


Euclid: Identifying the reddest high-redshift galaxies in the *Euclid* Deep Fields with gradient-boosted trees[★]

T. Signor^{1,2,3} , G. Rodighiero^{3,4} ^{★★}, L. Bisigello^{3,4}, M. Bolzonella⁵, K. I. Caputi^{6,7}, E. Daddi⁸, G. De Lucia⁹, A. Enia^{10,11}, L. Gabarra¹², C. Gruppioni⁵, A. Humphrey^{13,14}, F. La Franca¹⁵, C. Mancini¹⁶, L. Pozzetti⁵, S. Serjeant¹⁷, L. Spinoglio¹⁸, S. E. van Mierlo⁶, S. Andreon¹⁹, N. Auricchio⁵, M. Baldi^{10,5,20}, S. Bardelli⁵, P. Battaglia⁵, R. Bender^{21,22}, C. Bodendorf²¹, D. Bonino²³, E. Branchini^{24,25,19}, M. Brescia^{26,27,28}, J. Brinchmann¹³, S. Camera^{29,30,23}, V. Capobianco²³, C. Carbone¹⁶, J. Carretero^{31,32}, S. Casas³³, M. Castellano³⁴, S. Cavuoti^{27,28}, A. Cimatti³⁵, R. Cledassou^{36,37} [†], G. Congedo³⁸, C. J. Conselice³⁹, L. Conversi^{40,41}, Y. Copin⁴², L. Corcione²³, F. Courbin⁴³, H. M. Courtois⁴⁴, A. Da Silva^{45,46}, H. Degaudenzi⁴⁷, A. M. Di Giorgio¹⁸, J. Dinis^{46,45}, F. Dubath⁴⁷, X. Dupac⁴¹, S. Dusini⁴⁸, A. Ealet⁴⁹, M. Farina¹⁸, S. Farrens⁵⁰, S. Ferriol⁴², S. Fotopoulou⁵¹, E. Franceschi⁵, S. Galeotta⁹, B. Garilli¹⁶, W. Gillard⁵², B. Gillis³⁸, C. Giocoli^{5,53}, A. Grazian⁴, F. Grupp^{21,54}, L. Guzzo^{55,19,56}, S. V. H. Haugan⁵⁷, I. Hook⁵⁸, F. Hormuth⁵⁹, A. Hornstrup^{60,61}, K. Jahnke⁶², M. Kümmel²², S. Kermiche⁵², A. Kiessling⁶³, M. Kilbinger⁸, T. Kitching⁶⁴, H. Kurki-Suonio^{65,66}, S. Ligi²³, P. B. Lilje⁵⁷, V. Lindholm^{65,66}, I. Lloro⁶⁷, D. Maino^{55,16,56}, E. Maiorano⁵, O. Mansutti⁹, O. Marggraf⁶⁸, N. Martinet⁶⁹, F. Marulli^{70,5,20}, R. Massey⁷¹, E. Medinaceli⁵, M. Melchior⁷², Y. Mellier^{73,74}, M. Meneghetti^{5,20}, E. Merlin³⁴, M. Moresco^{70,5}, L. Moscardini^{70,5,20}, E. Munari⁹, R. C. Nichol⁷⁵, S.-M. Niemi⁷⁶, C. Padilla³¹, S. Paltani⁴⁷, F. Pasian⁹, K. Pedersen⁷⁷, V. Pettorino⁷⁸, S. Pires⁵⁰, G. Polenta⁷⁹, M. Poncet³⁶, L. A. Popa⁸⁰, F. Raison²¹, A. Renzi^{3,48}, J. Rhodes⁶³, G. Riccio²⁷, E. Romelli⁹, M. Roncarelli⁵, E. Rossetti¹⁰, R. Saglia^{22,21}, D. Sapone⁸¹, B. Sartoris^{22,9}, P. Schneider⁶⁸, T. Schrabback⁸², A. Secroun⁵², G. Seidel⁶², S. Serrano^{83,84,85}, C. Sirignano^{3,48}, G. Sirri²⁰, L. Stanco⁴⁸, C. Surace⁶⁹, P. Tallada-Crespí^{86,32}, H. I. Teplitz⁸⁷, I. Tereno^{45,88}, R. Toledo-Moreo⁸⁹, F. Torradeflot^{32,86}, I. Tutusaus⁹⁰, E. A. Valentijn⁶, T. Vassallo^{22,9}, A. Veropalumbo^{19,25}, Y. Wang⁸⁷, J. Weller^{22,21}, O. R. Williams⁹¹, J. Zoubian⁵², E. Zucca⁵, C. Burigana^{92,93}, and V. Scottez^{73,94}

(Affiliations can be found after the references)

Received 27 November 2023 / Accepted 4 February 2024

ABSTRACT

Context. ALMA observations show that dusty, distant, massive ($M_* \gtrsim 10^{11} M_\odot$) galaxies usually have a remarkable star-formation activity, contributing of the order of 25% of the cosmic star-formation rate density at $z \approx 3-5$, and up to 30% at $z \sim 7$. Nonetheless, they are elusive in classical optical surveys, and current near-IR surveys are able to detect them only in very small sky areas. Since these objects have low space densities, deep and wide surveys are necessary to obtain statistically relevant results about them. *Euclid* will potentially be capable of delivering the required information, but, given the lack of spectroscopic features at these distances within its bands, it is still unclear if *Euclid* will be able to identify and characterise these objects.

Aims. The goal of this work is to assess the capability of *Euclid*, together with ancillary optical and near-IR data, to identify these distant, dusty, and massive galaxies based on broadband photometry.

Methods. We used a gradient-boosting algorithm to predict both the redshift and spectral type of objects at high z . To perform such an analysis, we made use of simulated photometric observations that mimic the *Euclid* Deep Survey, derived using the state-of-the-art Spectro-Photometric Realizations of Infrared-selected Targets at all- z (SPRITZ) software.

Results. The gradient-boosting algorithm was found to be accurate in predicting both the redshift and spectral type of objects within the simulated *Euclid* Deep Survey catalogue at $z > 2$, while drastically decreasing the runtime with respect to spectral-energy-distribution-fitting methods. In particular, we studied the analogue of HIEROs (i.e. sources selected on the basis of a red $H - [4.5] > 2.25$), combining *Euclid* and *Spitzer* data at the depth of the Deep Fields. These sources include the bulk of obscured and massive galaxies in a broad redshift range, $3 < z < 7$. We find that the dusty population at $3 \lesssim z \lesssim 7$ is well identified, with a redshift root mean squared error and catastrophic outlier fraction of only 0.55 and 8.5% ($H_E \leq 26$), respectively. Our findings suggest that with *Euclid* we will obtain meaningful insights into the impact of massive and dusty galaxies on the cosmic star-formation rate over time.

Key words. methods: statistical – galaxies: active – galaxies: evolution – galaxies: high-redshift – infrared: galaxies

1. Introduction

In the last few decades, a major effort has been dedicated to the statistical identification of galaxies over a wide range of

redshifts. Multi-wavelength observational campaigns (from the X-ray to the radio spectral regime) in the deepest extragalactic fields have allowed the reconstruction of the average properties of various galaxy populations and their evolution. A fundamental result is the measurement of the star-formation rate density (SFRD) of the Universe (e.g. Madau & Dickinson 2014). The SFRD reached a peak at $z \approx 1-3$ (so-called ‘cosmic noon’),

[★] This paper is published on behalf of the *Euclid* Consortium.

^{★★} e-mail: giulia.rodighiero@unipd.it

[†] Deceased.

rapidly declining to the current value. Several works have shown that the fraction of the SFRD obscured by dust, and therefore not accounted for by optical/UV surveys at $z > 2$, is likely not negligible and increases with redshift at least up to $z \approx 5-7$ (e.g. Novak et al. 2017; Gruppioni et al. 2020; Topping et al. 2022; Barrufet et al. 2023b; Fujimoto et al. 2023; Algera et al. 2023). A comprehensive study of high-redshift galaxies (well before cosmic noon) is of fundamental importance for our understanding of the early epochs of galaxy stellar mass assembly.

The classic technique for selecting sources at $z > 3$ relies on their broadband colours: the drop in brightness caused by the Lyman break (at 912 Å in the rest frame) and/or the Lyman forest (between 912 and 1216 Å in the rest frame) is measured. The selected objects are referred to as Lyman-break galaxies (LBGs). However, while this approach is straightforward to apply, it is also affected by significant incompleteness and contamination; in particular, as a consequence of their redder UV slopes and relative faintness, LBG selection is known to be notably biased against massive galaxies ($M_* \gtrsim 10^{11} M_\odot$; van Dokkum et al. 2006; Bian et al. 2013). Indeed, various massive, non-UV-selected galaxy populations have been detected and spectroscopically confirmed at $z \gtrsim 3$ (e.g. Daddi et al. 2009; Huang et al. 2014). Among these, optically faint sub-millimetre galaxies (SMGs; i.e. galaxies discovered at sub-millimetre wavelengths) have been particularly interesting, as they can be undetectable at high redshifts, even with the deepest optical/near-IR imaging (e.g. Frayer et al. 2000; Wang et al. 2019; Smail et al. 2021).

These massive and dusty galaxies have low space densities and a remarkable star-formation activity, contributing up to $\approx 20-25\%$ of the cosmic star-formation rate density (CSFRD) at $z \approx 3-5$ (Gruppioni et al. 2020; Talia et al. 2021; Enia et al. 2022; Xiao et al. 2023). In particular, the CSFRD estimated for H -faint galaxies ($H \gtrsim 26.4$, $< 5\sigma$) in Sun et al. (2021) is $\approx 8\%$ of the CSFRD at this epoch (Madau & Dickinson 2014); the values suggested by Williams et al. (2019) and Gruppioni et al. (2020) are approximately 2 to 3 times larger. Despite the importance of these galaxies, given their faintness and non-detection at most wavelengths, most of their physical properties remain highly uncertain, except for in a very few cases with spectroscopic confirmations (e.g. Wang et al. 2019; Caputi et al. 2021). More recently, some attempts to characterise these galaxy populations have been performed thanks to the *James Webb* Space Telescope (JWST; e.g. Pérez-González et al. 2023; Barrufet et al. 2023a; Rodighiero et al. 2023; Barro et al. 2024; Bisigello et al. 2023), but the areas observed remain small. It is thus clear that mapping the full cosmic star-formation history and understanding the early phases of massive galaxy formation require the study of the star formation in massive galaxy populations to be as comprehensive as possible at $z > 3$.

To this end, several colour-selection methods have been proposed to identify this optically faint massive galaxy population. In particular, Wang et al. (2016) present a method based on the $H - [4.5]/J - H$ diagram that enables a rather clean selection of $z > 3$ galaxies, which are usually called HIEROs (Extremely Red Objects with H and IRAC colours). However, the separation of high-redshift galaxies from low-redshift contaminants remains difficult. Multi-wavelength observations that sample the spectral energy distribution (SED) of the sources are also widely used to measure their photometric redshifts and physical parameters (such as stellar mass, star-formation rate, stellar age, and extinction; e.g. Weaver et al. 2022; Laigle et al. 2016; Ilbert et al. 2009). The most used technique that leverages this kind of information is template-fitting (e.g. Benitez 2000), which uses a set of theoretical or empirical SED templates for the estimations.

In recent years, the large amount of data provided by a wealth of extragalactic surveys has enabled the use of supervised machine learning techniques, in which the mappings between inputs (photometry in different bands) and outputs (redshift or other physical properties) are learned through a reference, or training, sample. Their most obvious advantage is the much higher efficiency in memory usage and computational time with respect to SED-fitting techniques, for example seconds instead of days when dealing with millions of objects.

Nonetheless, although these new empirical methods were found to outperform even the accuracy of template-based methods (Abdalla et al. 2011) and although their use has become very common (e.g. Ball et al. 2008; Pasquet et al. 2019; Liu et al. 2019; Euclid Collaboration 2023b), no detailed study has as yet focused on distant galaxies ($z > 3$). In fact, the photometric redshift accuracy of these objects is not well determined because of a lack of sufficiently large reference samples with spectroscopic redshifts and the paucity of deep near-IR data. Furthermore, massive galaxies are rare, and wide fields are needed to obtain statistically relevant results.

Taking all these issues into account, the upcoming *Euclid* Space Telescope (Laureijs et al. 2011) will open up new possibilities for the study of these objects by observing a large area of the sky at near-IR wavelengths. Most of the mission's observations will comprise a wide survey, covering approximately 15 000 deg² down to a 10σ depth of 24.5 mag in the visible filter and down to a 5σ depth of 24 mag at near-IR wavelengths. A deep survey two magnitudes deeper than the wide survey will also be conducted over 50 deg² in the *Euclid* Deep Fields (Euclid Collaboration 2022a). The *Euclid* Deep Fields are expected to contain millions of $z > 3$ galaxies and therefore enable studies of early galaxy formation and evolution with unprecedented statistical significance.

Given the limited number of *Euclid* photometric bands (a visible band, I_E , and three near-IR bands, Y_E , J_E , and H_E), efforts are being made to complement *Euclid* space-based data with ground-based data in the UV to visible spectral range (Laureijs et al. 2011; Ibata et al. 2017). Combined with existing *Spitzer* Space Telescope surveys (near-IR light; Capak et al. 2016; Masters et al. 2019, e.g. Euclid Collaboration 2022b), these data will provide a comprehensive description of the SEDs of galaxies up to the epoch of reionisation.

It is thus essential to assess the capability of the *Euclid* filters (combined with the ancillary data) to identify high-redshift galaxies, in particular the massive and dusty population, through precise photometric redshifts, and to characterise their spectral types. In this work we propose performing such a task by using simulated *Euclid* photometric observations. We adopted simulated data from Spectro-Photometric Realizations of Infrared-selected Targets at all- z (SPRITZ; Bisigello et al. 2021). This choice was based on the fact that this phenomenological simulation, which includes both star-forming galaxies and active galactic nuclei (AGN), reproduces the statistics and the evolution of the far-IR sources well (number counts and luminosity functions), that is to say, the dusty massive population.

Since the methodology proposed in this work (i.e. a gradient-boosting algorithm) returns a photometric redshift for each source selected within the simulated lightcone, we present the general performance for the optimised $z > 2$ range (where the algorithm is trained). However, in this paper we focus our attention on a specific class of objects: the HIEROs. They represent an ideal population for testing the ability of *Euclid*, combined with ancillary data, to recover dark galaxies beyond cosmic noon.

The structure of the paper is as follows. In Sect. 2 the simulated catalogue and the data used are introduced. In Sect. 3 our main methods are described, and the obtained results are presented in Sect. 4. Finally, Sect. 5 summarises the main results and the perspectives for the *Euclid* mission.

Throughout the paper, we assume a Λ cold dark matter cosmology with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.27$, and $\Omega_\Lambda = 0.73$. All magnitudes are in the AB system (Oke & Gunn 1983).

2. Data

2.1. The SPRITZ simulation

SPRITZ (Bisigello et al. 2021) is a state-of-the-art simulation based on a set of observed galaxy stellar mass functions and luminosity functions, mainly in the IR, derived for different galaxy populations. In this work we consider the version 1.13 of the simulation, with updates the dwarf irregular galaxy stellar mass function used in input and includes CO and [C II] line luminosities, as presented in Bisigello et al. (2022). Briefly, each simulated galaxy is assigned a unique SED template, taken from a set of 35 empirical templates (Polletta et al. 2007; Rieke et al. 2009; Gruppioni et al. 2010; Bianchi et al. 2018). These templates are of low- z galaxies, but they represent a good description of galaxies observed by *Herschel* up to $z = 3.5$ (Gruppioni et al. 2013) and by the Atacama Large Millimeter/submillimeter Array (ALMA) at $z = 6$ (Gruppioni et al. 2020). A set of empirical and theoretical relations is then used to link each source to its physical properties, such as stellar mass, star-formation rate and AGN contribution. The galaxy populations included in the simulations are spirals, starbursts (SBs), ellipticals, dwarf irregulars, AGN, and composite AGN. The AGN population includes both type-1 and type-2 AGN, and this classification is based on the optical/UV part of their spectrum. Composite AGN include objects with an AGN component that is not the dominant source of bolometric emission, because they are intrinsically faint – referred to as star-forming AGN (SF-AGN) – or because they are extremely obscured by dust – referred to as starburst AGN (SB-AGN). The choice of such galaxy populations was motivated by the variety of galaxies observed by *Herschel* up to $z = 3.5$.

The resulting mock catalogues are found to be consistent with a large variety of observations, including the stellar mass versus star-formation rate relation, luminosity functions (LFs) and number counts from X-ray to radio, demonstrating that this simulation is suitable for making predictions for a set of future surveys operating particularly, but not only, at IR wavelengths. In particular, we highlight that the simulation is in agreement with the IR LF observed at $z \sim 6$ (Gruppioni et al. 2020) as well as the [C II] and CO LFs at $z = 4\text{--}6$ (e.g. Riechers et al. 2019; Loiacono et al. 2021; Boogaard et al. 2023). For more details on SPRITZ, we refer to Bisigello et al. (2021).

In this work, we made use of this simulation for forecasting high-redshift galaxy identification in the *Euclid* Deep Fields, noting that the results obtained in this work are an optimistic version of the ones that will be obtained from real data.

2.2. The simulated *Euclid* Deep Fields catalogue

As reported in Sect. 1, the *Euclid* Deep Survey will observe an area of at least 40 deg^2 in four filters: I_E from the visible instrument (VIS, Cropper et al. 2016) and Y_E , J_E , and H_E from the Near Infrared Spectrometer and Photometer (NISF,

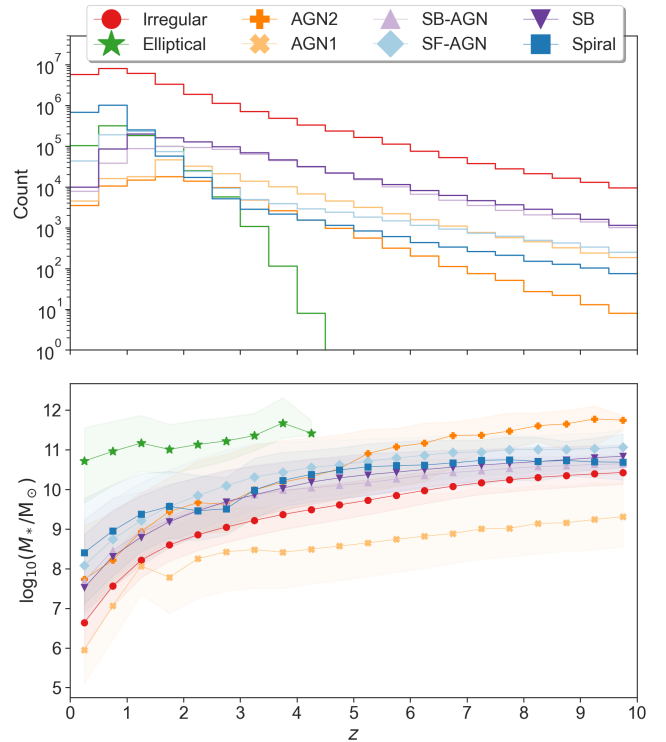


Fig. 1. Redshift distribution (top) and stellar mass as a function of redshift (bottom) of the simulated galaxies obtained with SPRITZ for the *Euclid* Deep Survey, colour-coded by their SED type (see the legend). Markers represent the median value in the redshift bin and shaded areas the corresponding 68% coverage interval. For details about the mass function distribution at different redshift ranges, we refer the reader to Bisigello et al. (2021).

Euclid Collaboration 2022d), with 5σ depths between 26 and 27.3 AB magnitudes.

Including galaxies with at least one *Euclid* filter with signal-to-noise ratio ≥ 2 , this combination of depth and area results in a simulated catalogue with a total of 33 650 754 objects with redshifts from 0 to 10. The redshift distribution and the stellar mass as a function of redshift are shown in Fig. 1. Irregular galaxies with stellar masses below $M_* \approx 10^{11} M_\odot$ are expected to dominate the number counts of the survey. At the same time, given the brightness of AGN1, we expect to observe them down to $M_* \approx 10^8 M_\odot$ even at the highest redshifts.

In this simulated catalogue, we also included photometry in additional filters, from *Rubin* and the *Spitzer* Infrared Array Camera (IRAC, Fazio et al. 2004), as these ancillary observations will be available to complement *Euclid* space-based data (Euclid Collaboration 2022a), at least for a fraction of the observed fields. It should be noted, however, that the IRAC photometry is affected by source confusion at faint magnitudes. Specifically, at $\text{IRAC}/3.6\mu\text{m} > 22$, a de-blending based on the available *Euclid* data (Euclid Collaboration 2022b) will be necessary when dealing with real observations. The expected 5σ and 2σ depths of the *Euclid*, *Rubin* and IRAC bands are reported in Table 1.

Photometric errors

In the SPRITZ lightcone used in this work, only the *Euclid* photometric bands have an associated observational uncertainty (i.e. the errors on the simulated magnitudes). We perturbed the *Rubin* and *Spitzer* fluxes to mimic realistic observations. In particular,

Table 1. Depths of the filters considered in this work.

Band	5σ Depth	2σ Depth
I_E	28.2	29.2
Y_E	26.3	27.3
J_E	26.5	27.5
H_E	26.4	27.4
<i>Rubin/u</i>	26.8	27.8
<i>Rubin/g</i>	28.4	29.4
<i>Rubin/r</i>	28.5	29.5
<i>Rubin/i</i>	28.3	29.3
<i>Rubin/z</i>	28.0	29.0
IRAC/3.6 μm	24.8	25.8
IRAC/4.5 μm	24.7	25.7

Notes. Expected 5 and 2σ depths of four *Euclid* filters, *Rubin ugriz* filters for the Deep Drilling Fields (Foley et al. 2018) and *Spitzer* IRAC 3.6, 4.5 μm (Euclid Collaboration 2022b) filters, as considered in this work.

the *Rubin* total photometric error has both a systematic and a random contribution and can be written as (Ivezić et al. 2019)

$$\sigma^2 = \sigma_{\text{sys}}^2 + \sigma_{\text{rand}}^2, \quad (1)$$

where σ_{rand} is the random photometric error and σ_{sys} is the systematic one. Given that the *Rubin* telescope is designed to have a systematic error below 0.005 mag, we decided to neglect it in this work. The random photometric uncertainty can be written as a function of the magnitude (Ivezić et al. 2019):

$$\sigma_{\text{rand}}^2 = (0.04 - \gamma)x + \gamma x^2 \quad (2)$$

$$x \equiv 10^{0.4(m-m_5)}, \quad (3)$$

where m_5 is the 5σ depth (see Table 1) and γ is a parameter equal to 0.039 for the *g*, *r*, *i*, and *z* bands and 0.038 for *u*. The photometry for each galaxy in each *Rubin* filter was then derived by randomly sampling a Gaussian distribution with mean equal to the true value and standard deviation σ_{rand} .

Similarly, for the two *Spitzer* bands (IRAC/3.6 μm and IRAC/4.5 μm), we considered errors from Laigle et al. (2016) and parametrised them using Eqs. (2)–(3) with m_5 as in Table 1 and $\gamma = 0.038$.

2.3. Photometric selections

The first step in our analysis was to validate the simulated *Euclid* Deep Field catalogue’s compatibility with a set of observed photometric diagnostics available from the literature, with a focus on high- z galaxies. The main focus in this work is the dusty and massive galaxy populations at $3 \lesssim z \lesssim 7$. For their selection, we relied on the evolutionary $H - [4.5]$ tracks of a set of theoretical galaxy SED templates for $z > 3$, assuming the colour cut-off suggested by Wang et al. (2016):

$$H - [4.5] > 2.25. \quad (4)$$

This colour was proposed to select old or dusty galaxies at $z > 3$. Objects satisfying this criterion are referred to as HIEROs. In particular, it was found that almost none of the spectroscopically confirmed LBGs at $z > 3$ satisfies this criterion. These are examples of the types of objects missed by conventional UV/optical surveys that we aim to recover with the future *Euclid* Deep Survey. Furthermore, based on the colour tracks of theoretical models and their photometric redshifts, HIEROs can be separated in

two main classes:

$$\text{blue HIEROs,} \quad H - [4.5] > 2(J - H) + 1.45; \quad (5)$$

$$\text{red HIEROs,} \quad H - [4.5] \leq 2(J - H) + 1.45. \quad (6)$$

According to Wang et al. (2016), blue HIEROs are dominated by normal massive and dusty star forming galaxies at $z \gtrsim 3$. Red HIEROs, instead, include a mix of lower $z \sim 2$ –3 dusty star forming objects and passive galaxies at $z \sim 3$ –4. Passive galaxies are expected to be the most massive systems at any cosmic epoch, and are thus relevant for our study.

Figure 2 (left panel) shows the $H_E - [4.5]/J_E - H_E$ diagram for the *Euclid* Deep Survey simulated catalogue with $2 \leq z \leq 8$ and fluxes brighter than their 5σ detection limits, coloured by redshift. To better highlight the different populations, in the top-right panel we report their redshift distributions: a clear difference arises between HIEROs and other objects, with 99% of HIEROs being at $z > 3$ and 43% of non-HIEROs being above the same redshift.

The bottom-right panel shows, instead, the extinction properties for red and blue HIEROs (in terms of A_V distributions). As expected, their red colours are mostly due to the presence of dust, with the bulk of $A_V \sim 4$ mag and spanning values up to 5.5 mag. We note that by selection both blue and red HIEROs include highly extinguished objects (Wang et al. 2016), so it is natural to observe consistent A_V distributions (while keeping in mind that blue and red HIEROs peak at different cosmic epochs).

Given that in our simulation red HIEROs dominate the number densities of red sources at $z \sim 4$, while blue HIEROs populate the higher redshift queue up to $z \sim 7$ –8, we included in the following discussion the study of the overall class of *Euclid* sources with $H - [4.5] > 2.25$, focusing on the $3 < z < 7$ redshift range (where the bulk of HIEROs lie). To check the representativeness of our HIERO mock catalogue, we compared the overall number densities with the observations of Wang et al. (2016), who selected HIEROs with $[4.5] < 24$ over an area of 350 arcmin² and detected a cleaned sample of 285 sources. Restricting our selection (i.e. requiring at least four detections at $S/N > 2$ in the considered bands; see Sect. 3.2.2) to the same magnitude limit of Wang et al. (2016), we identify 155 objects over the same area. However, if we mimic the Wang et al. (2016) selection by simply requiring a 5σ detection at 4.5 μm , and $[4.5] < 24$, the number of predicted HIEROs increases to 425 (over 350 arcmin²). The variance on the predicted space densities, related to the assumed selection function, shows that the our mock catalogue provides a statistical sampling of the HIERO population consistent with the real world.

To understand the role of HIEROs in the stellar mass assembly, we report in Fig. 3 (black lines) the overall stellar mass function for galaxies in the redshift range $4 < z < 6$ from SPRITZ, split in two z bins. The mass function limited to HIEROs is also reported (red lines), showing the dominant contribution of this population to the massive end of the stellar mass function at $z > 4$. Both at $z = 4.5$ and 5.5, HIEROs constitute approximately 93% of the galaxies with $M_* > 10^{11.5} M_\odot$ and 32% of the galaxies with $M_* > 10^{10.5} M_\odot$. The figure includes the observed mass function of HIEROs at $z \sim 4$ –5 from Wang et al. (2016), confirming the consistency of our model with the observations, within the uncertainties.

Finally, in Fig. 4 we show the most representative SED templates in SPRITZ for this population of galaxies, specifically the four most numerous templates used to generate red and blue HIEROs.

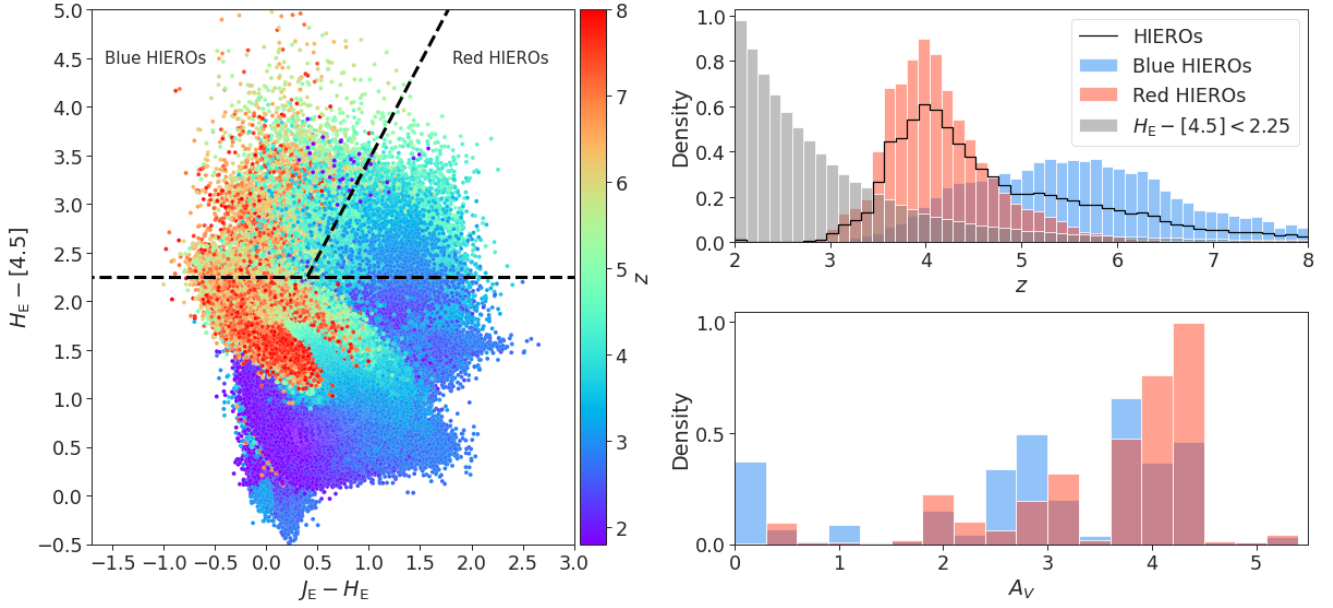


Fig. 2. HIEROs selection, redshift and extinction in the Simulated *Euclid* Deep Survey Catalogue. Left: $H_E - [4.5]$ versus $J_E - H_E$ diagram for the simulated *Euclid* Deep survey catalogue (5σ depth) colour-coded by redshift. Top right: Redshift distribution of red HIEROs (red), blue HIEROs (blue), and galaxies with $H_E - [4.5] < 2.25$ (grey). Bottom right: Optical extinction distribution of red HIEROs (red) and blue HIEROs (blue). We normalised each histogram to obtain an area equal to unity.

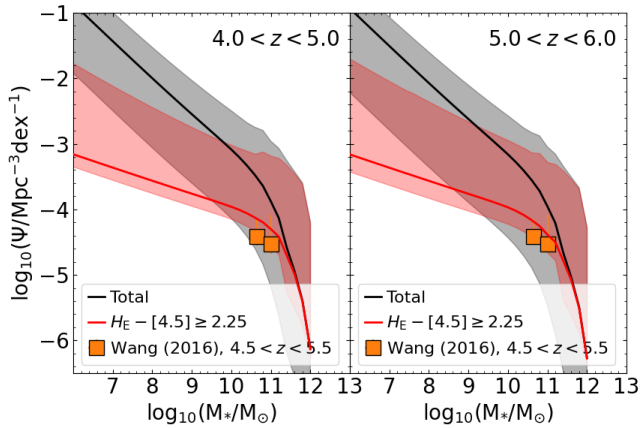


Fig. 3. Contribution of HIEROs to the galaxy stellar mass function in SPRITZ at $z > 4$ (red line). The grey and red shaded areas include the errors of the initial luminosity functions used to derive simulated galaxies and the uncertainties due to the high- z extrapolation ($z > 3$). The observed stellar mass function for HIEROs measured by Wang et al. (2016) at $4.5 < z < 5.5$ is reported in both panels as filled orange squares.

3. Methods

In this section, we introduce the methods and the metrics that we used to estimate the capability of the future *Euclid* Space Telescope to identify high-redshift galaxies.

3.1. Gradient-boosted trees

We considered a gradient-boosting approach to independently predict both the redshifts and the SED types, based on the observed fluxes in different bands. This corresponds to a regression and a classification problem, respectively. Other methods have been proposed in the literature to classify differ-

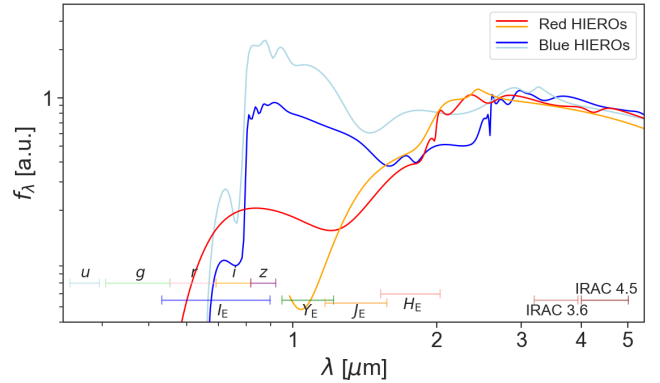


Fig. 4. The four most numerous SED templates used in SPRITZ for the generation of red HIEROs (red and orange) as considered at $z = 4$, and blue HIEROs (blue and light blue) at $z = 5.5$. The flux is in arbitrary units. We also show the wavelength coverage of the photometric bands considered in this work.

ent galaxy populations in *Euclid* (e.g. Bisigello et al. 2020, *Euclid* Collaboration 2023a) or to derive photometric redshifts (e.g. *Euclid* Collaboration 2020, Bisigello et al. in prep.), but none of these is focused on high- z galaxies.

The method proposed in this paper is a type of ensemble algorithm, in which the relationships between the features \mathbf{x} and the target variables $y = f(\mathbf{x})$ are learned by sequentially fitting new models: new decision trees (a representation of the decision tree model is shown in Fig. 5) are constructed to be maximally correlated with the negative gradient of the loss function associated with the ensemble. The loss function is chosen according to the task, while structural and learning parameters (referred to as hyperparameters) have to be tuned in a data-driven fashion.

The algorithm used in this work is implemented in the software package XGBoost (Chen & Guestrin 2016). Here we briefly summarise its main features while for more details we

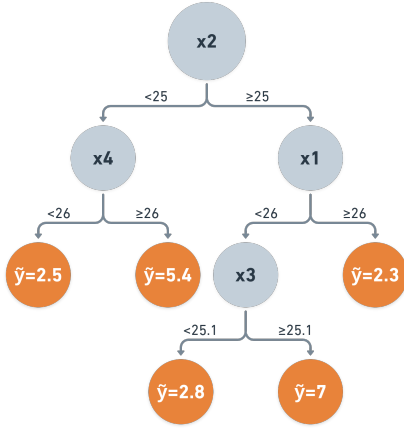


Fig. 5. Decision tree model. Grey nodes are called internal nodes and the orange ones, which represent the tree predictions, leaf nodes. Features are indicated with $x1$, $x2$, $x3$, or $x4$ and predictions with \tilde{y} . Every internal node is labelled with an input feature. The arcs coming from a node labelled with an input feature (for example $x4$) to a leaf node (orange) are labelled with each of the possible values of the target feature (for example $\tilde{y} = 2.5$ or $\tilde{y} = 5.4$); otherwise the arc leads to a subordinate decision node on a different input feature. For example the arcs starting from the node labelled $x2$ lead to decisions based on the value of $x4$ and $x1$.

refer to the paper in which the library is presented or the online documentation¹.

There are four main reasons to use gradient-boosted trees, and in particular XGBoost, in this work:

1. Execution speed: Generally, XGBoost is faster than back-propagation-based models (e.g. neural networks), and even than some other gradient-boosting implementations (Pafka 2015). In addition, and of particular interest for galaxy identification, its support for hardware acceleration makes the speed difference with any SED-fitting algorithm very advantageous (see Sect. 3.5).

2. Model performance: Gradient-boosting dominates structured and tabular datasets on both regression and classification predictive modelling problems. It is the go-to algorithm for competition winners on Kaggle².

3. Decision-based splits: In the decision tree model, the splitting occurs at certain thresholds for every feature (see Fig. 5). This makes the model more robust to outliers and to the presence of only upper or lower limits in certain attributes, because it does not make any difference how far a point is from such thresholds.

4. Missing values handling: The dataset used in this work includes missing data (NaNs) in one or more of the features. Traditionally, to handle NaNs one can specify a fixed value to replace missing numbers, or impute them with either the mean or the median of that feature. It is obvious that this approach might not always be the best choice. XGBoost enhances the function class to learn the best way to handle missing values: the idea is to learn a ‘default direction’ for each node and guide the sample with missing values along the default directions. This approach can be seen as an implicit way of imputing missing numbers.

¹ <https://xgboost.readthedocs.io>

² Kaggle (<https://www.kaggle.com/>) is a community of data scientists and a place to find and publish datasets, explore and build models in a web environment, and participate in data science and machine learning competitions (with prizes).

We might also have considered other recent gradient-boosting implementations, for example CatBoost (Prokhorenkova et al. 2017) or LightGBM (Ke et al. 2017), which should perform very similarly to XGBoost. However, a systematic comparison of different gradient-boosting libraries is beyond the scope of this work. We opted for XGBoost due to its longer presence in the field, which has allowed for a more mature development of features and a wider user community.

XGBoost hyperparameters

A hyperparameter is a parameter whose value is used to control the learning process, which, as a consequence, cannot be learned, but has to be set by the user by evaluating the machine performance while varying its value.

The most impactful XGBoost hyperparameters are as follows:

- The number of estimators refers to the number of gradient-boosted trees fitted during the learning process. Larger values lead to more complex models, which, however, are more prone to overfitting³.

- The learning rate is the rate at which new trees are added to the ensemble. Lower values lead to a slower addition of new trees, thus preventing (or at least slowing down) overfitting.

- The maximum depth refers to that of a tree. Increasing this value will make the model more complex and more likely to overfit.

- γ is defined as the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger γ is, the more conservative the algorithm will be.

- λ is the L2 (squared norm) regularisation term on the weights. Increasing this value will make the model more conservative.

- The column sample by tree is the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.

The values used in this work and the optimisation method used to derive them are reported in Sect. 3.5.

3.2. Data preprocessing and feature engineering

Data preprocessing is a fundamental step in machine learning, since the quality of data greatly impacts the capability of a model to learn. Therefore, before feeding the data to the machine learning model and tuning its hyperparameters, we preprocessed it by performing feature engineering and data cleaning. To begin, the magnitude cut applied to the simulated *Euclid* lightcone was set to a depth of 2σ , and we defined magnitudes larger than their 2σ limit (see Table 1) as missing numbers.

In the following analysis, we considered only objects at $2 \leq z \leq 8$. This selection yields a dataset with 6 304 179 galaxies. We did not consider galaxies at $z < 2$ since the official photo- z pipelines are very well optimised for identifying low-redshift galaxies (see e.g. Euclid Collaboration 2023b, 2020). By utilising them, we assumed that we would have a clean selection of high-redshift objects.

To test the validity of this assumption, an analysis addressing potential contamination from low-redshift sources in the selection of HIEROs candidates is presented in Appendix D. In the same appendix we also assess the contamination by brown dwarfs.

³ Overfitting is the production of an analysis corresponding too closely to a particular dataset, and might thus fail to predict new observations reliably.

3.2.1. Features engineering

In this work, each element in the data refers to the simulated photometry of a particular galaxy. To provide more information for the objects, some derived features are also included: the more that is known about the SED of a galaxy, the better the inference will be. The number of features provided in the dataset is limited to 11 (four *Euclid*, five *Rubin*, and two IRAC bands; see Table 1); therefore, we decided to include some additional features to add more information for the training. In particular, we included the following features.

- Differences: Pairwise (without permutation) differences of the magnitudes.

- Ratios: Pairwise ratios between magnitudes without permutation. Even though they have no physical meaning, they are used because we empirically found that they help the training, increasing (albeit slightly) the performance.

- Errors: Parametric photometric errors associated with each band, as given by Eq. (2) (applied to the perturbed magnitudes). This parametrisation is also applied for the *Euclid* magnitudes even though their uncertainties are provided in the catalogue, as these are computed analytically starting from the true flux, unaffected by photometric errors.

This process generates a total of 132 features, whose importance is reported and discussed in Appendix C.

3.2.2. Data cleaning

To have a more reliable set of measurements, only objects detected (i.e. $S/N > 2$) in at least four bands are used for estimating photometric redshifts; their counts for different z ranges are reported in Table 2 and the fraction of detections per band (computed after the cleaning procedure) in Table 3. The choice of going to such a low S/N is explained in Sect. 3.4. We remind the reader that the starting catalogue contains galaxies with at least one detection in a *Euclid* filter.

This cleaning procedure removes roughly 18% of the initial data, yielding a total of 5 174 988 galaxies. To show how the redshift distribution is not strongly affected by the cleaning performed, the percentage reduction of objects in different redshift ranges is also reported in Table 2.

In the following photometric redshift estimation procedure, for bands with missing numbers (which replaced magnitudes fainter than their 2σ detection limits), 2σ magnitude lower limits are used (see Table 1). Missing colours and ratios of magnitudes are also replaced with the 2σ magnitudes lower limits of the missing band (if only one) instead of their upper or lower limit, or left missing (if both magnitudes are missing).

These choices were taken in order to avoid the contamination of detected colours and ratios with other lower or upper bounded ones; they furthermore provided slightly better performance. Some clarifying examples are shown in Table A.1.

3.3. Performance metrics

Three metrics were used to evaluate the redshift prediction performance (m indicates the number of objects in the sample, z the true redshift, and \tilde{z} the model-predicted redshift), the root mean squared error (RMSE), the bias ($\langle \Delta z \rangle$), and the normalised median absolute deviation (NMAD), and the and the catastrophic outlier fraction (OLF):

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (z_i - \tilde{z}_i)^2};$$

$$\langle \Delta z \rangle = \frac{1}{m} \sum_{i=1}^m (\tilde{z}_i - z_i);$$

$$\text{NMAD} = 1.48 \text{ median}(|z - \tilde{z}|/(1+z));$$

$$\text{OLF} = \frac{1}{m} |\{\tilde{z} : |z - \tilde{z}|/(1+z) > 0.15\}|.$$

For spectral type classification, to evaluate the model we used a simple accuracy metric:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}.$$

3.4. The training set

To train and evaluate the model, the dataset must be divided into two subsets. The first one, of size N_{train} , is used to fit the model and is referred to as the training set. The second, of size N_{test} , is not used to train the model; instead, its input elements are provided to the model to make predictions. This second subset is referred to as the test set. The objective of such a division was to estimate the performance of the machine learning model on new data. This procedure is called a train-test split and depends on the percentage size of the training and test sets with respect to the initial dataset.

To ensure the model is trained effectively, in real-world observations a training set must be built with the most robust measurements and accurate redshift estimations available. The redshifts in a training set are thus required to be estimated spectroscopically or with very reliable photometry derived from a larger number of bands than those that will be available with *Euclid* + ancillary data. Consequently, as detailed in Sect. 3.2.2, we decided to extend our analyses to objects with $S/N > 2$, as in real observations these objects will be supplemented with additional information that compensates for such a low S/N .

While in practical applications one typically has limited control over the size of the training set N_{train} , when forecasting future surveys observations it is useful to assess what dimension of the training set is required to obtain a given prediction performance. In this first part, the minimum training set size required to obtain good test performance is found by performing a baseline XGBoost regression on different numbers of training points N_{train} and then comparing the results obtained.

Figure 6 shows the RMSE, OLF, and NMAD improvement (evaluated on a test set of 1 000 000 galaxies subdivided into different redshift ranges) for different training set sizes⁴, as a function of $N_{\text{train}}/(N_{\text{train}} + N_{\text{test}})$. By analysing such curves, we came to the following conclusions:

- All three performance metrics reported increase as N_{train} increases, as expected, but they begin to flatten for large N_{train} values. In general, the size of the training set required to obtain a certain precision depends on the diversity built into the dataset: for photometric redshifts it depends on the redshift range, meaning that the narrower the redshift range, the smaller the N_{train} required;

- The performance gain is in every case larger for higher redshift objects with respect to lower redshift ones; for example, the RMSE improvement for objects at $6 < z \leq 7$ becomes more than five times larger than the one for $z \leq 3$ galaxies.

⁴ The improvement of a metric for a training set of size N_{train} , with respect to the initial training set size one, M_0 , is defined as $M_0 - M(N_{\text{train}})$, where $M(N_{\text{train}})$ is the metric evaluated after training with N_{train} data points.

Table 2. Counts per squared degree for objects detected in at least four bands.

z	Total	Spiral	SB	SF-AGN	SB-AGN	AGN1	AGN2	Elliptical	Irregular	HIEROs
$2 \leq z < 3$	73 906 (17.08%)	508 (9.47%)	4612 (18.32%)	784 (6.08%)	3498 (20.19%)	1267 (5.22%)	480 (18.22%)	598 (22.92%)	62 159 (17.13%)	152 (27.17%)
$3 \leq z < 4$	30 205 (17.38%)	111 (11.1%)	2447 (14.97%)	205 (7.05%)	2195 (18.11%)	568 (6.0%)	165 (10.94%)	10 (67.0%)	24 504 (17.87%)	1107 (7.81%)
$4 \leq z < 5$	14 027 (19.31%)	62 (7.54%)	1179 (12.41%)	125 (5.9%)	1092 (17.03%)	258 (8.89%)	57 (9.88%)	0 (100%)	11 253 (20.61%)	1979 (4.55%)
$5 \leq z < 6$	6554 (22.45%)	34 (6.36%)	595 (12.66%)	78 (6.66%)	526 (16.26%)	120 (10.72%)	20 (9.46%)	0 (0%)	5182 (24.54%)	1443 (3.57%)
$6 \leq z < 7$	3144 (21.49%)	18 (7.55%)	311 (13.79%)	49 (5.92%)	255 (10.58%)	60 (10.7%)	7 (10.48%)	0 (0%)	2444 (23.91%)	801 (3.94%)
$7 \leq z < 8$	1540 (25.85%)	11 (10.13%)	175 (16.18%)	31 (8.7%)	125 (19.66%)	30 (9.79%)	3 (16.67%)	0 (0%)	1166 (28.5%)	356 (17.88%)

Notes. Counts per squared degree for different redshift intervals and spectral types for objects detected at least in four bands. In parentheses we indicate the percentage of objects lost during data cleaning, which does not show a strong dependence on redshift. We lose 100% of elliptical galaxies at $4 \leq z < 5$, but the starting catalogue contained only eight of them in this redshift range. We also present the counts for HIEROs, which are also only marginally affected by our data cleaning process.

Table 3. Fraction of detections, i.e. $S/N > 2$, per considered band for objects detected in at least four bands.

<i>Rubin/u</i>	<i>Rubin/g</i>	<i>Rubin/r</i>	<i>Rubin/i</i>	<i>Rubin/z</i>	I_E	Y_E	J_E	H_E	IRAC/3.6 μm	IRAC/4.5 μm
16.4%	75.8%	94.6%	96.2%	93.2%	94.5%	46.3%	71.6%	87.2%	60.1%	61.6%

The number of detected galaxies at $z < 4$ is about four times the number of higher- z galaxies (see Table 2). This means that larger training sets are important to constraint the redshifts of rare objects: the majority of galaxies in the dataset (and, consequently, in the training set) are at low redshifts, so the training process gives more weight to the predictions for galaxies at low redshifts with respect to the other galaxies. Furthermore, for machine learning models, high- z objects are more difficult to identify than lower- z ones, given their larger photometric uncertainties.

To conclude, all the curves reported in Fig. 6 show an onset of a plateau at an N_{train} corresponding to roughly 10% of the data size. This 10%–90% train–test split will then be used in the following analyses.

3.5. Hyperparameter tuning and time performance

A Bayesian hyperparameters optimisation with cross-validation is then run on a training set large 10% of the total data size. The hyperparameters delivering the lowest RMSE thus found are reported in Table 4.

Timed on a workstation with a 2.20 GHz Intel Xeon CPU and a 16 GB Tesla P100-PCIE GPU, an XGBoost regressor with these hyperparameters takes 70 s to train on a dataset with 517 498 samples and 132 features, and 40 s to estimate the redshift for the 4.6 million galaxies in the test set. To demonstrate the difference in speed with SED-fitting methods, performing this operation with LEPHARE (Arnouts & Ilbert 2011) on similar hardware requires approximately 0.23 s per object (considering 463 680 different combinations of SED template, redshift, age and dust extinction). Estimating the photometric redshift for all the galaxies in the test set would then need approximately 12 days.

4. Results

During the testing phase, we set a higher significance threshold by considering only observations with at least one band having $S/N > 5$. This additional criterion, in conjunction with the requirement of at least four detections at $S/N > 2$ (see Sect. 3.2.2), aimed to minimise inclusions of spurious objects such as noise spikes, halos, and artefacts that would result from real observations. This removes roughly 21.728% of the initial test data N_{test} , yielding a total of 3 645 481 galaxies. The results reported hereafter are thus relative to this dataset.

4.1. Photometric redshifts

The photo- z performance for the test set is summarised in Table B.1 and shown in Fig. 7.

This figure clearly illustrates the trend of the performance metrics:

1. Top-left panel: The errors increase with redshift, as expected and noted in Sect. 3.4.

2. Bottom panels: The precision of the predicted redshift decreases for fainter objects, as these are usually found at higher z and they come with detections in fewer bands; this practically means the input features vectors contain less information to learn and infer from.

3. Top-right panel: The lowest errors (RMSE = 0.22) are obtained for elliptical galaxies, and the largest for AGN1, SF-AGN, and SBs (RMSE > 0.5). This behaviour may be related to their variability⁵ and (for SBs) their number density evolution with redshift: while ellipticals quickly drop from $z = 2$ to disappear at $z \gtrsim 3.5$, SBs are observed at any redshift.

Furthermore, by removing the IRAC bands from the input features, we note (Fig. 7, top-left panel) how mid-IR data improves redshift estimation especially for $z \gtrsim 4$ galaxies.

⁵ Data variability refers to how spread out a set of data is.

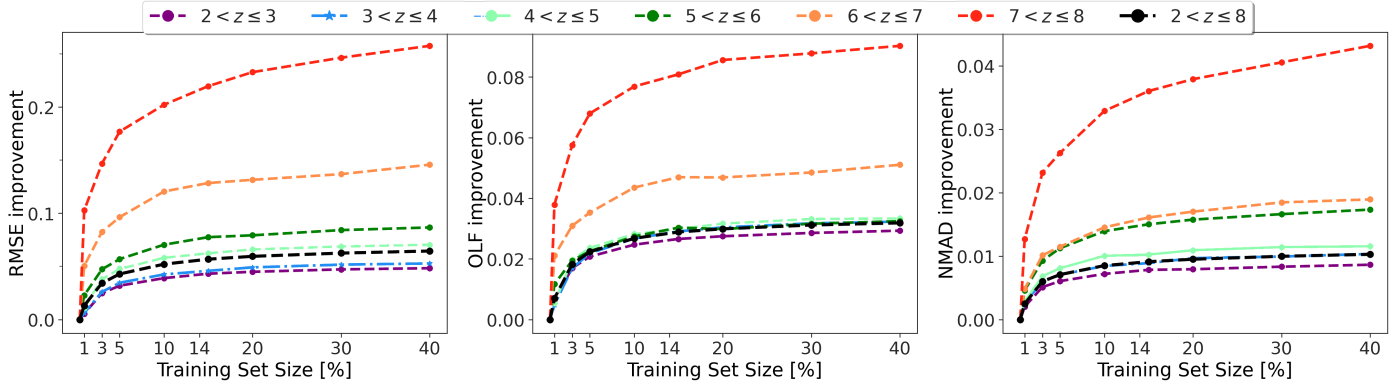


Fig. 6. Photometric redshift prediction performance improvement (as defined in Footnote 4) with respect to the metrics with the initial training set size (0.5% of the total number of galaxies), as a function of the training set size, evaluated on a sample of 1 000 000 test galaxies. From left to right we plot: the RMSE, OLF, and NMAD for different redshift intervals, as indicated in the legend.

Table 4. Best XGBoost hyperparameters.

Number of estimators	530
Learning rate	0.01
Max depth	12
γ	0
λ	7.7×10^{-5}
Column sample by tree	0.6

Comparison with previous results

To give context to the results obtained, we compared them to previous photometric redshift performance, in a similar redshift range, as reported in Weaver et al. (2022, hereafter W22). There, the precision of the photometric redshifts obtained with LePHARE using 39 bands, and included in the COSMOS2020 catalogue, is assessed against spectroscopic ones over the Cosmic Evolution Survey (COSMOS, Scoville et al. 2007) field ($0 < z < 6$).

The photometric catalogues created in COSMOS, with their rich multi-wavelength coverage, have for years constituted the state of the art reference to predict the quality of photo-zs. Within the *Euclid* Collaboration (Euclid Collaboration 2020), this assessment has been done using the COSMOS2015 catalogue (Laigle et al. 2016). Improving on previous releases, the COSMOS2020 catalogue features significantly deeper optical, IR, and near-IR data and thus gains almost one order of magnitude in photometric redshift precision compared to its predecessor COSMOS2015.

Comparing the results reported in W22 and those obtained in this work using a gradient-boosting approach on the simulated *Euclid* Deep Survey catalogue, we find that even in the faintest $25 < i < 27$ bin, gradient-boosting provides a lower NMAD (0.019 versus 0.044), and also a lower OLF (0.005 versus 0.204). Overall, the performance of the XGBoost method presented in this work is comparable to the performance of previous SED-fitting at each magnitude range, but our process is based on a notably smaller number of filters as inputs (i.e. 11 instead of 39).

It is important to remind the reader that our results are derived from simulated data, which is a simplified representation of reality, and therefore may not reflect actual performance accurately and could potentially overestimate it. The aim of this comparison is therefore to show that our performance is in line with

previous studies without stressing much the one-to-one direct comparison.

To conclude this photo- z section, we estimate that in the range $z > 6$ the fraction of contaminants is relatively low, at 12%, while the completeness is around 72%. For comparison, Euclid Collaboration (2022c), using the same set of photometric bands (*Euclid*+ *Rubin* + *Spitzer*) and estimating with LePHARE the photometric redshifts of mock galaxies created from the Ultra Deep Survey with the VISTA telescope (UltraVISTA, McCracken et al. 2012), obtained at $z > 6$ a contamination fraction of 12% for bright UltraVISTA-like galaxies, with a $z > 6$ completeness of 95%. For fainter sources ($25.3 \leq H_E < 27.0$), contamination is more prevalent at 35%, with a $z > 6$ completeness of 88%. Our smaller completeness and contaminant fractions are mostly due to the unbalanced redshift distribution in the dataset (and consequently in the training set), as pointed out in Sect. 3.4. In fact, objects at $z > 6$ are systematically placed at lower redshifts, as their average offset $\bar{z} - z$ is ≈ -0.483 , while for objects at $z < 6$ it is ≈ 0.001 . However, this paper focuses mainly at $3 < z < 7$, where the XGBoost performance is better, with a completeness of 89.3% and a contaminant fraction of 7.73%.

4.2. SED type classification

We also performed an independent experiment, applying the same methodology with the aim of recovering the classification of the spectral class of each source. The prior knowledge on the redshift derived as in the previous sections is ignored here.

When used as a classifier, when a feature vector is fed to it, the XGBoost output is a vector of probabilities, in this case with eight entries, each one corresponding to an SED type. The predicted class is thus the one corresponding to the maximum value. Each simulated galaxy is assigned an SED template taken from a set of 35 templates, divided into eight spectral types, as reported in Sect. 2.1.

Since the XGBoost input data and the train-test split used are the same as in the photometric redshift estimation, we used the same hyperparameters (Table 4). The accuracy obtained on the test set is 96.8%, meaning that 3 529 289 galaxies out of 3 645 481 are correctly classified. Some significant misclassifications (see the confusion matrix in Fig. 8) are explained by considering the nature of the SEDs and the photometry available. For example, it is reasonable that AGN2 and SB-AGN, both obscured in the optical, are misclassified as SBs, since it would be necessary to

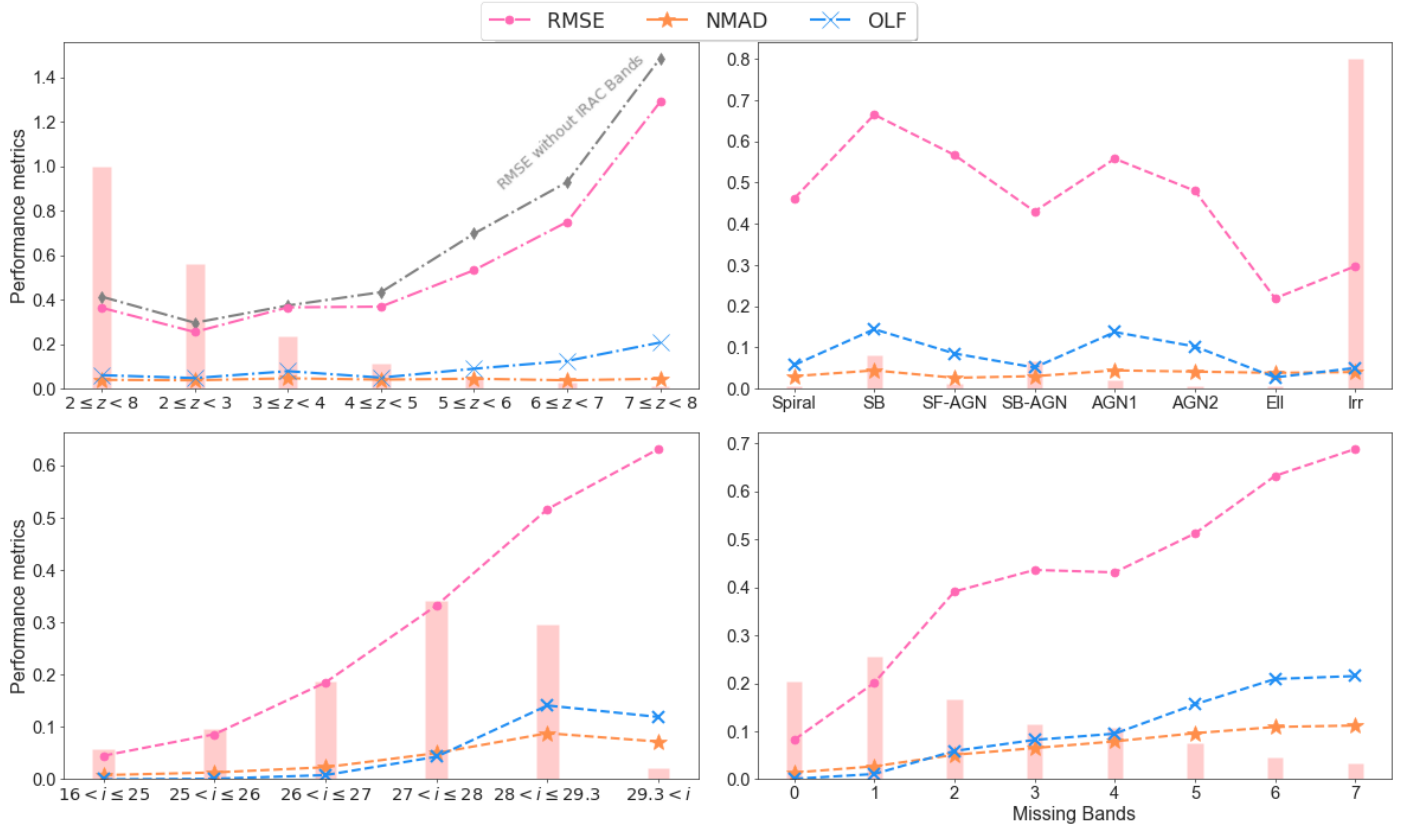


Fig. 7. XGBoost test set RMSE, OLF, and NMAD (starting from the upper-left panel and going clockwise) for different redshift intervals, SED types, numbers of missing bands, and i -band magnitudes. The grey curve shows the RMSE obtained by removing the two IRAC bands from the XGBoost input. The vertical pink bands indicate the fraction of objects belonging to the group with respect to the test set size. For details, refer to Table B.1.

have spectroscopic data or photometric observations in the mid-IR (rest-frame) and X-ray, to identify their AGN nature.

The distribution of the maximum probability per object for correct and wrong classifications is shown in Fig. 9. While for the mistakenly classified samples it is a rather flat distribution, for correct predictions it is strongly peaked at values of maximum probability larger than ≈ 0.95 . This means it is possible to improve the spectral type prediction accuracy even more, by simply considering objects with large maximum probability, while losing a negligible number of correct predictions. For example, by keeping only objects with maximum probability > 0.8 , the number of misclassified objects goes down by 61%, while dropping only 3.7% of the correctly classified ones. The resulting accuracy is 98.7%.

4.3. Identifying HIEROs

Having determined the general framework for the estimations of photometric redshifts and spectral types, in this section we present the results of the approach applied to the most relevant population in our study, namely red and massive galaxies at distances greater than $z \approx 3$. We selected them as HIEROs (Eq. (4) and Fig. 2).

As these massive galaxies are challenging to spectroscopically confirm, to obtain a more realistic estimate of the identification performance of this population, we trained the gradient-boosting algorithm with a more realistic spectroscopic completeness: we utilised the training set as described in the previous sections but retaining only 500 HIEROs, which accounts for

approximately 1/46 of the total number of HIEROs that would be available. We consider this to be conservative approach since:

- We assume that a robust sample of HIERO that we could use as training set will be available in the COSMOS field, thanks to the very precise and accurate photometric redshifts derived from the already available and future multi-wavelength observations (W22; Casey et al. 2023).
- We assume that in the next few years we will have access to spectroscopic redshifts for HIEROs thanks to JWST, as there are already photometric candidates suitable for follow-up (i.e. 138 objects similar to HIEROs in an area of 38.8 arcmin^2 Pérez-González et al. 2023).

As there is clearly some stochasticity in the sampling of the HIEROs in the training set, the results reported are the average across multiple runs.

In Table 5 their identification performance is shown in terms of photo- z and SED type classification. The regression performance is compared to that reported in Wang et al. (2016), where the NMAD is approximately 0.1. Even for this population, the performance obtained in this work (HIEROs NMAD = 0.068) is encouraging, especially for the brightest galaxies ($H_E \leq 26$), as shown in Figs. 10 and 11. Clearly, the advantage of the *Euclid* Deep Fields survey will be the HIEROs' much larger statistical relevance compared to previous surveys.

We also report in Table 6 the completeness and fraction of contaminants in different redshift ranges. The redshift range where we expect to find the vast majority of these objects (i.e. $3 \leq z < 7$) is complete at 99.4%, with a contaminant fraction of only 5%, mostly coming from higher-redshift galaxies.

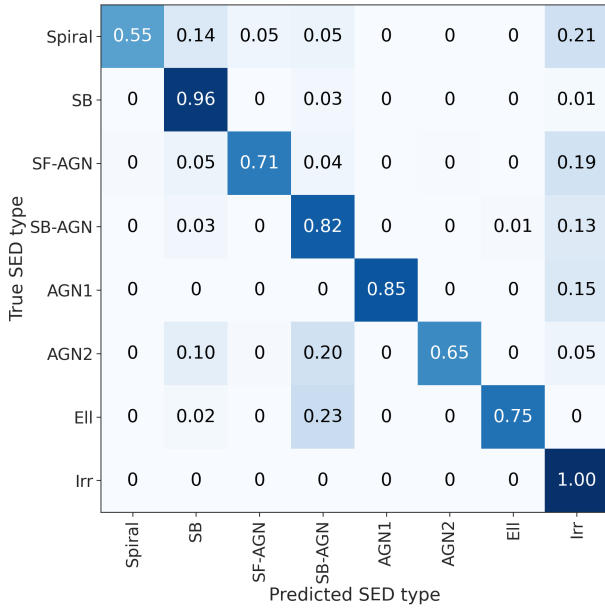


Fig. 8. Test set confusion matrix for SED type classification. Each row of the matrix represents the fraction of samples in an actual SED type, while each column represents the fraction of samples in a predicted SED type. The superimposed numbers indicate the fraction of objects of a class in that particular position. For example, considering the second row (SBs): 96% of SBs are correctly classified as such, 3% as SB-AGN, and 1% as Irregular.

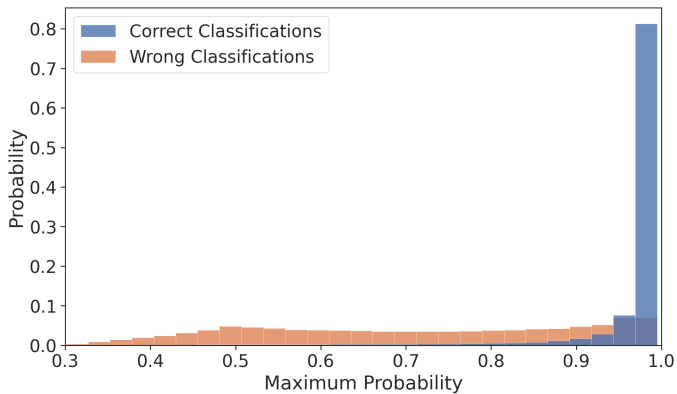


Fig. 9. Maximum probability distribution for correct (blue) and wrong (orange) classifications. The sum of the heights of each histogram is one.

Lastly, the spectral type classification, although less accurate than for non-HIEROs (86.20% versus 97.20%), shows (Fig. 12) a similar confusion between classes to the classifications for all the objects in the catalogue (Fig. 8).

5. Summary and discussion

5.1. A new approach to identifying the most elusive and massive systems with *Euclid* at high z

In this work we have taken a photometric approach to predicting whether *Euclid* will provide enough information to resolve one of the key scientific unknowns in galaxy formation and evolution studies: the role of distant obscured galaxies in the buildup of today's large-scale structures.

Such sources have proven to be elusive in optical surveys, and current near-IR surveys are deep enough to detect them only in very small sky areas. Since these optically faint objects are rare, large-area surveys that go to sufficient depths are necessary to provide a statistical census of such a population. *Euclid* will potentially provide all the ingredients to recover these missing galaxies at high redshifts (e.g. $z > 3$). However, it has been unclear if the photometric information it will deliver will be sufficient to identify and characterise them, given the lack of bright spectroscopic features observable by *Euclid* at these distances (i.e. $\lambda_{\text{rest-frame}} < 0.5 \mu\text{m}$).

Our goal was therefore to assess the capability of *Euclid* in combination with ancillary data to identify these distant obscured objects. We carried out the following analysis:

- We adopted the simulated *Euclid* Deep Fields catalogue from the SPRITZ simulation as a basis for our analysis and show that it includes a massive and dusty population, as selected by the criterion $H_E - [4.5] > 2.25$. A total of 98% of these simulated objects (the HIEROs) are indeed at $3 < z < 7$ and contribute significantly ($\approx 93\%$) to the massive end of the stellar mass function at $z > 4$.

- We implemented a general, fast, and accurate machine learning technique optimised for $z \geq 2$ galaxy identification, based on photometric data from *Euclid*, *Rubin*, and *Spitzer*, finding that only a $\approx 10\%$ subset of the total observed galaxies with spectroscopic redshifts (or reliable photometric redshifts) is required to obtain a good performance (RMSE = 0.363, OLF = 0.061, $\langle \Delta z \rangle = -0.017$, and NMAD = 0.039 over $2 \leq z < 8$). This photo- z precision is comparable to that of the COSMOS2020 catalogue, obtained with state-of-the-art traditional SED-fitting methods with almost four times the number of photometric bands.

- We applied the same methods for spectral type classification, distinguishing between eight different spectral types (i.e. spiral, SB, star-forming AGN, SB-AGN, type-1 AGN, type-2 AGN, elliptical, and irregular), obtaining an accuracy of 96.8%.

- We evaluated the identification performance for objects within the HIEROs selection and found a photo- z OLF = 0.123 and NMAD = 0.068; the SED classification accuracy is 86.2%. Evaluating only the brightest HIEROs ($H_E < 26$), we obtained an OLF = 0.085, an NMAD = 0.055, and a 92.68% SED classification accuracy.

- We determined the completeness and the fraction of contaminants for the HIERO photo z . In the range $3 \leq z < 7$, we estimated a completeness of 99.4% and a contaminant fraction of 5%. The majority of the HIEROs are at $z > 7$.

All these results suggest that by leveraging *Euclid*, *Rubin*, and *Spitzer* photometric data, and by taking the approach described in this paper, it will be possible to constrain the contribution of the dusty and massive galaxies at $z \approx 3 - 7$ to the mass functions and, hopefully, to the SFRD.

5.2. Future perspectives

Under suitable conditions, gradient-boosting (like many other supervised machine learning techniques) is a very competitive tool for photometric redshift and spectral type estimation. However, its successful application depends on the availability of a large enough training set that is representative of the populations under consideration. It is therefore most effective when applied to large photometric surveys, some of which include spectroscopic data for subsets of the photometric catalogues. One considerable problem for these methods is the difficulty in extrapolating to regions of the feature space not properly represented

Table 5. Redshift and spectral type prediction performance for HIEROs.

	N	Photo- z			SED classification	
		RMSE	$\langle \Delta z \rangle$	OLF	NMAD	Accuracy
HIEROs	184 441	0.654	-0.084	0.123	0.068	86.20%
Red HIEROs	51 092	0.668	-0.059	0.143	0.073	86.28%
Blue HIEROs	89 996	0.690	-0.119	0.126	0.061	86.28%
Other	3 461 040	0.341	-0.013	0.057	0.038	97.20%
$H_E \leq 26$ HIEROs	30 632	0.545	-0.131	0.085	0.055	92.68%
$H_E \leq 26$ Red HIEROs	24 645	0.477	-0.097	0.07	0.052	94.44%
$H_E \leq 26$ Blue HIEROs	5987	0.766	-0.275	0.149	0.076	85.44%

Notes. N indicates the number of objects in the selection. ‘Other’ indicates the classes of objects with $H_E - [4.5] < 2.25$.

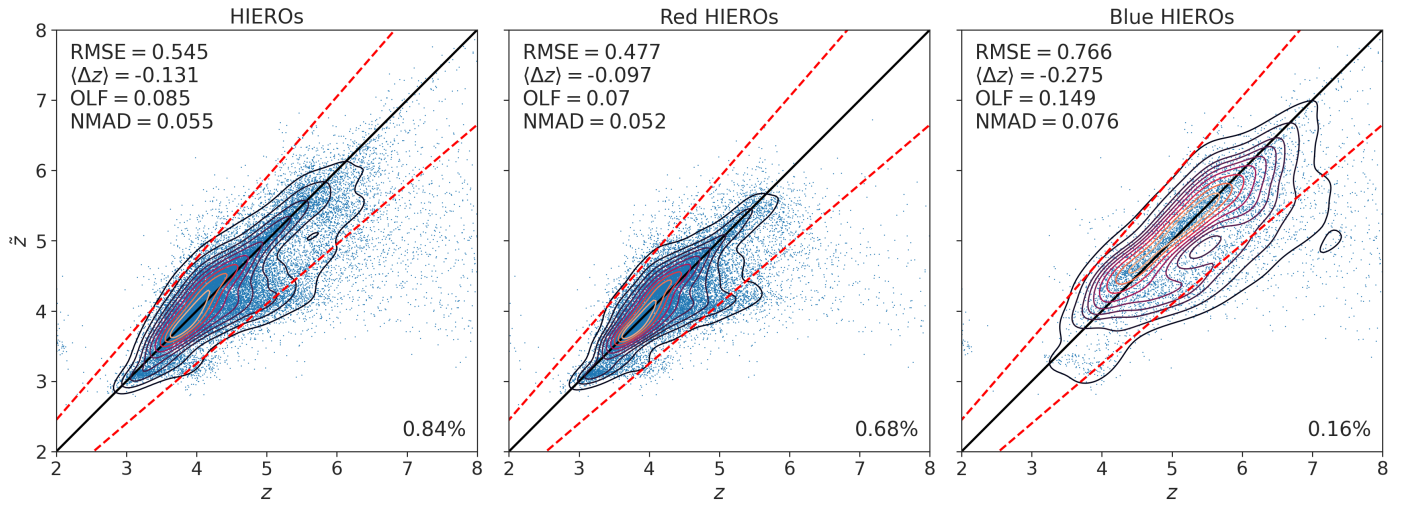


Fig. 10. Contour plot of XGBoost predictions (\tilde{z}) versus catalogue redshifts (z) coloured by density in the (z, \tilde{z}) space for HIEROs (left panel), red HIEROs (middle panel), and blue HIEROs (right panel), all of which have $H_E \leq 26$. Lighter contour colours indicate higher densities. The dashed red lines show $\tilde{z} = z \pm 0.15(1+z)$. The percentage of objects with respect to N_{test} is indicated in the bottom-right corner of each panel.

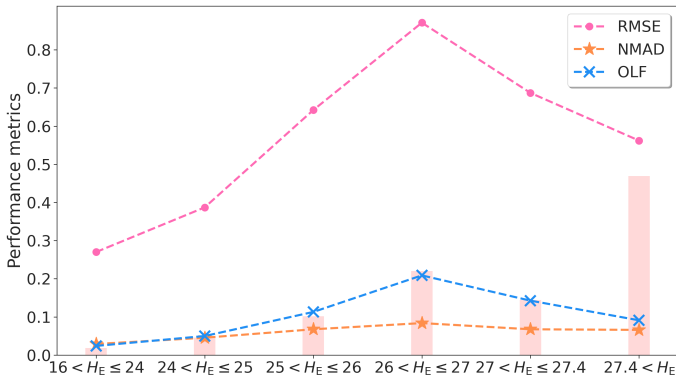


Fig. 11. RMSE, OLF, and NMAD values for HIEROs of different H_E -band magnitudes. The vertical pink bands indicate the fraction of objects belonging to the group with respect to the total number of HIEROs.

in the training data; the distribution of magnitudes and colours of the training set has to be as close as possible to those in the target set.

Lastly, the good performance demonstrated here relies heavily on the representativeness of the SPRITZ simulation with respect to the *Euclid* observations. It has been shown (Bisigello et al. 2021, 2022) that SPRITZ results are in agreement

Table 6. Confusion matrix for photometric redshift ranges for HIEROs.

	$z < 3$	$3 \leq z < 7$	$7 \leq z$
$\tilde{z} < 3$	37.2% (1537)	22.1% (438)	0% (9)
$3 \leq \tilde{z} < 7$	1.4% (2590)	99.4% (169 707)	3.6% (6442)
$7 \leq \tilde{z}$	0% (2)	15.2% (566)	32.8% (3150)

Notes. z and \tilde{z} indicate true and predicted redshifts, respectively. The diagonal entries represent the completeness of the samples, while the off-diagonal terms represent their contaminants. The sum along rows of the off-diagonal terms equals 1-purity. In parenthesis we report the numbers of objects

with a large set of observations at different wavelengths. However, the simulation also has some limitations. For example, the number of SED templates included in the simulation is limited, and results at number densities $z > 3$ are, though widely tested, mainly based on extrapolation. Therefore, the results shown here may be prone to biases or may be slightly optimistic compared to those that will be obtained once real data become available. To mitigate this limitation and enable a more robust forecast, future work could apply the methods of this paper to different simulated datasets, for example MAMBo (Mocks with Abundance Matching in Bologna; Girelli et al. 2020), and compare the results obtained with those of the official *Euclid* pipelines.

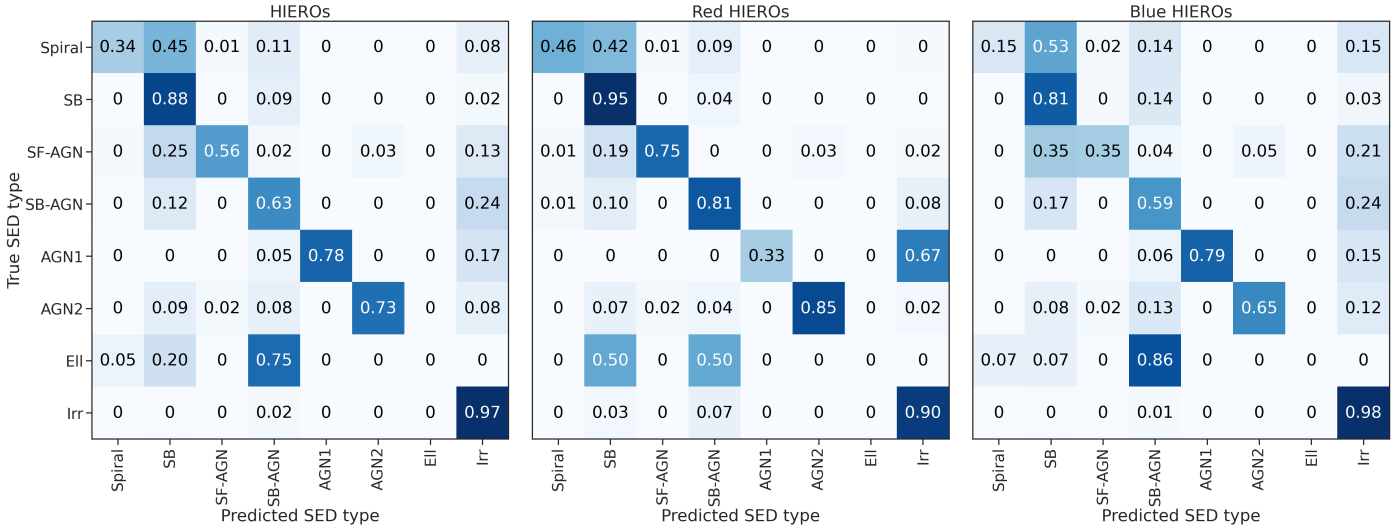


Fig. 12. SED type classification confusion matrix for HIEROs, red HIEROs, and blue HIEROs. The high confusion for ellipticals is explained by noting that there are only 20 of them in the HIERO selection (six in the red HIEROs and 14 in the blue HIEROs), none of them correctly identified.

Furthermore, some additional strategies can be used to improve the predictions. The gradient-boosting algorithm can easily be extended to provide a probability distribution even on redshifts.

After applying the most effective photometric pipeline to real data, a promising selection of candidate galaxies will be obtained. To confirm their nature, spectroscopic follow-up will be necessary, for example via the Extremely Large Telescope, ALMA, or JWST.

Acknowledgements. We thank Tao Wang for sharing his data for the stellar mass function for HIEROs. 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 French Centre National d’Etudes Spatiales, the Deutsches Zentrum für Luft- und Raumfahrt, the Danish Space Research Institute, the Fundação para a Ciência e a Tecnologia, the Ministerio de Ciencia e Innovación, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Nederlandse 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>).

References

Abdalla, F. B., Banerji, M., Lahav, O., & Rashkov, V. 2011, *MNRAS*, **417**, 1891
 Algera, H. S. B., Inami, H., Oesch, P. A., et al. 2023, *MNRAS*, **518**, 6142
 Arnouts, S., & Ilbert, O. 2011, *Astrophysics Source Code Library* [record ascl:[1108.0091](https://arxiv.org/abs/1108.0091)]
 Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, *ApJ*, **683**, 12
 Barro, G., Pérez-González, P. G., Kocevski, D. D., et al. 2024, *ApJ*, **963**, 128
 Barrufet, L., Oesch, P. A., Bouwens, R., et al. 2023a, *MNRAS*, **522**, 3926
 Barrufet, L., Oesch, P. A., Weibel, A., et al. 2023b, *MNRAS*, **522**, 449
 Benitez, N. 2000, *ApJ*, **536**, 571
 Bian, F., Fan, X., Jiang, L., et al. 2013, *ApJ*, **774**, 28
 Bianchi, S., De Vis, P., Viaene, S., et al. 2018, *A&A*, **620**, A112
 Bisigello, L., Kuchner, U., Conselice, C. J., et al. 2020, *MNRAS*, **494**, 2337
 Bisigello, L., Gruppioni, C., Calura, F., et al. 2021, *PASA*, **38**, e064
 Bisigello, L., Vallini, L., Gruppioni, C., et al. 2022, *A&A*, **666**, A193
 Bisigello, L., Gandolfi, G., Grazian, A., et al. 2023, *A&A*, **676**, A76
 Boogaard, L. A., Decarli, R., Walter, F., et al. 2023, *ApJ*, **945**, 111
 Brandt, W. N., Ni, Q., Yang, G., et al. 2018, *ArXiv e-prints* [[arXiv:1811.06542](https://arxiv.org/abs/1811.06542)]
 Burrows, A., Sudarsky, D., & Hubeny, I. 2006, *ApJ*, **640**, 1063

Capak, P., Arendt, R., Arnouts, S., et al. 2016, *Spitzer Proposal ID 13058*, <https://ui.adsabs.harvard.edu/abs/2016sptz.prop13058C>
 Caputi, K. I., Caminha, G. B., Fujimoto, S., et al. 2021, *ApJ*, **908**, 146
 Casey, C. M., Kartaltepe, J. S., Drakos, N. E., et al. 2023, *ApJ*, **954**, 31
 Chen, T., & Guestrin, C. 2016, *ArXiv e-prints* [[arXiv:1603.02754](https://arxiv.org/abs/1603.02754)]
 Cropper, M., Pottinger, S., Niemi, S., et al. 2016, *SPIE Conf. Ser.*, **9904**, 99040Q
 Daddi, E., Dannerbauer, H., Stern, D., et al. 2009, *ApJ*, **694**, 1517
 Enia, A., Talia, M., Pozzi, F., et al. 2022, *ApJ*, **927**, 204
 Euclid Collaboration (Desprez, G., et al.) 2020, *A&A*, **644**, A31
 Euclid Collaboration (Scaramella, R., et al.) 2022a, *A&A*, **662**, A112
 Euclid Collaboration (Moneti, A., et al.) 2022b, *A&A*, **658**, A126
 Euclid Collaboration (van Mierlo, S. E., et al.) 2022c, *A&A*, **666**, A200
 Euclid Collaboration (Schirmer, M., et al.) 2022d, *A&A*, **662**, A92
 Euclid Collaboration (Humphrey, A., et al.) 2022a, *A&A*, **671**, A99
 Euclid Collaboration (Bisigello, L., et al.) 2023b, *MNRAS*, **520**, 3529
 Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, *ApJS*, **154**, 10
 Foley, R. J., Koekemoer, A. M., Spergel, D. N., et al. 2018, *ArXiv e-prints* [[arXiv:1812.00514](https://arxiv.org/abs/1812.00514)]
 Frayer, D. T., Smail, I., Ivison, R. J., & Scoville, N. Z. 2000, *AJ*, **120**, 1668
 Fujimoto, S., Kohno, K., Ouchi, M., et al. 2023, *ApJS*, submitted [[arXiv:2303.01658](https://arxiv.org/abs/2303.01658)]
 Girelli, G., Pozzetti, L., Bolzonella, M., et al. 2020, *A&A*, **634**, A135
 Gruppioni, C., Pozzi, F., Andreani, P., et al. 2010, *A&A*, **518**, L27
 Gruppioni, C., Pozzi, F., Rodighiero, G., et al. 2013, *MNRAS*, **432**, 23
 Gruppioni, C., Béthermin, M., Loiacono, F., et al. 2020, *A&A*, **643**, A8
 Huang, J. S., Rigopoulou, D., Magdis, G., et al. 2014, *ApJ*, **784**, 52
 Ibata, R. A., McCannachie, A., Cuillandre, J.-C., et al. 2017, *ApJ*, **848**, 128
 Ilbert, O., Capak, P., Salvato, M., et al. 2009, *ApJ*, **690**, 1236
 Ivezić, Z., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, **873**, 111
 Ke, G., Meng, Q., Finely, T., et al. 2017, *Advances in Neural Information Processing Systems 30 (NIP 2017)*
 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](https://arxiv.org/abs/1110.3193)]
 Liu, R. H., Hill, R., Scott, D., et al. 2019, *MNRAS*, **489**, 1770
 Loiacono, F., Decarli, R., Gruppioni, C., et al. 2021, *A&A*, **646**, A76
 Lundberg, S. M., Erion, G., Chen, H., et al. 2020, *Nat. Mach. Intell.*, **2**, 2522
 Madau, P., & Dickinson, M. 2014, *ARA&A*, **52**, 415
 Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2019, *ApJ*, **877**, 81
 McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, *A&A*, **544**, A156
 Novak, M., Smolčić, V., Delhaize, J., et al. 2017, *A&A*, **602**, A5
 Oke, J. B., & Gunn, J. E. 1983, *ApJ*, **266**, 713
 Paflka, S. 2015, *Benchmarking Random Forest Implementations*, <https://www.r-bloggers.com/2015/05/benchmarking-random-forest-implementations/>
 Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, *A&A*, **621**, A26
 Pérez-González, P. G., Barro, G., Annunziatella, M., et al. 2023, *ApJ*, **946**, L16
 Polletta, M., Tajer, M., Maraschi, L., et al. 2007, *ApJ*, **663**, 81

- Prokhorenkova, L., Gusev, G., Vorobev, A., Veronika Drogush, A., & Gulín, A. 2017, ArXiv e-prints [arXiv:1706.09516]
- Riechers, D. A., Pavesi, R., Sharon, C. E., et al. 2019, *ApJ*, **872**, 7
- Rieke, G. H., Alonso-Herrero, A., Weiner, B. J., et al. 2009, *ApJ*, **692**, 556
- Rodighiero, G., Bisigello, L., Iani, E., et al. 2023, *MNRAS*, **518**, L19
- Scoville, N., Aussel, H., Brusa, M., et al. 2007, *ApJS*, **172**, 1
- Smail, I., Dudzevičiūtė, U., Stach, S. M., et al. 2021, *MNRAS*, **502**, 3426
- Sun, F., Egami, E., Pérez-González, P. G., et al. 2021, *ApJ*, **922**, 114
- Talia, M., Cimatti, A., Giuliotti, M., et al. 2021, *ApJ*, **909**, 23
- Topping, M. W., Stark, D. P., Endsley, R., et al. 2022, *MNRAS*, **516**, 975
- van Dokkum, P. G., Quadri, R., Marchesini, D., et al. 2006, *ApJ*, **638**, L59
- Wang, T., Elbaz, D., Schreiber, C., et al. 2016, *ApJ*, **816**, 84
- Wang, T., Schreiber, C., Elbaz, D., et al. 2019, *Nature*, **572**, 211
- Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, *ApJS*, **258**, 11
- Williams, C. C., Labbe, I., Spilker, J., et al. 2019, *ApJ*, **884**, 154
- Xiao, M. Y., Elbaz, D., Gómez-Guijarro, C., et al. 2023, *A&A*, **672**, A18
- ¹ Instituto de Estudios Astrofísicos, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile
- ² Inria Chile Research Center, Av. Apoquindo 2827, piso 12, Las Condes, Santiago, Chile
- ³ Dipartimento di Fisica e Astronomia “G. Galilei”, Università di Padova, Via Marzolo 8, 35131 Padova, Italy
- ⁴ INAF-Osservatorio Astronomico di Padova, Via dell’Osservatorio 5, 35122 Padova, Italy
- ⁵ INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
- ⁶ Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
- ⁷ Cosmic Dawn Center (DAWN), Copenhagen, Denmark
- ⁸ AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, 91191 Gif-sur-Yvette, France
- ⁹ INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy
- ¹⁰ Dipartimento di Fisica e Astronomia, Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
- ¹¹ INAF-IASF Bologna, Via Piero Gobetti 101, 40129 Bologna, Italy
- ¹² Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
- ¹³ Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
- ¹⁴ DTx – Digital Transformation CoLAB, Building 1, Azurém Campus, University of Minho, 4800-058 Guimarães, Portugal
- ¹⁵ Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, 00146 Rome, Italy
- ¹⁶ INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
- ¹⁷ School of Physical Sciences, The Open University, Milton Keynes MK7 6AA, UK
- ¹⁸ INAF-Istituto di Astrofisica e Planetologia Spaziali, Via del Fosso del Cavaliere, 100, 00100 Roma, Italy
- ¹⁹ INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
- ²⁰ INFN-Sezione di Bologna, Viale Bertini Pichat 6/2, 40127 Bologna, Italy
- ²¹ Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
- ²² Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany
- ²³ INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy
- ²⁴ Dipartimento di Fisica, Università di Genova, Via Dodecaneso 33, 16146 Genova, Italy
- ²⁵ INFN-Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy
- ²⁶ Department of Physics “E. Pancini”, University Federico II, Via Cinthia 6, 80126 Napoli, Italy
- ²⁷ INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
- ²⁸ INFN section of Naples, Via Cinthia 6, 80126 Napoli, Italy
- ²⁹ Dipartimento di Fisica, Università degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
- ³⁰ INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
- ³¹ Institut de Física d’Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
- ³² Port d’Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
- ³³ Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, 52056 Aachen, Germany
- ³⁴ INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
- ³⁵ Dipartimento di Fisica e Astronomia “Augusto Righi” – Alma Mater Studiorum Università di Bologna, Viale Bertini Pichat 6/2, 40127 Bologna, Italy
- ³⁶ Centre National d’Etudes Spatiales – Centre spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
- ³⁷ Institut national de physique nucléaire et de physique des particules, 3 rue Michel-Ange, 75794 Paris Cédex 16, France
- ³⁸ Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
- ³⁹ Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
- ⁴⁰ European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
- ⁴¹ ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
- ⁴² University of Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
- ⁴³ Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
- ⁴⁴ UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, 4 rue Enrico Fermi, 69622 Villeurbanne, France
- ⁴⁵ Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal
- ⁴⁶ Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
- ⁴⁷ Department of Astronomy, University of Geneva, ch. d’Ecogia 16, 1290 Versoix, Switzerland
- ⁴⁸ INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
- ⁴⁹ Univ Claude Bernard Lyon 1, CNRS, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
- ⁵⁰ Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191 Gif-sur-Yvette, France
- ⁵¹ School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
- ⁵² Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille, France
- ⁵³ Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy
- ⁵⁴ University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 Munich, Germany
- ⁵⁵ Dipartimento di Fisica “Aldo Pontremoli”, Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
- ⁵⁶ INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
- ⁵⁷ Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029, Blindern 0315, Oslo, Norway
- ⁵⁸ Department of Physics, Lancaster University, Lancaster LA1 4YB, UK
- ⁵⁹ von Hoerner & Sulger GmbH, Schloßplatz 8, 68723 Schwetzingen, Germany
- ⁶⁰ Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark
- ⁶¹ Cosmic Dawn Center (DAWN), Copenhagen, Denmark
- ⁶² Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

- ⁶³ Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
- ⁶⁴ Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
- ⁶⁵ Department of Physics, PO Box 64, 00014 University of Helsinki, Finland
- ⁶⁶ Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
- ⁶⁷ NOVA Optical Infrared Instrumentation Group at ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
- ⁶⁸ Universität Bonn, Argelander-Institut für Astronomie, Auf dem Hügel 71, 53121 Bonn, Germany
- ⁶⁹ Aix-Marseille Université, CNRS, CNES, LAM, Marseille, France
- ⁷⁰ Dipartimento di Fisica e Astronomia “Augusto Righi” – Alma Mater Studiorum Università di Bologna, Via Piero Gobetti 93/2, 40129 Bologna, Italy
- ⁷¹ Department of Physics, Institute for Computational Cosmology, Durham University, South Road, DH1 3LE, UK
- ⁷² University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland
- ⁷³ Institut d’Astrophysique de Paris, 98bis Boulevard Arago, 75014 Paris, France
- ⁷⁴ Institut d’Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France
- ⁷⁵ School of Mathematics and Physics, University of Surrey, Guildford, Surrey GU2 7XH, UK
- ⁷⁶ European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
- ⁷⁷ Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, 8000 Aarhus C, Denmark
- ⁷⁸ Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France
- ⁷⁹ Space Science Data Center, Italian Space Agency, Via del Politecnico snc, 00133 Roma, Italy
- ⁸⁰ Institute of Space Science, Str. Atomistilor, nr. 409 Măgurele, Ilfov 077125, Romania
- ⁸¹ Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
- ⁸² Universität Innsbruck, Institut für Astro- und Teilchenphysik, Technikerstr. 25/8, 6020 Innsbruck, Austria
- ⁸³ Institut d’Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain
- ⁸⁴ Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
- ⁸⁵ Satlantis, University Science Park, Sede Bld 48940, Leioa-Bilbao, Spain
- ⁸⁶ Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
- ⁸⁷ Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
- ⁸⁸ Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda 1349-018, Lisboa, Portugal
- ⁸⁹ Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, Plaza del Hospital 1, 30202 Cartagena, Spain
- ⁹⁰ Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
- ⁹¹ Centre for Information Technology, University of Groningen, PO Box 11044, 9700 CA Groningen, The Netherlands
- ⁹² INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
- ⁹³ INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
- ⁹⁴ Junia, EPA department, 41 Bd Vauban, 59800 Lille, France

Appendix A: Missing detection handling

We show some clarifying examples of the handling of missing detections used in this work in Table A.1.

Table A.1. Handling of missing detections.

M_1	M_2	$M_1 - M_2$	M_1/M_2
NaN	27.17	NaN	NaN
25.72	NaN	NaN	NaN
NaN	NaN	NaN	NaN
↓			
M_1	M_2	$M_1 - M_2$	M_1/M_2
$M_1^{(2\sigma)}$	27.17	$M_1^{(2\sigma)}$	$M_1^{(2\sigma)}$
25.72	$M_2^{(2\sigma)}$	$-M_2^{(2\sigma)}$	$1/M_2^{(2\sigma)}$
$M_1^{(2\sigma)}$	$M_2^{(2\sigma)}$	NaN	NaN

Notes. Three examples of data points (rows) with non-detections to clarify the missing numbers handling adopted in this work. The upper table indicates the original values of magnitudes, colours and ratios, while the lower table give the corresponding values after the missing detection handling. M_1 and M_2 indicate two different bands magnitudes, while $M_1^{(2\sigma)}$ and $M_2^{(2\sigma)}$ are their 2σ detection limits, as reported in Table 1.

Appendix B: Photo- z performance

Here we report the detailed XGBoost results for photometric redshifts (evaluated on the test set), for different redshift bins, SED types, i -band AB magnitudes, and numbers of missing detections (shown graphically in Fig. 7).

Table B.1. XGBoost prediction performance.

z	N	RMSE	OLF	NMAD	$\langle \Delta z \rangle$
$2 \leq z < 8$	3645481	0.363	0.060	0.039	-0.017
$2 \leq z < 3$	2047751	0.254	0.047	0.037	0.057
$3 \leq z < 4$	860381	0.365	0.078	0.046	-0.049
$4 \leq z < 5$	407092	0.368	0.049	0.040	-0.067
$5 \leq z < 6$	195186	0.532	0.089	0.044	-0.230
$6 \leq z < 7$	91505	0.749	0.124	0.037	-0.372
$7 \leq z < 8$	43566	1.289	0.207	0.044	-0.718
SED Type	N	RMSE	OLF	NMAD	$\langle \Delta z \rangle$
Spiral	24719	0.461	0.057	0.030	-0.045
SB	293497	0.665	0.144	0.044	0.007
SF-AGN	42778	0.567	0.085	0.026	-0.109
SB-AGN	243550	0.430	0.051	0.030	-0.044
AGN1	72714	0.558	0.137	0.044	-0.062
AGN2	23009	0.480	0.102	0.041	-0.006
Ell	21535	0.219	0.027	0.038	0.052
Irr	2923679	0.297	0.050	0.040	-0.015
i -band	N	RMSE	OLF	NMAD	$\langle \Delta z \rangle$
$16 < i \leq 25$	209415	0.045	0.000	0.008	-0.010
$25 < i \leq 26$	352218	0.086	0.001	0.013	-0.013
$26 < i \leq 27$	683341	0.185	0.008	0.023	-0.018
$27 < i \leq 28$	1245136	0.332	0.043	0.050	-0.016
$28 < i \leq 29.3$	1080295	0.516	0.141	0.088	-0.019
$29.3 < i$	75076	0.631	0.119	0.072	-0.040
Missing Bands	N	RMSE	OLF	NMAD	$\langle \Delta z \rangle$
0	741823	0.082	0.001	0.014	-0.009
1	930708	0.201	0.011	0.027	-0.018
2	608234	0.391	0.059	0.051	-0.020
3	420111	0.436	0.082	0.065	-0.020
4	389306	0.431	0.095	0.079	-0.016
5	271637	0.512	0.156	0.096	-0.018
6	165679	0.632	0.209	0.109	-0.019
7	117983	0.688	0.215	0.112	-0.032

Notes. XGBoost Test set RMSE, bias, OLF, and NMAD for different redshift intervals, SED types, number of missing bands and i -band magnitudes. The redshift range $2 \leq z < 8$ comprises all the galaxies in the test set. Objects with $i > 29.3$ are not detected in the i -band.

Appendix C: Photo- z feature importance

Figure C.1 shows the most important features for redshift estimation (evaluated on the test set) for HIEROs and other objects, in terms of Shapley values (Lundberg et al. 2020). These can be both positive and negative, and indicate the contribution of every feature to the model output, as for every data point the sum of its Shapley values is the predicted redshift. Here the sum of the modulus of these values is normalised to one.

While the redshift prediction is mostly based on colours and magnitude ratios, the feature importance order of the two populations is quite different, with the HIEROs redshift estimates being more strongly reliant on the near-IR magnitudes. The colours and ratios of the same magnitude pairs are ranked closely. We note, however, the number of missing detections per band (Table 3) and the consequent reduced variability of the magnitudes, which clearly influences the importance.

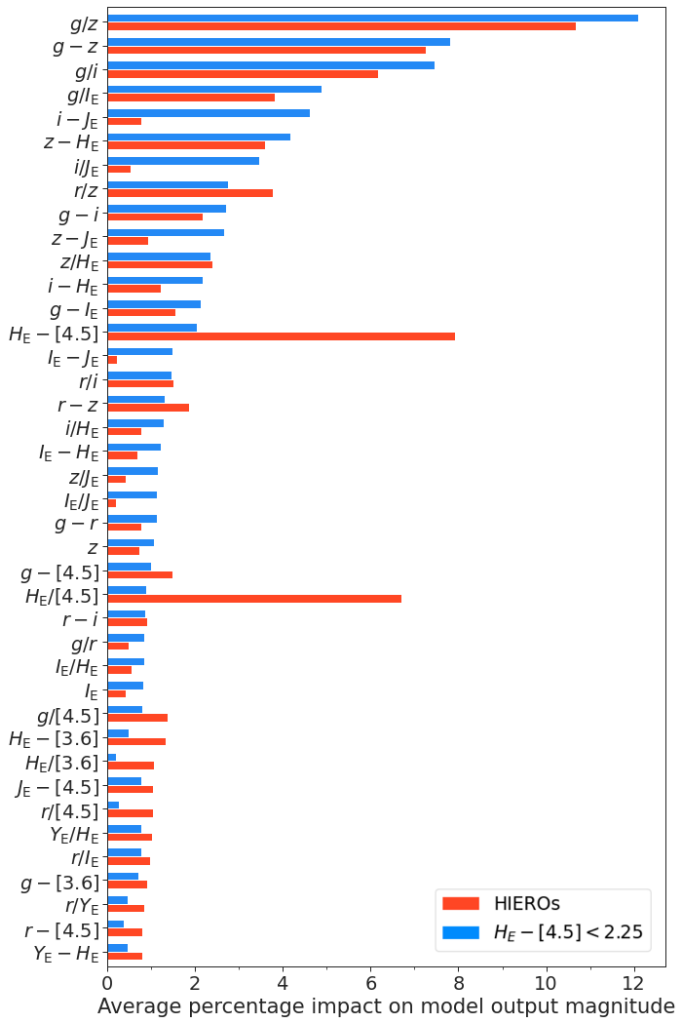


Fig. C.1. Feature rankings from the XGBoost model for HIEROs (red) and other objects (blue). We used the average of the absolute value of the normalised (by predicted redshift) Shapley values. This corresponds to the average percentage impact on model outputs. For example, on average, 8% of the model output for HIEROs is based on the $H_E - [4.5]$ colour, while for other objects the impact of this feature is only around 2%.

Appendix D: Contaminants analysis

Throughout the paper, we considered only objects at $z > 2$, under the assumption that the selection of high-redshift objects from official photo- z pipelines would yield a clean dataset. In this appendix we analyse potential contamination issues from low-redshift ($z < 2$) sources, as well as brown dwarfs, in the selection of $3 < z < 7$ HIEROs candidates.

To address the first concern, we applied a dedicated binary classifier using the same XGBoost algorithm and input features as described in the main body of the paper (see Sect. 4.2), with the goal of differentiating between objects at $z < 2$ and $z > 2$. This classifier is trained on a dataset comprehending objects down to $z = 0$ and 500 HIEROs (see Sect. 4.3). This training set consisted of 2 815 072 galaxies, the vast majority of which are at $z < 2$ (see, for example, Fig. 1, top panel).

The objective was to estimate the degree of contamination from low- z sources in the HIEROs selection (Eq. 4), essentially quantifying the number of galaxies with $z < 2$ that both fit within the colour-colour selection and that are predicted by the classifier at $z > 2$. We report the completeness and fraction of contaminants obtained from this analysis in Table D.1, reminding the reader that in the testing phase we consider only observations with at least four bands with $S/N > 2$ and at least one band with $S/N > 5$. These results underscore the classifier's effectiveness in distinguishing between $H_E - [4.5] > 2.25$ sources with redshifts above and below 2, showcasing high completeness at $z > 2$ (98.9%) and high purity, thereby minimising the risk of contamination (kept at only 1.1%).

Table D.1. Confusion matrix for photometric redshift ranges for HIEROs.

	$z < 2$	$z \geq 2$
$\tilde{z} < 2$	59.7% (2882)	39.9% (1914)
$\tilde{z} \geq 2$	1.1% (1948)	98.9% (182 527)

Notes. z and \tilde{z} indicate true and predicted redshift class, respectively. The diagonal entries represent the completeness of the samples, while the off-diagonal terms their contaminants. In parenthesis we report the numbers of objects.

To assess the contamination by brown dwarfs, we simulated *Euclid* magnitudes based on the models of L and T dwarf from Burrows et al. (2006). These templates have effective temperatures ranging from 700 K to 2300 K, metallicities from $[Fe/H] = -0.5$ to 0.5, and gravities from $10^{4.5}$ to $10^{5.5} \text{ cm s}^{-2}$. We did not simulate their spatial distribution, as this goes beyond the scope of this paper, but we derived magnitudes assuming they are located from 1 to 288 pc away from us, with steps of 1 pc. At larger distances, all brown dwarfs models correspond to $S/N < 3$ in all *Euclid* filter. We kept all mocks that results in a $S/N > 3$ in at least one *Euclid* filter, for a total of 33 866 simulated brown dwarfs. We then scattered their magnitudes using the same procedure reported in Sec. 2.2. We find that fewer than 1% of these brown dwarfs fall into the selection $H_E - [4.5] > 2.25$ and thus the contamination from these sources in the HIEROs selection (applied to all objects in our catalogue with $z > 2$) is below 0.01%.

Consequently, we argue that the conclusions drawn in the main body of the paper remain largely unaffected even with the possible inclusion of $z < 2$ galaxies and brown dwarfs in the analysis.

Appendix E: Identifying HIEROs with shallower *Rubin* depths

The analysis presented in this work assumes *Rubin* depths after 10 years of observations. To evaluate the efficacy of our search for the reddest high-redshift galaxies before that time, we run the same tests described above with shallower depths (Table E.1), corresponding to one and four years of *Rubin* observations (Brandt et al. 2018). In all our runs, we considered 500 HIEROs in the training set.

As expected, the photo- z performance improves with each passing year, while also benefiting from a larger sample of HIEROs in the test set. The decrease in RMSE and $\langle\Delta z\rangle$ from year one to year 10 is a clear indication that the accuracy of the predictions has improved significantly over time. However, the most notable improvement is in the catastrophic OLF metric, which decreased from 0.254 in year one to 0.141 in year 10. This is a significant improvement, and it suggests that deeper observations are particularly important in reducing exceptionally wrong predictions.

Table E.1. Expected 5σ *Rubin* depths after one and four years of observations.

<i>Rubin</i> year	<i>Rubin/u</i>	<i>Rubin/g</i>	<i>Rubin/r</i>	<i>Rubin/i</i>	<i>Rubin/z</i>
1	24.5	25.7	25.8	25.1	24.4
4	25.2	26.4	26.5	25.8	25.1

Notes. To be compared with the ones after 10 years, reported in Table 1.

Table E.2. Redshift prediction performance for HIEROs obtained considering *Rubin* depths after one and four years of observations.

<i>Rubin</i> year	N	RMSE	OLF	NMAD	$\langle\Delta z\rangle$
1	115 303	0.864	0.301	0.112	-0.064
4	115 371	0.858	0.229	0.110	-0.031
10	184 441	0.654	0.123	0.068	-0.084

Notes. N indicates the number of objects in the selection. For ease of comparison, we report again the results after 10 years of *Rubin* observations, already presented in Table 5.

Appendix F: *Rubin/y*

We did not include in our analysis the *Rubin/y* band (2σ depth of 27.2 mag), as it was determined to have negligible impact on our results. We address this point by showing in Table F.1 the improvement in performance between the dataset with and without the *Rubin/y* band. This results demonstrate that the inclusion of this band yields little improvement on our results across most redshifts and magnitude intervals.

Table F.1. XGBoost improvement in RMSE, OLF, and NMAD with the inclusion of the *Rubin/y* band.

z	RMSE	OLF	NMAD	$\langle\Delta z\rangle$
$2 \leq z < 8$	0.011	0.005	0	-0.003
$2 \leq z < 3$	0	0.001	0.001	0.001
$3 \leq z < 4$	0.010	0.006	0.001	0
$4 \leq z < 5$	0.007	0.004	0	0
$5 \leq z < 6$	0.026	0.012	0.001	-0.028
$6 \leq z < 7$	0.069	0.029	0.001	-0.051
$7 \leq z < 8$	0.140	0.037	0.004	-0.106
<i>i</i> -band	RMSE	OLF	NMAD	$\langle\Delta z\rangle$
$16 < i \leq 25$	-0.002	-0.001	0	-0.001
$25 < i \leq 26$	-0.001	0	0	0
$26 < i \leq 27$	0.015	0.002	0	-0.001
$27 < i \leq 28$	0.019	0.005	0.001	-0.002
$28 < i \leq 29.3$	0.020	0.014	0.004	-0.003
$29.3 < i$	-0.008	-0.009	0.005	-0.014

Notes. XGBoost Improvement in test set RMSE, OLF, and NMAD, bias for different redshift intervals *i*-band magnitudes, as obtained with the inclusion of the *Rubin/y* band. Objects with $i > 29.3$ are not detected in the *i*-band.