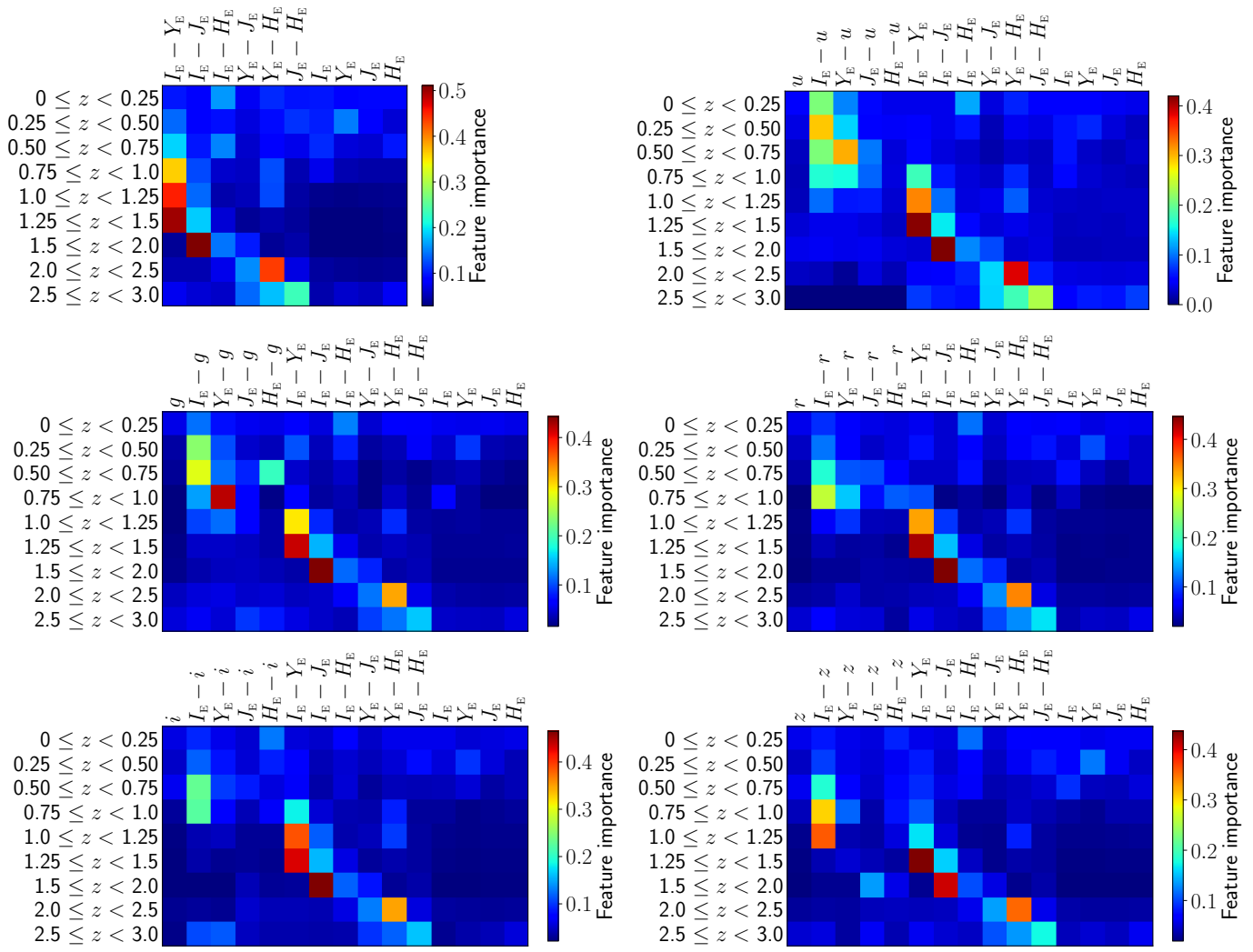

Fig. B.9. Visual representation of the feature importances derived from `XGBoostClassifier` models trained to select quiescent galaxies using only *Euclid* photometry and colours (top left panel), or *Euclid* photometry with the addition of one ground based optical band, and the relevant broad-band colours. The optical bands are u (top right), g (middle left), r (middle right), i (bottom left), and z (bottom right). The `XGBoostClassifier` models were trained without foreknowledge of the redshifts.

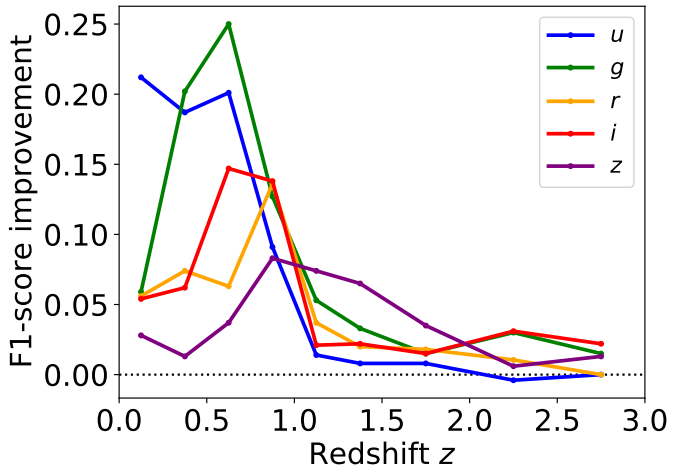

Fig. B.10. The effect on the F1-score due to the including one additional band with the *Euclid* photometry set. The `ARIADNE` pipeline was run in *fast mode* using the `LightGBMClassifier` base learner, or `XGBoostClassifier` at $2.5 < z < 3$, without foreknowledge of galaxy redshifts.

regime, our `XGBoostClassifier` models assign a fairly similar importance value to each feature.

Above $z = 0.4$, the 4000 Å break becomes redshifted into the I_E band, becoming potentially detectable via one or more of the *Euclid* broad-band colours. In the range $0.5 < z < 1.5$, the colour $I_E - Y_E$ is the most sensitive to the presence of the 4000 Å break, and our `XGBoostClassifier` models consider this feature to be the single most important for these redshifts. At $z > 1.5$, where the 4000 Å break is now redshifted into the near-IR bands, other colours become more important: $I_E - J_E$ at $1.5 < z < 2$, $Y_E - H_E$ at $2 < z < 2.5$, and $J_E - H_E$ at $2.5 < z < 3$.

The addition of u-band photometry allows the 4000 Å break to be detected at $z \lesssim 0.5$, resulting in significantly increased F1-scores in this redshift range (see Fig. B.10). In this case, $I_E - u$ and/or $Y_E - u$ become the most important in the redshift range $0 < z < 0.75$ by a large margin. Features using the u-band photometry continue to have significant (or non-zero) importance up to $z \sim 2$, before dropping to ~ 0 at $z > 2.5$.

In the case where models are trained using features derived from the g-band and *Euclid* photometry, colours involving g are found to be the most important within the range $0.25 < z < 1$. While the inclusion of g also helps significantly at $z < 0.25$, the

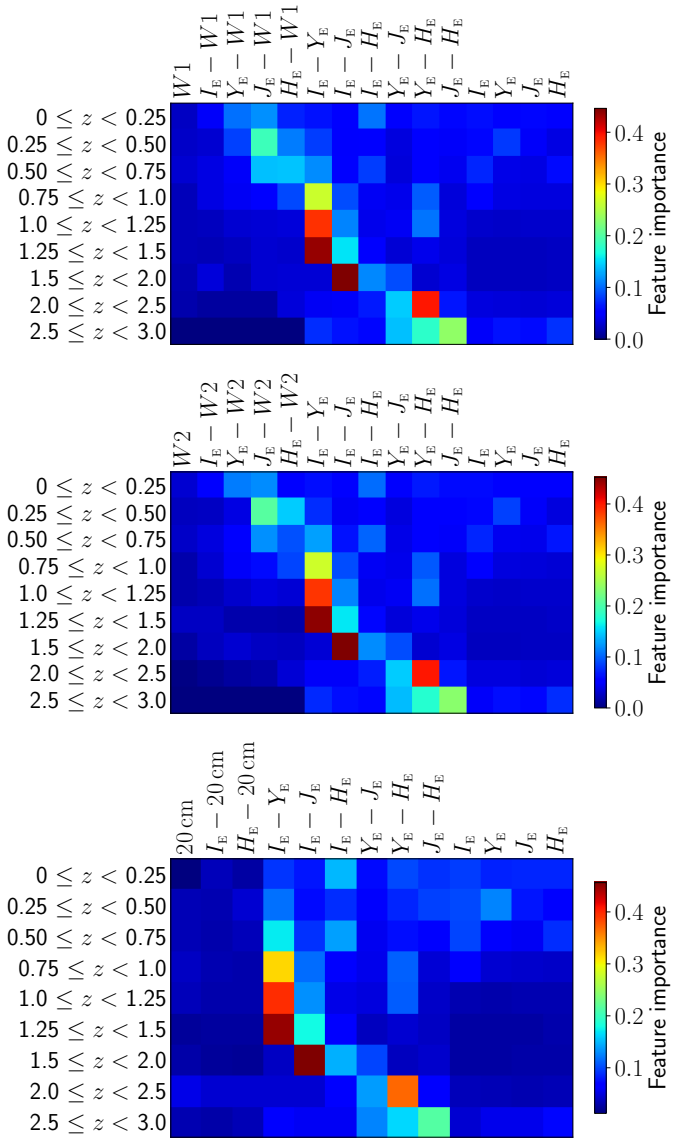

Fig. B.11. Similar to Fig. B.9, but showing the feature importances when the $W1$ (top), $W2$ (middle) or 20 cm (bottom) bands are included with the *Euclid* photometry.

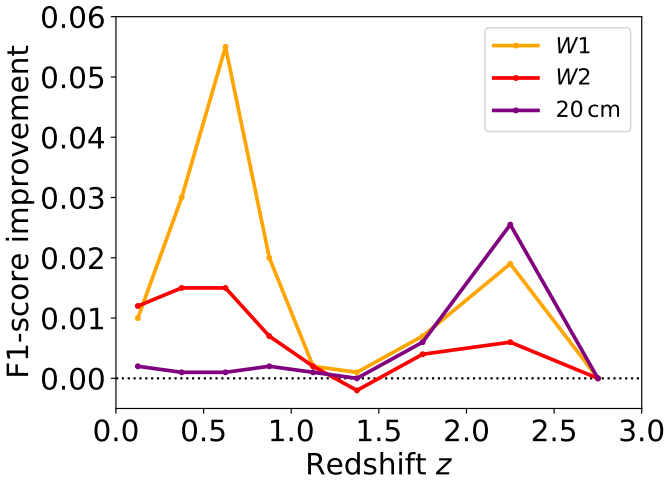

Fig. B.12. Similar to Fig. B.10, but showing the effect on the F1-score due to the including $W1$, $W2$, or the 20 cm radio band with the *Euclid* photometry when training `LightGBMClassifier` ($z < 2.5$) or `XGBoostClassifier` ($2.5 < z < 3$) models.

B.9.2. Importance of long wavelength bands

We repeat the process described in Appendix B.9.1, this time considering the addition of the $W1$, $W2$, or 20 cm radio band and related broad-band colours. The data for these three bands is very sparse, with detection fractions of 0.053, 0.024, and 0.0076 for $W1$, $W2$, and 20 cm, respectively.

The feature importance as a function of redshift is shown in Fig. B.11, and the improvement from including the $W1$, $W2$, or 20 cm bands with the *Euclid* photometry is shown in Fig. B.12. Despite being very sparse, each of the three bands provide a significant improvement in F1-score within the redshift ranges $0 < z < 1$ and $1.5 < z < 3$. We find little or no improvement evident in the redshift range $1 < z < 1.5$.

The usefulness of the 20 cm band for discriminating quiescent galaxies from star-forming galaxies is intriguing and somewhat surprising. Clearly, there is a correlation (or correlations) between the galaxy class and the presence of radio continuum emission. We speculate that this may be due to a subset of star-forming galaxies being detected in radio continuum, and/or the presence of radio-loud massive elliptical galaxies whose X-ray emission is below the COSMOS2015 detection threshold.

improvement is considerably smaller compared to when the u is used, because the 4000 Å break is blueward of the u-band's wavelength range. While the importance values of features using the g-band decrease substantially at $z > 1.25$, these features remain useful even in the $2.5 < z < 3$ bin.

When the r, i, or z band is used together with the *Euclid* bands, the situation is similar to that described above for g, with the main differences being the substantially lower feature importances at $0.25 < z < 0.5$ for colours that use either of the three bands, and the relatively high importance of $I_E - z$ at $1 < z < 1.25$.

In summary, there is no single 'ideal' optical band to include alongside the *Euclid* bands when selecting quiescent galaxies, and a trade-off should be made depending on whether low-redshift ($z \lesssim 0.25$) or high-redshift ($z \gtrsim 2$) galaxies are required. Of course, one may circumvent this choice if *ugriz* photometry is available.

B.10. Selection of quiescent galaxies from SPRITZ: Euclid Deep Survey

To complement the results presented in Sect. 5.1.1, we have also tested the selection of quiescent galaxies using the simulated Euclid Deep Survey from SPRITZ (Bisigello et al. 2021). Compared to the Euclid Deep Survey catalogues, the SPRITZ catalogue has the advantage of being complete down to the expected depth for the actual survey Deep Survey and the expected ancillary ground-based data. For the *Euclid* and *ugriz* bands, we adopt identical photometric uncertainties and detection limits to those used for the SED Deep catalogue given in Sect. 2. In the case of *Spitzer* Space Telescope IRAC bands, we adopt the following four cases based on Euclid Collaboration: Moneti et al. (2022):

- **Case 1**: no IRAC photometry;
- **Case 2**: IRAC photometry 3 σ depths of 24.55, 24.39, 22.61, and 22.17 mag in channel 1, 2, 3 and 4, respectively;

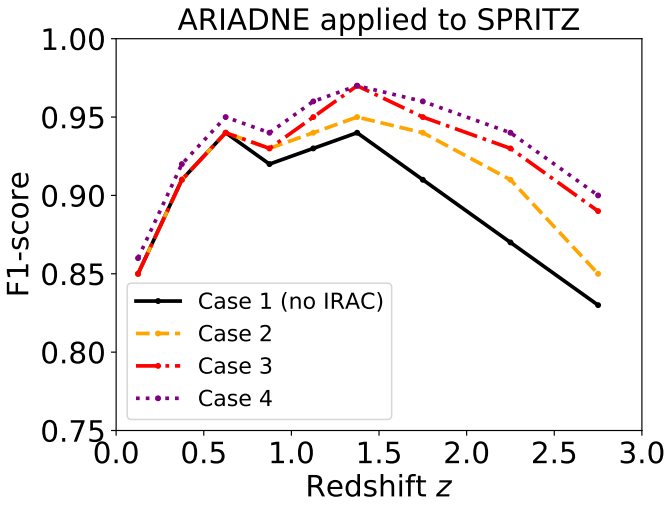

Fig. B.13. Results from applying the ARIADNE pipeline to the selection of quiescent galaxies from the SPRITZ Euclid Deep Survey simulated catalogue. Four cases described in Appendix B.10 are shown.

- **Case 3**: IRAC photometry 3 σ depths of 25.55, 25.39, 23.61, and 23.17 mag in channel 1, 2, 3 and 4, respectively;
- **Case 4**: IRAC photometry 3 σ depths of 26.55, 26.39, 23.61, and 23.17 mag in channel 1, 2, 3 and 4, respectively.

The ARIADNE pipeline was used in its default configuration, where five base-learners are employed. Galaxies not detected in one or more of the *Euclid* bands were removed from the dataset, as were galaxies containing an AGN. The results are shown in Table B.2 and Fig. B.13.

Appendix C: Number of detections in each catalogue and band

Table B.2. Selection of quiescent galaxies in the redshift range $0 \geq z \geq 3$, using the (1) Redshift range in which the test was conducted;
(2) information on whether photometric redshifts were used to restrict the dataset to source in the specified redshift range ('Yes' or 'No');
(3) precision P for quiescent galaxy selection for SPRITZ Case 1;
(4) recall R for quiescent galaxy selection for SPRITZ Case 1;
(5) the F1-score for quiescent galaxy selection for SPRITZ Case 1;
(6) precision P for quiescent galaxy selection for SPRITZ Case 2;;
(7) recall R for quiescent galaxy selection for SPRITZ Case 2;
(8) the F1-score for quiescent galaxy selection for SPRITZ Case 2;;
(9) precision P for quiescent galaxy selection for SPRITZ Case 3;;
(10) recall R for quiescent galaxy selection for SPRITZ Case 3;;
(11) the F1-score for quiescent galaxy selection for SPRITZ Case 3;;
(12) precision P for quiescent galaxy selection for SPRITZ Case 4;;
(13) recall R for quiescent galaxy selection for SPRITZ Case 4;;
(14) the F1-score for quiescent galaxy selection for SPRITZ Case 4.

		SPRITZ Case 1			SPRITZ Case 2			SPRITZ Case 3			SPRITZ Case 4		
Redshift range	Redshift binning	P	R	F1-score	P	R	F1-score	P	R	F1-score	P	R	F1-score
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
0 – 3	No	0.92	0.86	0.89	0.92	0.88	0.90	0.93	0.90	0.91	0.93	0.91	0.92
0 – 0.25	Yes	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.86	0.86
0.25 – 0.5	Yes	0.92	0.90	0.91	0.92	0.91	0.91	0.92	0.91	0.91	0.92	0.91	0.92
0.5 – 0.75	Yes	0.96	0.92	0.94	0.96	0.92	0.94	0.96	0.93	0.94	0.96	0.93	0.95
0.75 – 1	Yes	0.93	0.91	0.92	0.94	0.91	0.93	0.94	0.92	0.93	0.95	0.92	0.94
1 – 1.25	Yes	0.95	0.92	0.93	0.96	0.92	0.94	0.97	0.93	0.95	0.97	0.94	0.96
1.25 – 1.5	Yes	0.93	0.95	0.94	0.96	0.95	0.95	0.97	0.96	0.97	0.98	0.97	0.97
1.5 – 2	Yes	0.90	0.93	0.91	0.94	0.94	0.94	0.95	0.95	0.95	0.96	0.96	0.96
2 – 2.5	Yes	0.91	0.84	0.87	0.93	0.89	0.91	0.95	0.91	0.93	0.95	0.93	0.94
2.5 – 3	Yes	0.82	0.83	0.83	0.86	0.84	0.85	0.89	0.89	0.89	0.91	0.89	0.90

Table C.1. The number of $3\,\sigma$ detections in each of the bands, for each mock catalogue. Also shown are the numbers of quiescent galaxies. The SED catalogues do not contain 20 cm radio band.

Catalogue	I_E	Y_E	J_E	H_E	u	g	r	i	z	$W1$	$W2$	20 cm	Quiescent
Int Wide	315755	212019	231039	250077	140782	226514	198564	194912	204649	10476	4704	1536	21998 (7.0 %)
Int Deep	517890	486394	491588	500299	499565	504416	490457	493018	499140	10476	4704	1698	30990 (6.0 %)
SED Wide	3249101	2056800	2270138	2455887	1498518	2330338	2091921	2040916	2031115	131703	65904	–	213837 (6.6 %)
SED Deep	5121526	4763050	4890667	4963038	3971472	4796448	4828112	4802416	4766807	134656	69493	–	303761 (5.9 %)