A&A, 709, A79 (2026)
https://doi.org/10.1051/0004-6361/202658903

Astronomy & Astrophysics

Star–galaxy classification in deep LSST data with random forest

A pilot study on the Data Preview 1 release

M. Gatto[1,⋆], V. Ripepi[1], M. Bellazzini[2], C. Tortora[1], and M. Dall'Ora[1]

[1] INAF – Osservatorio Astronomico di Capodimonte, Salita Moiariello, 16, 80131 Napoli, Italy
[2] INAF – Osservatorio di Astrofisica e Scienza dello Spazio, Via Gobetti, 93/3, 40129 Bologna, Italy

Received 9 January 2026 / Accepted 24 March 2026

ABSTRACT

Context. The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) will produce unprecedentedly deep and wide photometric catalogues, enabling transformative studies of faint stellar systems such as the research of ultra-faint dwarf (UFD) galaxies. A critical challenge for these studies is reliable star–galaxy separation at faint magnitudes, where compact background galaxies increasingly contaminate stellar samples.

Aims. This work aims to assess the performance of supervised machine-learning techniques for star–galaxy separation in LSST-like data, to quantify the relative importance of morphological and photometric information, and to identify the most effective combinations of input features for minimizing galaxy contamination while preserving stellar completeness in the faint regime relevant for UFD searches.

Methods. We applied a Random Forest classifier to observations of the Extended Chandra Deep Field South from LSST Data Preview 1 (DP1), the deepest field observed within the DP1. We constructed a curated sample of bona fide stars and galaxies using spectroscopic data, Gaia DR3, and multi-band photometric catalogues. We trained and validated the classifier using several configurations of LSST-based input features, including multi-band colours, the LSST morphological parameter REFEXTENDEDNESS, and photometric uncertainties.

Results. We find that LSST multi-band photometry alone delivers a good star–galaxy separation, significantly outperforming morphology-based classification at faint magnitudes. Colours involving the u band are essential to provide a robust star-galaxy separation. Furthermore, explicitly including photometric uncertainties as input features yields the best overall performance. Across all configurations that include all the six LSST filters, galaxy contamination remains negligible almost the whole magnitude range probed in this work (i.e. $r \lesssim 27.5$ mag).

Conclusions. Our results demonstrate that supervised machine-learning methods, when combined with LSST multi-band photometry, can effectively suppress galaxy contamination in deep stellar catalogues, ensuring that searches for UFDs are not significantly compromised. Given that the DP1 data are shallower and have poorer seeing than the final LSST survey, our findings should be regarded as a conservative lower limit on the performance achievable with the full 10-year dataset. To facilitate further development, we will publicly release the curated star–galaxy sample used in this work.

Key words. methods: data analysis – methods: statistical – techniques: photometric – surveys – stars: general – galaxies: general

1. Introduction

Since the beginning of the new millennium, our understanding of the formation and evolution of the Milky Way (MW) has advanced substantially, largely owing to wide and deep panchromatic surveys conducted with both ground-based facilities, such as the Sloan Digital Sky Survey (SDSS[1]), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016), the Dark Energy Survey[2] (DES), and space missions such as Gaia (Gaia Collaboration 2016). These surveys have produced extensive and accurate stellar catalogues, enabling detailed reconstructions of the MW's past and ongoing mass assembly, and, more broadly, of the evolution of the Local Group. Most current wide-field surveys reach an average depth of $r \simeq 24$ mag, meaning they are able to resolve main-sequence turn-off (MSTO) stars that belong to an old stellar population ($t \sim 12$ Gyr) at a heliocentric distance of ∼150 kpc.

The Vera C. Rubin Observatory[3] (VRO), now in operation, will dramatically extend this capability through the 10-year Legacy Survey of Space and Time (LSST). The LSST will repeatedly image the Southern Hemisphere in six photometric bands (*ugrizy*), ultimately achieving a coadded depth of $r \simeq 27.5$ mag (Ivezić et al. 2019). At this depth, main-sequence stars will be resolved well beyond the presumed edge of the Galactic halo. An immediate science case enabled with this unprecedented vast, deep, and precise catalogue of stars will be the detection of dwarf galaxies, by means of their resolved stellar population, out to ∼5 Mpc across the entire southern sky (Mutlu-Pakdil et al. 2021).

However, realizing this potential requires an effective and reliable star–galaxy classifier, particularly at faint magnitudes where contamination from unresolved galaxies becomes severe. Fadely et al. (2012) demonstrated that, in the colour range typical of old, metal-poor MSTO stars inhabiting the MW halo ($g - r \lesssim 1.0$ mag), unresolved galaxies begin to outnumber stars

⋆ Corresponding author: massimiliano.gatto@inaf.it
[1] https://www.sdss.org/
[2] https://www.darkenergysurvey.org/
[3] https://rubinobservatory.org/

at $r \gtrsim 23.5$ mag, with their relative fraction increasing sharply towards fainter magnitudes. This contamination poses a major obstacle for the search of ultra-faint dwarf (UFD) galaxies, the most ancient, metal-poor, and dark matter-dominated galaxies known to date (Simon 2019). Detection techniques for UFDs rely on identifying overdensities of old, metal-poor stars over limited regions of the sky (e.g. Walsh et al. 2009), and inadequate star–galaxy separation can therefore lead to catastrophic misclassifications, such as mistaking a projected overdensity of unresolved background galaxies (e.g. a distant group or cluster) for a nearby dwarf galaxy.

In the absence of spectroscopic radial velocity measurements, whose acquisition over a large portion of the sky is prohibitively expensive in terms of telescope time, source classification generally relies on morphological diagnostics. Standard pipelines exploit shape parameters, such as source roundness, or flux ratio tests. In the LSST pipeline, for example, the OBJECT catalog[4] includes the REFEXTENDEDNESS parameter, which serves as a binary star–galaxy discriminator (NSF-DOE Vera C. Rubin Observatory 2025b). This parameter is assigned based on the ratio between cModel[5] and PSF fluxes, with sources classified as point-like if $f_{\rm psf} \geq 0.985 f_{\rm cmodel}$[6] (Choi et al. 2025). Nonetheless, the classification degrades at faint magnitudes: tests with artificial source injection in fields observed in LSST Data Preview 1 (DP1) indicate that the fraction of correctly classified stars drops to ∼50% at $i = 23.8$ mag, and at $i \simeq 24.5$–25 mag nearly 20% of all sources are incorrectly classified (see NSF-DOE Vera C. Rubin Observatory 2025b, their Section 5.6).

To improve upon such limitations, several studies have explored more sophisticated approaches, often leveraging multi-band photometry (e.g. Pennock et al. 2025). Fadely et al. (2012) compared maximum likelihood methods, hierarchical Bayesian classifiers, and support vector machines (SVMs), applying them to the 30-band COSMOS survey (Scoville et al. 2007), which reaches $r \simeq 25$ mag over 2 deg^2. They found that SVMs and hierarchical Bayesian classifiers yield the most effective star–galaxy separation. Logan & Fotopoulou (2020) applied the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to a dataset of about 50 000 spectroscopically labelled sources from the sample of Fotopoulou & Paltani (2018), aiming to distinguish stars, galaxies, and quasars in a multidimensional colour space ($ugrizYHJKW_1W_2$). They achieved an accuracy of ≥98% down to $i \simeq 24$–25 mag. Jeakel et al. (2026) applied an XGBoost classifier trained on the ∼60-band Javalambre Physics of the Accelerating Universe Astrophysical Survey[7] (J-PAS), combining photometric and morphological features to distinguish stars and galaxies down to $r \sim 23.5$ mag, showing that incorporating morphology significantly enhances performance. Zhang et al. (2025) combined a Bayesian photometric redshift estimator with a multilayer perceptron to separate artificially injected stars and galaxies in Kilo-Degree Survey (KiDS) Data Release 4-like images (Kuijken et al. 2019). They reported a negligible false positive rate (FPR) down to MAG_AUTO ≤ 25 mag, although the true positive rate (TPR) drops below 40% at the same magnitude. Khramtsov et al. (2019) applied CatBoost, a gradient-boosting decision-tree algorithm, to the nine-band ($ugriZYJHK_s$) KiDS DR4 + VIKING dataset, obtaining TPRs of ≃98–99% down to $r \sim 22$ mag for stars, galaxies, and quasars. More recently, Feng et al. (2025) employed a multimodal neural network on the nine-band KiDS Data Release 5 (Wright et al. 2024) + VIKING dataset, to classify stars, galaxies, and quasars, achieving an overall accuracy of ∼99% down to $r \sim$ 23 mag. Nevertheless, these efforts probe significantly shallower regimes than LSST's final depth ($r \simeq 27.5$ mag), and thus may not fully capture the classification challenges LSST will face.

The purpose of this work is to show that a straightforward, photometry-based machine-learning model can already suppress much of the unresolved galaxy contamination that would otherwise hinder UFD searches in LSST, even though future, more advanced techniques may further improve upon the method presented here. To this aim, we test the application of a random forest classifier, a supervised machine-learning algorithm, for star–galaxy separation in the recent released LSST Data Preview 1 (DP1; NSF-DOE Vera C. Rubin Observatory 2025b). Throughout this work, we explore multiple configurations of input features, primarily constructed from LSST multi-band photometry, and optionally incorporating additional features (including, but not limited to, the REFEXTENDEDNESS parameter provided in the DP1 catalog), as well as cases in which selected features are intentionally excluded to assess their impact on classification performance. Colour–colour diagrams are known to be powerful for this task, as stars occupy relatively narrow loci while galaxies are more broadly distributed, thereby enhancing separability. In particular, optical–infrared colour combinations have proven to be robust discriminants (e.g. Maddox et al. 2008; Tortora et al. 2018). To emulate LSST's final depth, we exploit the Extended Chandra Deep Field-South (ECDFS) region available within LSST DP1, which has been observed multiple times. Although this field represents the deepest dataset currently released by LSST, its properties remain significantly shallower than the expected 10-year survey. In particular, the ECDFS area reaches a 5σ point-source depth of only $r \simeq 26$ mag (NSF-DOE Vera C. Rubin Observatory 2025b), about 1.5 mag brighter than LSST's nominal coadded depth. Moreover, the median point-spread function FWHM is ≃ 1.14″(NSF-DOE Vera C. Rubin Observatory 2025b), considerably worse than the ≃ 0.65″median seeing anticipated for LSST operations (Ivezić et al. 2019). As a consequence, both the depth and image quality of DP1 fall short of what the final survey will deliver. These limitations imply that the performance we obtain should be regarded as a conservative lower bound on what a random forest classifier may achieve when applied to the full 10-year LSST dataset.

The paper is structured as follows: Sect. 2 presents the LSST DP1 dataset and the literature catalogues used to assemble the sample of bona-fide stars and galaxies. Section 3 details the optimization of the random-forest hyperparameters and discusses the resulting classification performance for different choices of input features. Finally, Sect. 5 provides a summary and outlines the implications of this work.

2. Data

2.1. The LSST Data Preview 1

The Data Preview 1 (DP1) represents the first data release obtained with the Vera C. Rubin Observatory (NSF-DOE Vera C. Rubin Observatory 2025b). DP1 includes raw and calibrated single-epoch images and their corresponding SOURCE

[4] The OBJECT catalogue provides measurements of objects detected in deep coadd images, with a signal-to-noise ratio (S/N) of >5 in at least one band, (NSF-DOE Vera C. Rubin Observatory 2025a).
[5] cModel is a composite flux obtained from the linear combination of de Vaucouleurs and exponential fluxes in each band (see https://dp1.lsst.io/tutorials/notebook/201/notebook-201-1.html).
[6] https://sdm-schemas.lsst.io/dp1.html
[7] https://www.j-pas.org

catalogues, as well as stacked deep coadd images and the associated OBJECT catalogues, among other products. A detailed description of the DP1 data products is provided in the official documentation (NSF-DOE Vera C. Rubin Observatory 2025). DP1 is based on 48 nights of observations collected during the first on-sky commissioning campaign with the LSSTComCam. The dataset covers a total area of ∼15 deg^2, distributed across seven distinct, non-contiguous fields. Details of the LSSTComCam campaign and of DP1 can be found in SITCOM-149[8] and in the DP1 overview paper (NSF-DOE Vera C. Rubin Observatory 2025b).

For the purposes of this work, we focus on the Extended Chandra Deep Field-South (ECDFS) field, which benefits from the most extensive and homogeneous temporal coverage among the DP1 fields. In particular, ECDFS was observed 43, 230, 237, 162, 153, and 30 times in the u, g, r, i, z, and y filters, respectively. This cadence enables a 5σ point-source depth of $r \simeq 26$ mag. We retrieved the data from the OBJECT table, which contains measurements of sources detected in the coadded images. The ADQL query used for the extraction is the following:

```
SELECT objectId, coord_ra, coord_dec, ebv,
refExtendedness, u_psfMag, u_psfMagErr,
u_psfFlux_flag, g_psfMag, g_psfMagErr,
g_psfFlux_flag, r_psfMag, r_psfMagErr,
r_psfFlux_flag, i_psfMag, i_psfMagErr,
i_psfFlux_flag, z_psfMag, z_psfMagErr,
z_psfFlux_flag, y_psfMag, y_psfMagErr,
y_psfFlux_flag
FROM dp1.Object
WHERE CONTAINS(POINT('ICRS',coord_ra,coord_dec),
CIRCLE('ICRS', 53.16, --28.1, 1.0)) = 1
```

We retained only sources for which the *_psfFlux_flag* was not set, as a triggered flag indicates that the PSF photometry in that band failed and was therefore replaced by forced photometry. In addition, we imposed a bright-end cut to remove saturated sources. Specifically, Choi et al. (2025) estimated the saturation bright limit to be ∼15.70, 15.08, 15.19 mag in the g, r, and i bands, respectively. We adopted these values as upper limits on magnitudes. This query returned 356 786 objects within a circular region of radius 40′ (area ∼1.4 deg^2), centred at (RA, Dec) = (53.16°, –28.1°).

2.2. Spectroscopic catalogues

Random Forest is a supervised machine-learning algorithm, meaning that it learns to classify sources based on a training set of objects with secure labels, i.e. sources confidently identified as stars or galaxies. To construct a robust training set, we cross-matched our LSST catalogue with spectroscopic surveys available in the literature that overlap with the ECDFS region. Spectroscopic redshifts provide an unambiguous physical discriminator between stars and galaxies, making them an ideal source of ground truth labels. Specifically, we used the following catalogues:

The 3D-HST survey is a 248-orbit Treasury program with the Hubble Space Telescope (HST) that employed the WFC3/G141 slitless grism to obtain near-infrared spectra of galaxies up to redshift $z < 3$ (Momcheva et al. 2016). The survey covers five deep extragalactic fields, including one centered on the Chandra Deep Field South (CDFS), and is complemented by extensive multiwavelength imaging from both ground- and space-based facilities, enabling high-quality photometric redshift estimates (see also Skelton et al. 2014). The spectroscopic catalogue contains 4795 spectroscopic redshifts, with stars identified by a value of $z_{\rm best} = -1$[9]. Cross-matching with ECDFS (within a tolerance of 1″) yielded 4294 objects, the majority being galaxies (see Table 1).

Table 1. Number of stars and galaxies retrieved from external catalogues.

Catalogue	N_{stars}	$N_{galaxies}$	Reference
3D-HST	374	3920	Momcheva et al. (2016)
GOODS/VIMOS	50	1105	Popesso et al. (2009); Balestra et al. (2010)
GOODS/FORS2	0	130	Vanzella et al. (2008)
GMASS	0	22	Kurk et al. (2013)
CANDELS	0	213	Kodra et al. (2023)
VANDELS	1	455	Talia et al. (2023)
MUSE HUDF	0	381	Bacon et al. (2023)
JADES	0	194	D'Eugenio et al. (2025)
VVDS	0	634	Le Fèvre et al. (2013)
MUSYC	39	143	Cardamone et al. (2010)
OzDES	146	972	Lidman et al. (2020)
ACES	125	3081	Cooper et al. (2012)
GAIA DR3	2004	0	Gaia Collaboration (2023)
SIMPLE	165	0	Damen et al. (2011)
COMBO-17	178	0	Wolf et al. (2004)
DESI	1279	0	Duncan (2022)
TOTAL	4361	11 250	

The Great Observatories Origins Deep Survey (GOODS) targeted two deep fields: Hubble Deep Field North (GOODS-N) and CDFS (GOODS-S), with multiwavelength coverage from infrared to X-ray, aiming to address fundamental questions of galaxy and AGN formation, as well as the distribution of baryonic and dark matter at high redshift (Vanzella et al. 2008; Popesso et al. 2009; Balestra et al. 2010). Spectroscopic data were obtained with VIMOS at VLT, yielding a total of 3218 redshifts, with stars classified as Star in the column COMM. We selected 2223 sources with high-quality flags (QF = A or B), of which 1155 matched counterparts in our ECDFS catalog, again predominantly galaxies.

As part of the GOODS project, Vanzella et al. (2008) carried out deep spectroscopic observations of GOODS-S with the FORS2 at VLT instrument. Of the 1165 available redshifts, (with stars classified as Star in the column comments), we selected 972 with quality column = A or B. The cross-match added 130 sources to our secure catalog, all of them are galaxies. The Galaxy Mass Assembly ultra-deep Spectroscopic Survey (GMASS) was designed to probe both massive quiescent and star-forming galaxies at $z > 1.4$, with particular focus on the peak epoch of galaxy assembly around $z \sim 2$ (Kurk et al. 2013). Targets were drawn from the CDFS/GOODS-S field. Of the 210 spectroscopic redshifts available, we selected 192 with reliable quality flags (q_zsp = 1). The crossmatch with ECDFS field yielded further 22 sources, all of which correspond to galaxies.

The Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) combines high spatial resolution imaging from HST with complementary intermediate-resolution

[8] https://sitcomtn-149.lsst.io/

[9] The full catalogue also includes photometric redshifts for more than 45 000 sources. We opted to restrict only to sources with spectroscopic measurements.

optical and IR imaging. Kodra et al. (2023) published the first large-scale release of photometric redshifts derived by the CANDELS collaboration, based on a spectroscopic training set of 5807 high-quality redshifts covering all five survey fields. Within the GOODS-S region, 2312 sources possess reliable spectroscopic redshifts. Our cross-match with the ECDFS catalogue identified 213 unique counterparts, all of which are galaxies.

The VANDELS survey is a deep spectroscopic campaign carried out with the VIMOS at VLT, designed to probe in detail the physical properties of high-redshift galaxies. It targeted approximately 2100 galaxies within the redshift range $1 < z < 6.5$, located in the CDFS and Ultra Deep Survey (UDS) regions. Talia et al. (2023) presented the final VANDELS data release, which includes 880 sources in the CDFS field, 679 of which have secure spectroscopic redshifts (column zflg = 3 or 4). Our cross-match with the ECDFS catalogue yielded 456 common sources,all but one of which are galaxies; the only stellar object in the sample corresponds to the source with spectroscopic redshift equal to zero.

The MUSE Hubble Ultra Deep Field (HUDF) survey represents one of the deepest spectroscopic explorations ever conducted of the GOODS-S field. The second public data release (Bacon et al. 2023) includes a total of 2221 extracted spectra, of which 1711 have secure redshift determinations (column zconf = 2 or 3). Our cross-match with the ECDFS catalogue yielded 381 new galaxies.

The JWST Advanced Deep Extragalactic Survey (JADES) is an observational program designed to particularly investigate galaxies at $z > 3$. D'Eugenio et al. (2025) presented the third data release of JADES, which combines deep NIRCam imaging with extensive NIRSpec spectroscopy over the GOODS-N and GOODS-S fields. From a total of 2525 extracted spectra, 914 have high-quality redshift flags (z_Spec_flag = a or b). Our cross-match with the ECDFS catalogue resulted in 194 associations, all corresponding to galaxies.

The VIMOS VLT Deep Survey (VVDS) represents one of the most extensive spectroscopic efforts to map galaxy evolution across cosmic time, since $z \sim 6.7$. Spectroscopic observations were carried out with the VIMOS at VLT. The final release includes more than 45 000 sources (Le Fèvre et al. 2013). Among these, 1323 sources in the ECDFS have reliable spectroscopic redshifts ($f_z > 1$). Our cross-match yielded 634 counterparts, all identified as galaxies.

The Multiwavelength Survey by Yale–Chile (MUSYC) provides deep optical and near infrared imaging across four different fields. In particular, Cardamone et al. (2010) presented the dataset covering the ECDFS region, obtained with the Subaru Telescope using 18 medium-band filters. These observations were complemented with broad-band optical and near-infrared data ($UBVRIzJHK$) as well as Spitzer/IRAC imaging, resulting in a uniform multiwavelength catalog. To assess the accuracy of the derived photometric redshifts, they adopted spectroscopic measurements available from the literature. From the spectroscopic subsample of 257 high confident redshifts (q_zsp $\geq$ 2), our cross-match with the ECDFS catalogue yielded 182 sources: 39 stars (sources with spectroscopic redshift equal to zero, and 143 galaxies.

The Australian Dark Energy Survey (OzDES) is a spectroscopic campaign conducted with the 2dF fibre positioner and the AAOmega spectrograph on the 3.9-m Anglo-Australian Telescope. Its primary goal was to complement the Dark Energy Survey (DES) by obtaining spectroscopic redshifts for transient hosts, active galactic nuclei, and galaxies located within the ten DES deep fields. Lidman et al. (2020) presented the second OzDES data release, comprising nearly 30 000 reliable redshifts ($q_{op} = 3, 4$ for galaxies and $q_{op} = 6$ for stars) from a total of 38 624 spectra. Our cross-match with the ECDFS catalogue returned 1118 unique sources, including 146 stars, and 972 galaxies. The Arizona CDFS Environment Survey (ACES) is a spectroscopic program carried out with the IMACS spectrograph on the Magellan-Baade telescope, designed to significantly improve the sampling of galaxy redshifts across the ECDFS. From the full ACES catalogue of 12 983 entries, 6601 objects have high-quality redshift classifications (stars with Q = –1, galaxies with q = 3, 4; Cooper et al. 2012). Our cross-match with the ECDFS catalogue identifies 3206 matched sources, including 125 stars and the remainder classified as galaxies. Table 1 summarizes the number of stars and galaxies obtained from each spectroscopic catalogue after cross-matching with our sample of sources.

2.3. Gaia Data Release 3

The cross-matched catalogue from spectroscopic surveys yielded a total of 11 985 sources, the vast majority of which – 11 250 objects, or ∼94% – are galaxies. Such an imbalanced training set is problematic, as it fails to adequately sample the photometric parameter space of both classes. To mitigate this imbalance and increase the number of securely identified stars, we cross-matched ECDFS with Gaia Data Release 3 (Gaia Collaboration 2023), which provides precise parallaxes and proper motions that are highly effective for distinguishing Galactic stars from distant galaxies. The cross-match yielded 2770 additional sources with no prior spectroscopic counterpart. Of these, 2088 have reliable astrometric solutions with RUWE <1.4. Restricting further to sources with a star probability $P_{SS} > 0.99$, we obtained a final subsample of 2004 high-confidence stars. The combined catalogue therefore comprises 13 989 sources, of which 2630 are stars (19%) and 11 359 are galaxies (81%).

2.4. Multi-photometric catalogues

Given the relatively bright limiting magnitude of Gaia, the resulting stellar subsample is dominated by sources with $r \leq 21$ mag. A potential concern in relying solely on these objects is that a machine-learning classifier trained only on bright stars may not generalize well to the much fainter regime, where the separation between stars and unresolved galaxies becomes significantly more challenging. To mitigate this limitation, we supplemented the stellar sample with stars identified through multi-band photometry, leveraging catalogues that include either a large number of photometric bands or mid-infrared observations, where stellar and galaxy spectral energy distributions diverge more strongly. It is worth noting that although some of these stars were originally classified based on high-dimensional photometric information, the goal of this study is to assess how effectively a machine-learning classifier can separate stars and galaxies, and can reduce the contamination from compact galaxies at faint magnitudes, using only LSST filters. In this sense, our approach should be regarded as a pilot investigation aimed at testing the viability and performance of such methods in the photometric regime relevant for LSST. The inclusion of stars identified from deep multi-band data simply allows the classifier to sample the regions of colour-colour space where Gaia becomes incomplete, thus ensuring that the model is trained across the full dynamic range of interest.

We adopted the following catalogues to increase the number of stars at faint magnitudes:

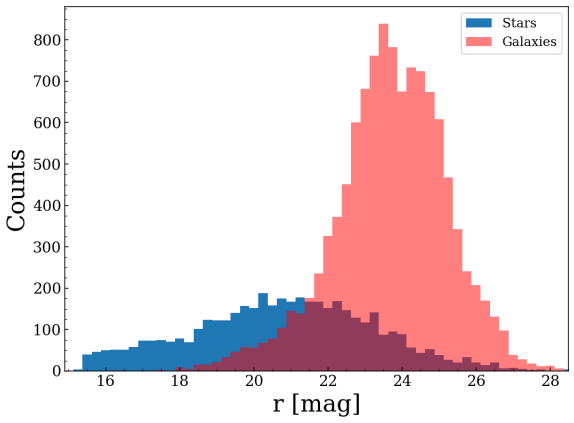

Fig. 1. r-band magnitude distribution of the full sample of bona fide stars and galaxies. Stars are shown in blue and galaxies in red.

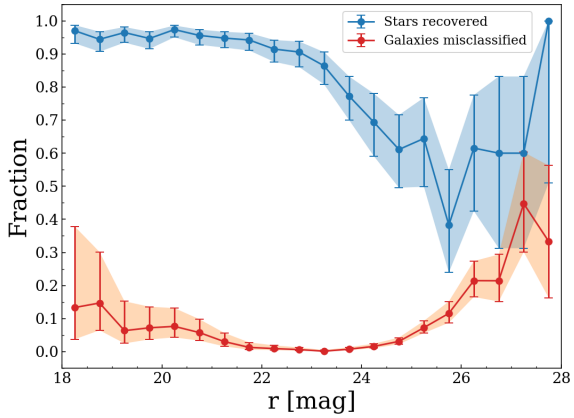

Fig. 2. Performance of the REFEXTENDEDNESS parameter as a function of the r-band magnitude for our catalog. The solid blue line shows the fraction of true stars correctly identified as stars, while the solid red line indicates the fraction of true galaxies misclassified as stars.

- The SIMPLE survey (Spitzer IRAC/MUSYC Public Legacy Survey in the Extended CDF-S) provides deep IRAC imaging (3.6, 4.5, 5.8, and 8.0 μm) over ≃1600 arcmin^2 surrounding the GOODS-S field. The IRAC data are matched to MUSYC optical and near-infrared imaging ($UBVRI_zJHK$), yielding a catalogue of 61 233 total sources, of which 43 782 are detected at $S/N > 5$ at 3.6 μm, with 19 993 objects having full 13-band photometry (Damen et al. 2011). Our cross-match with the ECDFS catalogue resulted in 165 stars.
- The COMBO-17 survey provides 17-band optical photometry (350–930 nm) covering the CDFS region. The catalogue contains 63 501 sources with a multi-colour classification into stars, galaxies, and QSOs (Wolf et al. 2004). The cross-match with this catalogue provided 178 stars.
- We also included stellar sources identified in the Dark Energy Spectroscopic Instrument (DESI) Legacy Imaging Surveys. The dataset combines optical (grz) and mid-infrared photometry (3.4, 4.6, 12, and 22 μm) from WISE and NEOWISE (Duncan 2022). From this catalog, we selected sources with a stellar probability $p\_star \geq 0.9$, providing further 1279 likely stars to our catalog.

To assess the purity of the multi-photometric catalogues, we cross-matched them with spectroscopic catalogues described in Sect. 2.2. A matching radius of 1″ was adopted. We obtained 2917 unique sources classified as stars in the multi-photometric catalogues and having a spectroscopic counterpart. Among them, 2488 are spectroscopically confirmed as stars, corresponding to a stellar purity of 85.3%.

2.5. The full sample of bona-fide stars and galaxies

Our final curated catalogue contains 15 611 sources, of which 4361 are classified as stars (28%) and 11 250 as galaxies (72%), as summarized in Table 1. This final catalogue will be made publicly available for future scientific use. Figure 1 displays the r-band magnitude distribution of this final sample. Stars (blue histogram) predominantly occupy the bright regime, with a peak at $r \simeq 21$ mag and a rapid decline beyond $r \approx 24$ mag. Conversely, galaxies (red histogram) peak at significantly fainter magnitudes ($r \sim 23$–24 mag). This behaviour is naturally linked to the characteristics of the input catalogues: many of the stars in our sample originate from Gaia DR3 or from high–quality multi-band surveys, both of which preferentially include relatively bright stellar sources. At the same time, this distribution is also representative of what is expected observationally: samples tend to be star-dominated at bright magnitudes, while progressively fainter magnitudes are increasingly populated by compact galaxies (Fadely et al. 2012). The decline in the number of stars at $r \gtrsim 24$ mag in our compiled catalogue therefore mirrors the observational regime in which searches for UFDs are typically conducted. Importantly, this does not hinder the goals of the present work. Our aim is to assess whether machine-learning classifiers trained on LSST colours can effectively reduce the contamination from unresolved galaxies in stellar samples. This aspect is particularly critical for detecting very low-surface-brightness systems in the Local Volume, where contamination, rather than completeness, might be the limiting factor.

Figure 2 illustrates the performance of the REFEXTENDEDNESS morphological classifier on our sample. In this case, we plotted only the 14 360 sources that have complete photometry in all six LSST filters and a valid value of the REFEXTENDEDNESS parameter. The fraction of stars correctly identified decreases steadily with magnitude, reaching ≃60% at $r \simeq 24.5$–25 mag, and is subject to large uncertainties due to the small number of stars in this regime. In contrast, galaxy contamination steadily increases from $r \simeq 24.5$ mag and remains below ∼20% up to $r \sim 26$ mag before rising sharply, exceeding ∼40% at $r \gtrsim 27$ mag. This confirms that REFEXTENDEDNESS is effective at bright and intermediate magnitudes but loses discriminating power near the LSST detection limit, where morphology alone becomes unreliable. Nevertheless, its relatively low galaxy contamination over a broad magnitude range persuaded us to test also the inclusion of REFEXTENDEDNESS as an additional feature in our random forest classifier.

3. The Random Forest classifier

Random Forest (RF) is a supervised ensemble learning algorithm widely used for classification tasks (Breiman 2001). It operates by constructing a large number of decision trees, each trained on a subset of the training data. At each split in a tree, only a random subset of features is considered, which introduces additional randomness and decorrelation among the trees, thus mitigating overfitting. RF classifiers have several properties that make them particularly well suited to star–galaxy separation. First, they can naturally handle heterogeneous and correlated input features (such as colours derived from LSST magnitudes), without requiring strong assumptions on the underlying data distribution. Second, they are robust against noisy measurements

and outliers, a key aspect given the faint magnitude regime probed by LSST. Third, RFs provide a quantitative estimate of feature importance, allowing us to identify which colour combinations contribute most effectively to the separation between stars and galaxies.

To optimize model performance during photometric analysis, we de-reddened the stellar magnitudes across all six filters using the coefficients provided by Schlafly & Finkbeiner (2011). In the next sections, we describe the procedures we adopt to set RF parameters, train the model, and test its performance.

3.1. Tuning the Random Forest parameters

To implement the RF classifier we used the Python library *scikit-learn* (Pedregosa et al. 2011), which provides a flexible and efficient framework for model training, evaluation, and prediction. The performance of a RF depends on a set of hyperparameters, such as the number of trees in the ensemble, the maximum depth of each tree, and the minimum number of samples required to split a node, that are not learned during training but must be specified by the user. A common strategy is to explore the hyperparameter space and identify the combination that yields the best-performing model. To assess model performance, a suitable evaluation metric must be chosen, such as accuracy, F1 score, or recall. In this work, we adopted the F1 score as the primary metric, which is defined as

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{1}$$

where

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

is the fraction of predicted positives that are correct, and

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

is the fraction of true positives that are successfully recovered. We split the dataset into a training set (70% of the sources), and an evaluation sample of the performance of the model (30% of the sources).

We employed the GridSearchCV tool available in scikit-learn, which systematically evaluates all possible hyperparameter combinations within a user-defined grid through an exhaustive search. Although computationally expensive, since the model is trained and validated for every parameter combination, this approach ensures that the global optimum within the explored grid is identified. For completeness, scikit-learn also provides RandomizedSearchCV, which instead samples random subsets of the parameter space, but we opted for the exhaustive GridSearchCV strategy to guarantee robustness.

The following hyperparameter grid was explored:

- n_estimators = 100, 300, 500, 1000;
- criterion = gini, entropy;
- max_depth = None, 10, 50, 100;
- max_features = sqrt, log2, None;
- min_samples_split = 2, 5, 10;
- min_samples_leaf = 1, 2, 4;
- bootstrap = True, False

This corresponds to a total of 1728 distinct parameter combinations. We emphasize that the goal of this work is not to provide an exhaustive discussion of the role of each parameter, for which we refer the interested reader to the official RandomForestClassifier documentation[10].

[10] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

To reliably estimate the predictive performance of the classifier and to prevent overfitting to a single train/test split, GridSearchCV internally adopts the k-fold cross-validation technique. In this approach, the training set is randomly divided into k equally sized subsets (folds). At each iteration, one fold is retained for validation while the remaining k-1 folds are used for training. This process is repeated k times, ensuring that each fold serves as validation exactly once. The final performance metric is then computed as the average across all folds, thus providing a more robust and less biased estimate of the model's generalization ability compared to a single split. In this work, we adopted k = 5.

In the following sections, we present the optimized hyperparameters and the corresponding performance metrics for each of the input feature configurations explored in this study. This allows us to assess how different combinations of LSST features influence the classifier's ability to separate stars from galaxies across the full magnitude range.

3.2. Random Forest performance on the independent validation sample

In this section, we evaluate the performance of the RF classifier across multiple configurations of input features. Additional experiments, including further feature combinations, are presented in Appendix A.

3.2.1. The reference feature set: all LSST colours + refExtendedness

Our primary experiment adopts what we refer to as the reference feature set, composed of all possible LSST colours together with the morphological parameter refExtendedness. Among the 15 611 sources of our catalog, 14 360 objects, 4002 stars (28%) and 10 358 galaxies (72%), have complete photometry in all six LSST filters and a valid value of the refExtendedness parameter. The fine-tuning analysis described in Sect. 3.1 revealed that six distinct combinations of hyperparameters yielded the same maximum F1 score of 93.3%. All these models share the same core configuration, namely n_estimators = 300, criterion = entropy, min_samples_leaf = 2, min_samples_split = 2, and bootstrap = False. The only variations occur in the choice of max_features (either sqrt or log2), and max_depth (None, 50 or 100).

As already anticipated in the previous sections, to assess the robustness and generalization capability of the classifier, we evaluated its performance on the independent validation sample, consisting of 30% of the reference dataset (4308 objects), which was not used during training. We first consider the overall accuracy, defined as

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

which reaches 97.8%. This means that, on average, the classifier correctly identifies 98 out of 100 objects. However, accuracy alone can be misleading in the presence of class imbalance, and therefore we also examine precision, recall, and F1 score separately for stars and galaxies. Table 2 summarizes these metrics. Both classes exhibit high precision (~98%), indicating that when the classifier assigns an object to a given class, the classification is correct in nearly all cases. The recall is very high for galaxies (99.1%), while slightly lower for stars (94.3%), meaning that a small fraction of stars (about 6%) are misclassified as galaxies. This behaviour is expected given the class imbalance and

Table 2. Performance metrics for the RF classifier on the validation sample for a different combination of features.

Features	Class	Precision %	Recall %	F1 score %	N objects
Reference set	Galaxies	97.8	99.1	98.5	3109
	Stars	97.6	94.3	95.9	1199
	Accuracy = 97.8%				
No refExtendness	Galaxies	96.8	98.5	97.7	3109
	Stars	96.0	91.7	93.8	1199
	Accuracy = 96.6%				
Ref. Set + Phot. Err.	Galaxies	98.2	99.5	98.8	3109
	Stars	98.6	95.2	96.9	1199
	Accuracy = 98.3%				
No u-band	Galaxies	97.1	98.9	97.9	3271
	Stars	96.8	92.5	94.6	1291
	Accuracy = 97.0%				
No y-band	Galaxies	97.5	98.7	98.1	3202
	Stars	96.3	93.1	94.5	1182
	Accuracy = 97.2%				
Phot. Err. – No refExt.	Galaxies	97.3	99.0	98.1	3109
	Stars	97.2	92.9	95.0	1199
	Accuracy = 97.3%				

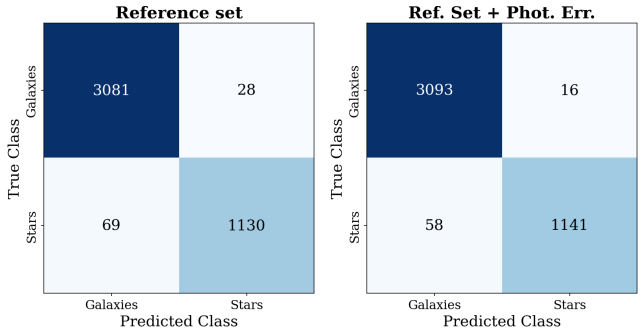

Fig. 3. Confusion matrix on the validation sample for the reference set (left matrix) and with the inclusion of photometric uncertainties (right matrix).

the intrinsic difficulty of distinguishing faint stars from compact galaxies. While looking for UFDs this issue will slightly reduce the discovery power (since a few stars of a real UFD may be missed due to misclassification) but will not enhance false detections, that are driven by galaxies misclassified as stars, not the opposite. These results are also reflected in the confusion matrix, displayed in the top panel of Figure 3, which shows that the model correctly classifies the vast majority of galaxies (3081 out of 3109) and stars (1130 out of 1199), with most of the misclassifications occurring in the form of stars incorrectly labeled as galaxies. Overall, the F1 score is high for both classes (98.5% for galaxies and 95.9% for stars), demonstrating that the classifier achieves a strong balance between completeness and purity. These results confirm that the RF model is robust and effective in separating stars from galaxies with LSST photometry.

In Figure 4, we compare the performance of our RF classifier (leftmost panel) with that of the REFEXTENDEDNESS parameter alone (left central panel), as a function of the r-band magnitude, using only objects in the validation sample. Since the leftmost panel is constructed in the same manner as Figure 2 but restricted to the validation set, the overall trends remain consistent, with fluctuations due to the smaller number of objects in some magnitude bins. To reduce the statistical noise at faint magnitudes, we adopted wider magnitude bins for $r \geq 24$ mag for stars and $r \geq 26$ mag for galaxies. As seen in the leftmost panel, the performance of REFEXTENDEDNESS for stars remains high down to $r \simeq 24$ mag, after which the stellar recovery fraction decreases and fluctuates around ∼60%. Conversely, the contamination from compact galaxies misclassified as stars remains below ∼20% down to $r \sim 26$ mag, but then increases sharply, exceeding ∼40% by $r \sim 27.5$ mag. This behaviour mirrors what was observed in the full sample (Figure 2), confirming that REFEXTENDEDNESS loses discriminating power at faint magnitudes.

The left central panel demonstrates the improvement introduced by the RF classifier adopting this set of features. At bright magnitudes ($r \lesssim 24$ mag), we can observe a slight improvement with respect to the REFEXTENDEDNESS classifier, as expected since most bright sources are clearly resolved. However, at fainter magnitudes ($r \geq 24$ mag), the RF substantially increases the stellar recovery fraction, as the latter stabilizes between 70% and 85%, compared to the ∼50%-75% obtained using REFEXTENDEDNESS alone. We note, however, that the performance at these faint magnitudes is affected by small-number statistics: the validation set contains only 31 stars with $r \geq 25$ mag and 16 with $r \geq 26$ mag.

For galaxies, the improvement is even more pronounced. The Random Forest keeps the fraction of galaxies misclassified as stars below ∼5% up to $r \sim 26.5$ mag, and only ∼14% at $r \sim 27.5$ mag. This is in stark contrast with REFEXTENDEDNESS, for which galaxy contamination rises rapidly beyond $r \sim 26$ mag. These results demonstrate that the multi-band colour information is the key factor enabling a significant reduction of galaxy contamination at faint magnitudes, with the Random Forest serving as an effective tool to exploit this information beyond what is achievable with morphology alone.

3.2.2. Removing morphology information: LSST colours only

We also evaluated the performance of the Random Forest classifier when the REFEXTENDEDNESS parameter is excluded from the input feature set, in order to assess the capability of the model to distinguish stars and galaxies using only LSST photometry. We repeated the optimization procedure described in Sect. 3.1 and found that six distinct combinations of hyperparameters yield the same maximum F1 score of 90.1%. All the best–performing models share the following configuration: n_estimators = 300, criterion = entropy, min_samples_leaf = 1, min_samples_split = 2, and bootstrap = True. The only differences among these top models lie in the choice of max_features (either sqrt or log2) and in the adopted max_depth, which can be None, 50, or 100. The corresponding metrics are reported in the second row of Table 2. The overall accuracy is slightly lower than in the case where REFEXTENDEDNESS is included, reaching 96.6%. The decrease primarily affects the stellar classification: while the metrics for the galaxy class degrade by less than ∼1%, those for the stars worsen by about 2%. In particular, the stellar recall drops to 91.7%, meaning that approximately 1 out of 10 stars is misclassified as a galaxy. The performance as a function of magnitude, shown in the right central panel of Figure 4, confirms these trends. Overall, the stellar recall is lower than in the configuration including REFEXTENDEDNESS,

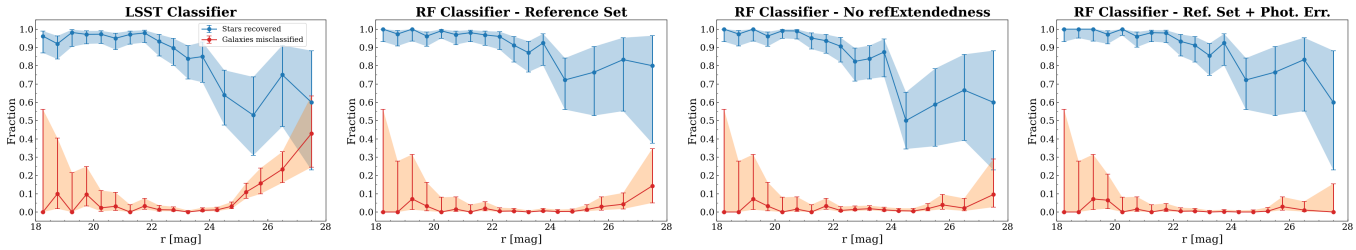

Fig. 4. Performance of the REFEXTENDEDNESS parameter (leftmost panel) and the RF classifier (left central panel), evaluated on the validation sample, as a function of the r-band magnitude, for the reference set. For both panels, the blue curves show the fraction of true stars correctly classified as stars (stellar completeness), while the red curves show the fraction of true galaxies misclassified as stars (galaxy contamination). The last two panels displays the performance of the RF classifier obtained by removing the REFEXTENDEDNESS parameter (right central panel) and by adding photometric uncertainties (rightmost panel) to the reference set.

and, in particular, the random forest without the morphological parameter does not provide a clear improvement over the internal LSST stellar classification alone in terms of star recovery. Conversely, the galaxy contamination fraction obtained with the random forest remains comparable between this experiment and the previous one. These results show that LSST multi-band colour information alone is already sufficient to keep the contamination from background galaxies at very low levels across the full magnitude range probed, including the faint end where morphological information becomes progressively ineffective. At the same time, the absence of the morphological parameter leads to a reduced recovery of stars, which becomes more evident at intermediate and faint magnitudes ($r \gtrsim 24$ mag). In this regime, REFEXTENDEDNESS still provides a measurable benefit in terms of stellar recall, even though its overall discriminating power is diminished. We note, however, that the stellar validation sample at these magnitudes is affected by low-number statistics. An expanded training and validation set with securely identified faint stars may therefore improve the stellar recall even in configurations that rely exclusively on multi-band photometric information. Interestingly, part of this loss in stellar recall can be mitigated when photometric uncertainties are explicitly included among the input features, even without the use of REFEXTENDEDNESS. The detailed results of this experiment are discussed in Appendix A.3.

3.2.3. Adding photometric uncertainties

In this third experiment, we augmented the reference feature set (all LSST colours + REFEXTENDEDNESS) by adding the photometric uncertainties in each filter. The motivation behind this test is to evaluate whether incorporating measurement errors can improve classification performance at faint magnitudes, where photometric noise becomes significant and sources of different classes tend to scatter into overlapping regions of colour-colour space (see also Fig. 6). A new hyperparameter optimization following the procedure described in Sect. 3.1 identified two distinct configurations achieving the same maximum F1 score of 93.5%, the highest obtained among the three experiments discussed so far. Both optimal models share the following parameters: n_estimators = 1000, criterion = gini, min_samples_leaf = 1, min_samples_split = 2, max_depth = 10, and bootstrap = True, differing only in the choice of max_features (either sqrt or log2).

The performance metrics reported in the third row of Table 2 confirm that this configuration yields the best overall results. The accuracy reaches 98.3%, outperforming both the reference model and all other feature combinations. While the improvement for galaxies is modest (a few tenths of a percent across precision, recall, and F1 score), the gain for stars is more substantial: stellar precision increases to almost 99%, recall to ≃95%, and F1 score to ≃97%. These improvements are also evident in the confusion matrix (bottom panel of Fig. 3), where the number of misclassified galaxies is reduced by almost a factor of two, and the number of misclassified stars decreases by ≃15% compared to the reference set. The magnitude-dependent performance (rightmost panel of Fig. 4) shows that galaxy contamination remains negligible across the entire magnitude range probed, while the stellar recall is comparable (within uncertainties) to that of the reference configuration. Overall, these results indicate that incorporating photometric uncertainties provides valuable additional information to the classifier, enabling it to better account for the scatter of faint sources away from their intrinsic loci in colour–colour space. This effect is particularly important for stars, whose locus is defined by narrow sequence (e.g. Fadely et al. 2012, and see also Fig. 6); poorly measured stars are more likely to deviate from this tight sequence, and including their uncertainties helps the classifier correctly recover them. Conversely, galaxies occupy a broader and more diffuse region of colour–colour space, so the benefit for this class is more limited, but still appreciable.

3.2.4. Key findings from the feature set tests

To summarize the results of these three experiments, our analysis shows that, for the present sample of bona fide stars and galaxies, LSST multi-band colours are sufficient to maintain negligible galaxy contamination across the entire magnitude range (e.g. $r \lesssim 27.5$ mag). While the inclusion of REFEXTENDEDNESS further enhances stellar recall, particularly at the faintest magnitudes, it offers no additional gain in reducing galaxy misclassification, as evidenced by the consistently low contamination levels even in the absence of morphological data (right central panel of Fig. 4). Finally, incorporating photometric uncertainties as input features enhances the overall performance of the random forest classifier, especially for stars, whose large colour errors can otherwise scatter them away from the narrow stellar locus in colour–colour diagrams.

We performed an additional test using the XGBoost algorithm (Chen & Guestrin 2016; Chen et al. 2026), which is widely adopted in classification tasks and is known for its computational efficiency. The performance metrics obtained with XGBoost are widely consistent with those of RF, with marginally lower metrics and a slightly higher contamination at the faint end. Given our primary scientific objective of minimizing galaxy contamination in stellar catalogues, we retain the RF classifier as our reference model. A detailed comparison is provided in Appendix B.

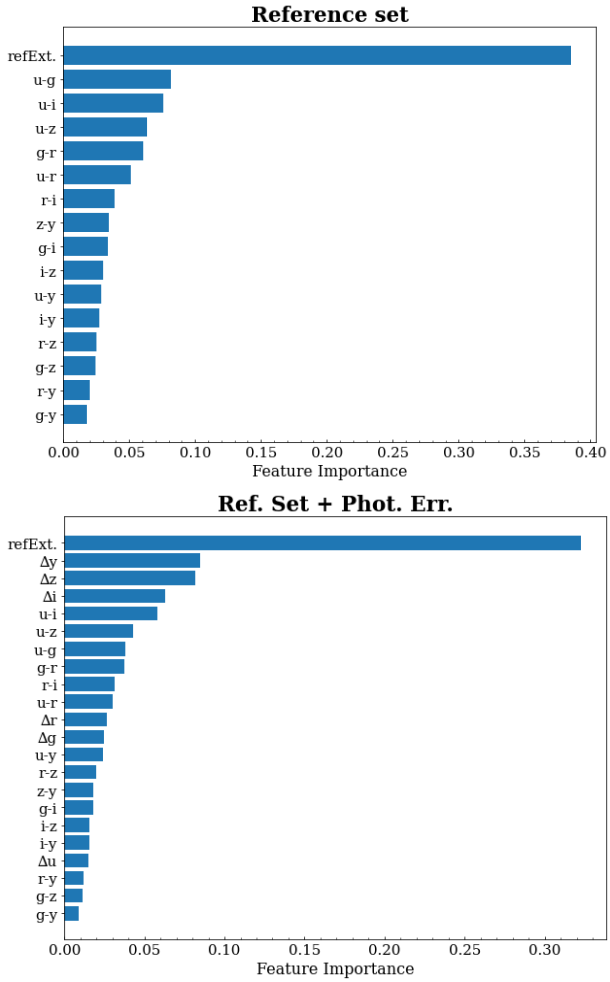

Fig. 5. Relative importance of each feature for the reference set (top panel) and with the inclusion of photometric uncertainties (bottom panel).

3.3. Feature importance

A key advantage of Random Forest classifiers is their ability to quantify the relative importance of the input features used during the classification process. We computed the feature importance values using the *feature_importances_* attribute provided by *scikit-learn*, which measures the average reduction in node impurity contributed by each feature across all decision trees in the ensemble. The top panel of Figure 5 displays the resulting ranking of features, from the most to the least relevant, by adopting the reference set of features. The REFEXTENDEDNESS parameter clearly dominates the feature space, contributing alone to nearly 40% of the total importance. This strong discriminating power is already evident at bright magnitudes (Figure 2), where morphological information efficiently separates stars from galaxies. The dominance of the morphological parameter in our feature-importance analysis is consistent with previous studies. For example, Jeakel et al. (2026) showed that the most informative predictors in their XGBoost model are the concentration index and the normalized peak surface brightness (see their Figure 17), while Khramtsov et al. (2019) found that the KiDS CLASS_STAR parameter overwhelmingly dominates the information content (their Figure 4). It is worth noting, however, that both works operate at substantially shallower depth than the LSST regime explored here. Indeed, in our experiments, as both Figure 2 and the leftmost panel of Figure 4 demonstrate, the effectiveness of REFEXTENDEDNESS alone decreases at intermediate and faint magnitudes. In this regime, where morphological information becomes progressively less reliable, multi-band colour information plays an increasingly dominant role. Indeed, at the faintest magnitudes probed in this work, the classifier achieves comparable, or even slightly improved, performance, in terms of galaxy classification, when REFEXTENDEDNESS is excluded from the feature set (see the right central panel of Figure 4). This indicates that LSST multi-band photometry alone already carries most of the discriminative power required to control galaxy contamination in the faint regime. It is important to note that the feature-importance analysis presented here is global, i.e. it does not explicitly account for magnitude-dependent effects. As a result, parameters that are highly effective at bright magnitudes, such as REFEXTENDEDNESS, naturally dominate the ranking, even though their discriminating power decreases toward the faint end. Consistently with this interpretation, we find that the reduced stellar recall observed at faint magnitudes when excluding REFEXTENDEDNESS can be largely recovered by incorporating photometric uncertainties into the feature set (see Appendix A.3), indicating that colour information, when properly weighted by its uncertainties, is sufficient to drive the classification in the low signal-to-noise regime.

Among the colour indices, the most relevant feature is $(u-g)$, which contributes almost 10% of the total importance. Other significant features include $(u-i)$, $(u-z)$, $(g-r)$, and $(u-r)$, each contributing more than 5%. The importance then gradually decreases for the remaining colour combinations, all below the 5%. Overall, the u-band emerges as a key player in the classification, being involved in nearly all the top-ranked colour indices. It is worth noting that this trend is broadly consistent with previous studies. For instance, Khramtsov et al. (2019) found that the most discriminating colour for their CatBoost classifier was the near-infrared index $(H-K_s)$, followed by $(u-g)$ as the second most informative feature. This result highlights two complementary points: (i) extending the wavelength coverage deeper into the near-infrared (beyond LSST's y band) can significantly enhance star–galaxy separation, suggesting that infrared imaging from missions such as Euclid (though lacking K_s, but including J and H) may provide valuable additional leverage; (ii) in the absence of such deep infrared information, the $(u-g)$ colour remains the strongest optical discriminant, fully consistent with what we observe in our feature set. Finally, apart from the presence of the z-filter in the third-ranked colour combination, the LSST infrared bands contribute relatively little to the classification task. This indicates that the combination of REFEXTENDEDNESS and u-band colours already captures most of the discriminative information needed, reducing the relative impact of redder bands in this particular feature set.

The bottom panel of Figure 5 displays the feature importance ranking obtained when photometric uncertainties are added to the reference feature set. Interestingly, immediately after REFEXTENDEDNESS which remains the single most informative feature, photometric uncertainties, particularly those in the y, z, and i bands, rank from second to fourth in importance. We interpret this result in light of the relative depths of the LSST filters in DP1: the z and y bands are shallower by approximately 1–3 magnitudes compared to g and r (i.e. $g = 26.18$ mag, $r = 25.96$ mag, $z = 25.07$ mag, $y = 23.10$ mag; NSF-DOE Vera C. Rubin Observatory 2025b). As a consequence, for colour combinations involving these redder bands, photometric uncertainties carry significant information that helps the classifier correctly assign sources to their respective classes, particularly for stars,

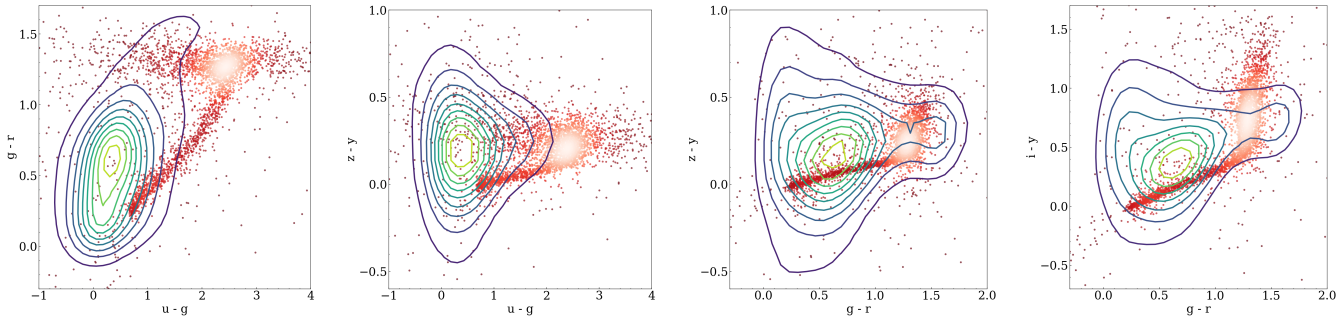

Fig. 6. Colour–colour diagram showing the distribution of stars and galaxies in our sample for a different combination of colour pairs. Stellar sources (in red) are displayed as individual points colour-coded by their local density. Galaxies are represented by isodensity contours (in viridis), highlighting the underlying structure of the galaxy locus in colour–colour space.

which occupy intrinsically narrow loci in colour-colour space (see Sect. 4). Consistent with this interpretation, the photometric uncertainties in the deeper g and r bands do not appear among the highest-ranked features. Interestingly, the uncertainty in the u band lies among the least informative features, despite the u band being shallower than z (i.e. $u = 24.55$ mag). This likely reflects the fact that colour–colour diagrams involving the u band already provide strong intrinsic separation between stars and galaxies (see Sect. 4), such that additional information from photometric uncertainties does not substantially improve the classification in this case. Indeed, even in this extended feature set, colours involving the u filter (e.g. $u - i$, $u - z$, and $u - g$) remain among the most discriminative photometric features, further highlighting the critical role of the u band in star–galaxy separation tasks at faint magnitudes.

4. Stellar and galaxy loci in LSST colour–colour diagrams

Figure 6 presents a set of colour–colour diagrams where the separation between the two populations becomes more evident. The leftmost panel shows $(u - g)$ versus $(g - r)$, the two most discriminant colour features of the classifier. In this diagram, nearly all stars (red points) occupy a well-defined region distinct from galaxies. Specifically, the galaxy distribution, represented by isodensity contours, is largely confined to $-0.5 \leq (u - g) \leq 1.0$ and $0 \leq (g - r) \leq 1.0$, whereas stars follow a narrow curved sequence extending from $(u - g, g - r) \simeq (0.7, 0.2)$ to $(2.4, 1.3)$, where thick- and thin-disk MW stars converge into a compact clump (see also Smolčić et al. 2004, for a similar plot made with ∼2 million stars from SDSS Data Release 1). The combination of $(u - g)$ with infrared colours also provides a clear separation. For instance, the left central panel shows $(u - g)$ versus $(z - y)$, where stars trace a narrow diagonal sequence from $(u - g, z - y) \simeq (0.7, 0.0)$ to $(2.4, 0.2)$, again terminating in a clump associated with thick- and thin-disk MW stars. Most of the galaxies, by contrast, are displaced to the left of this sequence. Similar diagrams can be obtained using $(u - r)$, or $u - i$, instead of $(u - g)$; in these cases, all sources (including galaxies) are shifted toward redder colours, with the MW stellar clump located at $u - r \simeq 3.7$ mag, or $u - i \simeq 4.7$ mag, respectively.

Even without the u filter, certain colour combinations could be more effective in separating stars from galaxies, particularly those involving optical–infrared pairs. The last two panels show $(g - r)$ versus $(z - y)$ (right central panel) and $(i - y)$ (rightmost panel). In both cases, the stellar locus forms a narrow, well-defined diagonal sequence, while the galaxy population occupies a broader region of the diagram. Nonetheless, despite its compactness, this stellar sequence partially overlaps the region where the galaxy density increases significantly, around $(g - r, z - y) \simeq (0.6, 0.1)$ and $(g - r, i - y) \simeq (0.6, 0.2)$. This overlap is particularly problematic because the same colour range $(g - r \lesssim 1.0$ mag) is where old, metal-poor stars characteristic of UFDs are found, hampering detection algorithms that rely on isolating these populations (e.g. Walsh et al. 2009). Moreover, the stellar clump associated with thick- and thin-disk MW stars itself intersects zones of the diagram that are populated, albeit at relatively low density, by compact galaxies, leading inevitably to some degree of contamination. These colour–colour diagrams therefore illustrate that in the absence of deeper near-infrared information (e.g. $J - K_s$, see Tortora et al. 2018, their Figure 1), which would more effectively separate stars and galaxies, optical–infrared LSST colours alone cannot fully disentangle the overlapping loci of stars and galaxies. Consequently, the inclusion of the u-band becomes particularly advantageous for enhancing star–galaxy separability.

As an additional robustness check, we verified that the machine-learning classification preserves the intrinsic stellar locus in colour–colour space by directly comparing the predicted stellar distribution with that defined by our sample of confirmed stars. The two loci show a good agreement. The details of this comparison are presented in Appendix C.

5. Summary

In this pilot study, we investigated the performance of supervised machine-learning techniques, specifically a Random Forest classifier, for separating stars from galaxies in deep, LSST data. Using observations of the ECDFS from LSST DP1, we constructed a heterogeneous sample of bona fide stars and galaxies that approaches, though does not fully reach, the expected 10-year LSST depth, and whose median seeing is worse than the survey's design goals. Consequently, the results presented here should be regarded as a conservative lower limit on the classification performance achievable with the final LSST dataset.

Our analysis demonstrates that incorporating LSST multi-band photometry substantially improves star–galaxy separation compared to using morphological information alone, especially at faint magnitudes, the regime in which LSST's discovery potential will be greatest. The study also highlights that while REFEXTENDEDNESS is the most informative parameter, its discriminating power rapidly declines toward the faint end. In contrast, colour information, particularly u-band colours, provides a robust classification even in the faint regime, when

morphological separation alone becomes ineffective. Moreover, providing full six-band LSST photometry is more effective for ensuring reliable star–galaxy separation than substituting a filter with the REFEXTENDEDNESS parameter, particularly when high stellar purity is a priority (see Appendix A). Finally, explicitly including photometric uncertainties as input features provides additional discriminative power, especially at faint magnitudes where increased photometric scatter causes sources to migrate across colour-colour loci. This leads to a measurable improvement in stellar classification, and a slight decreasing in galaxy contamination as well. This confirms that information from multi-band photometry is essential for reliable star–galaxy separation in LSST-like data.

A comparison with the XGBoost algorithm shows fully consistent results, with marginally lower performance. Specifically, it shows a slight increased contamination at the faint end, with more galaxies misclassified as stars at the last magnitude bins. Overall, our results demonstrate that supervised machine-learning methods can keep galaxy contamination at negligible levels down to very faint magnitudes, ensuring that searches for UFDs will not be significantly hindered by clustered misclassified background galaxies. A limitation instead might lie in stellar completeness, particularly due to the scarcity of well-characterized faint stars in current training sets. These findings provide a solid foundation for optimizing classification strategies in LSST and other multi-band surveys. Further performance gains will likely require access to larger, reliable training samples that adequately cover the faint end of the magnitude distribution, an observationally challenging but worthwhile goal. To support future work in this direction, we will publicly release the curated star–galaxy sample assembled for this study, enabling the community to extend, and refine classification models as LSST data quality and depth improve.

Data availability

The catalogue of bona-fide stars and galaxies built in this work is available at the CDS via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/709/A79

Acknowledgements. In this work, we made use of the following softwares: TOPCAT (Taylor 2011), Scikit-learn (Pedregosa et al. 2011), NumPy (Harris et al. 2020), pandas (Wes McKinney 2010; pandas development team 2020), matplotlib (Hunter 2007), and XGBoost (Chen & Guestrin 2016; Chen et al. 2026). We thank the anonymous referee for the constructive comments, which have significantly improved the quality of the manuscript. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This work is based on observations taken by the 3D-HST Treasury Program (GO 12177 and 12328) with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This research is supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI is provided by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico (CONACYT); the Ministry of Science and Innovation of Spain, and by the DESI Member Institutions. The DESI collaboration is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. We thank M. D'addona for his valuable insights on the random forest classifier. M.G. acknowledges "Partecipazione LSST – Large Synoptic Survey Telescope (ref. A. Fontana)" (Ob. Fu.: 1.05.03.06).

References

Bacon, R., Brinchmann, J., Conseil, S., et al. 2023, A&A, 670, A4
Balestra, I., Mainieri, V., Popesso, P., et al. 2010, A&A, 512, A12
Breiman, L. 2001, Mach. Learn., 45, 5
Cardamone, C. N., van Dokkum, P. G., Urry, C. M., et al. 2010, ApJS, 189, 270
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints [arXiv:1612.05560]
Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785
Chen, T., He, T., Benesty, M., et al. 2026, xgboost: Extreme Gradient Boosting, r package version 3.3.0.0
Choi, Y., Olsen, K. A. G., Carlin, J. L., et al. 2025, arXiv e-prints [arXiv:2507.01343]
Cooper, M. C., Yan, R., Dickinson, M., et al. 2012, MNRAS, 425, 2116
Damen, M., Labbé, I., van Dokkum, P. G., et al. 2011, ApJ, 727, 1
D'Eugenio, F., Cameron, A. J., Scholtz, J., et al. 2025, ApJS, 277, 4
Duncan, K. J. 2022, MNRAS, 512, 3662
Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15
Feng, H.-C., Li, R., Napolitano, N. R., et al. 2025, ApJS, 279, 26
Fotopoulou, S., & Paltani, S. 2018, A&A, 619, A14
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Vallenari, A., et al.) 2023, A&A, 674, A1
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jeakel, A. P., Vieira dos Santos, G., Marra, V., et al. 2026, The miniJPAS and J-NEP Surveys: Machine Learning for Star-Galaxy Separation
Khramtsov, V., Sergeyev, A., Spiniello, C., et al. 2019, A&A, 632, A56
Kodra, D., Andrews, B. H., Newman, J. A., et al. 2023, ApJ, 942, 36
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, 625, A2
Kurk, J., Cimatti, A., Daddi, E., et al. 2013, A&A, 549, A63
Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14
Lidman, C., Tucker, B. E., Davis, T. M., et al. 2020, MNRAS, 496, 19
Logan, C. H. A., & Fotopoulou, S. 2020, A&A, 633, A154
Maddox, N., Hewett, P. C., Warren, S. J., & Croom, S. M. 2008, MNRAS, 386, 1605
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27
Mutlu-Pakdil, B., Sand, D. J., Crnojević, D., et al. 2021, ApJ, 918, 88
NSF-DOE Vera C. Rubin Observatory 2025, Legacy Survey of Space and Time Data Preview 1 [Data set]
NSF-DOE Vera C. Rubin Observatory 2025a, Legacy Survey of Space and Time Data Preview 1: Object searchable catalog [Data set]
NSF-DOE Vera C. Rubin Observatory 2025b, The Vera C. Rubin Observatory Data Preview 1, Rubin Technical Note RTN-095, NSF-DOE Vera C. Rubin Observatory
pandas development team, T. 2020, pandas-dev/pandas: Pandas
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pennock, C. M., van Loon, J. T., Cioni, M.-R. L., et al. 2025, MNRAS, 537, 1028
Popesso, P., Dickinson, M., Nonino, M., et al. 2009, A&A, 494, 443
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Simon, J. D. 2019, ARA&A, 57, 375
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, ApJS, 214, 24
Smolčić, V., Ivezić, Ž., Knapp, G. R., et al. 2004, ApJ, 615, L141
Talia, M., Schreiber, C., Garilli, B., et al. 2023, A&A, 678, A25
Taylor, M. 2011, TOPCAT: Tool for OPerations on Catalogues And Tables, Astrophysics Source Code Library [record ascl:1101.010]
Tortora, C., Napolitano, N. R., Spavone, M., et al. 2018, MNRAS, 481, 4728
Vanzella, E., Cristiani, S., Dickinson, M., et al. 2008, A&A, 478, 83
Walsh, S. M., Willman, B., & Jerjen, H. 2009, AJ, 137, 450
Wes McKinney 2010, in Proceedings of the 9th Python in Science Conference, eds. Stéfan van der Walt, & Jarrod Millman, 56
Wolf, C., Meisenheimer, K., Kleinheinrich, M., et al. 2004, A&A, 421, 913
Wright, A. H., Kuijken, K., Hildebrandt, H., et al. 2024, A&A, 686, A170
Zhang, S., Hildebrandt, H., Yan, Z., et al. 2025, A&A, 698, A108

Appendix A: Random Forest performance with alternative photometric feature sets

In this section, we present a complementary suite of Random Forest experiments in which the classifier is trained and evaluated by excluding individual filters to quantify the robustness of our main results against changes in the available photometric information.

Appendix A.1: Random Forest performance without the u band

Our first experiment consisted of removing the *u*-band from the reference input feature. In this configuration, the sample increases to 15 207 sources (about 6% more than the 14 360 objects with complete six-band coverage). As before, we allocate 70% of the sample to the training set and the remaining 30% to the validation set. We repeated the hyperparameter optimization described in Sect. 3.1, maximizing the F1 score on the validation sample. Two distinct hyperparameter combinations achieved the best F1 score of 92.9%, slightly below the value obtained with the reference feature set (93.3%) but above the score obtained when removing the morphological parameter (90.1%). The two optimal configurations share the following parameters: bootstrap = False, criterion = gini, max_depth = 10, min_samples_leaf = 2, min_samples_split = 5, n_estimators = 100, while the only varying hyperparameter is max_features (sqrt or log2).

The complete set of performance metrics is reported in Table 2. Overall, these metrics lie between those of the reference feature set and those obtained without REFEXTENDEDNESS, suggesting that omitting the *u*-band is less disadvantageous than omitting the morphological information, at least when evaluated through global metrics alone. However, Figure A.1 reveals a more sophisticated picture. As in the leftmost panel of Fig. 4, the upper panel illustrates the classification capability of REFEXTENDEDNESS alone. Because of the slightly larger sample, this plot differs marginally from the one presented earlier, although the global trends remain unchanged. The morphological classifier has a good performance in identifying stars down to $r \lesssim 23.5$ mag, beyond which its discriminating power steadily declines, reaching a stellar recovery fraction of only ∼50% in the faintest bins. For galaxies, REFEXTENDEDNESS remains highly reliable down to $r \lesssim 25$ mag, after which the misclassification rate rapidly increases, consistent with the behaviour seen in Fig. 4. The application of the Random Forest improves upon the performance of the morphological classifier alone. Nevertheless, in the absence of the *u* band, the contamination from misclassified galaxies at faint magnitudes is higher than in any experiment involving the full set of LSST colours, whether or not REFEXTENDEDNESS is included. In the last magnitude bin, the contamination from galaxies rises to ∼20%.

This result suggests that the *u*-band is essential for suppressing galaxy contamination in the faint regime where LSST will achieve its full scientific reach. Although the global metrics might suggest that removing REFEXTENDEDNESS is more disadvantageous than removing the *u* band, the magnitude dependent behaviour reveals the opposite: for applications such as UFD searches, where faint-end purity is critical, ultraviolet information outperforms morphological information. In other words, retaining the *u* band is more important than retaining the REFEXTENDEDNESS parameter for preserving classification quality at the faint photometric limits. Finally, we note that in the faintest two magnitude bins, the stellar recovery fraction increases. This behaviour likely reflects the very small number of stars in these bins (46 stars for $r \geq 25$ mag and 21 stars for $r \geq 26$ mag), combined with the model's tendency to classify sources in regions of strong colour degeneracy as stars. This effect enhances stellar recall but simultaneously inflates galaxy misclassification. This experiment reinforce the conclusion that the *u* band provides the most critical information for maintaining effective star–galaxy separation at faint magnitudes.

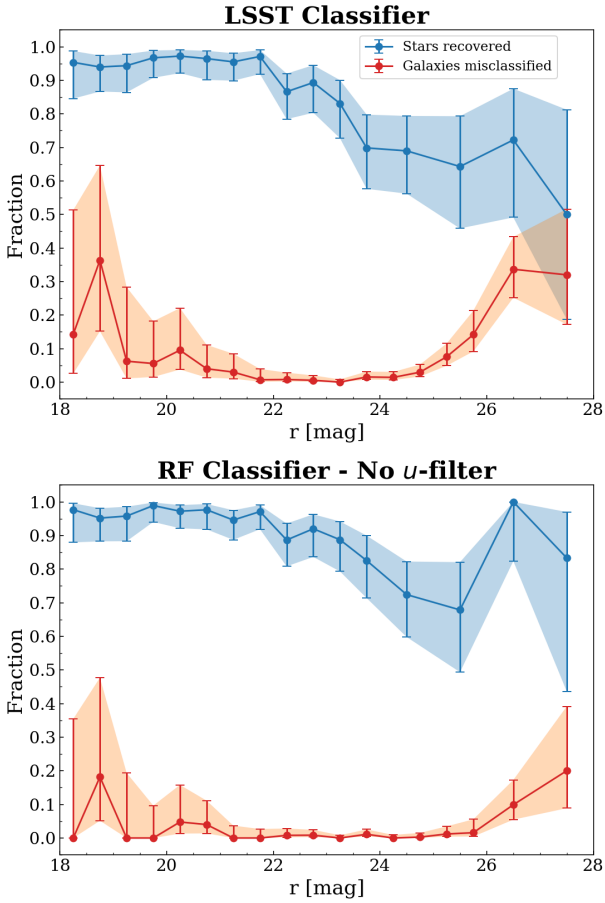

Fig. A.1. Same as Figure 4, but computed on the validation sample for the experiment in which the *u* band is excluded from the input feature set.

Appendix A.2: Random Forest performance without the y band

In this experiment, we evaluate the performance of the Random Forest classifier when the *y*-band is removed from the reference feature set. The resulting sample contains 14 612 sources. As in the previous tests, we performed a full hyperparameter optimization following the procedure described in Sect. 3.1. Six distinct hyperparameter configurations achieved the highest F1 score of 93.4%, a value marginally higher than that obtained with the reference feature set (93.3%). All optimal configurations share the following parameters: bootstrap = False, criterion = gini, min_samples_leaf = 2, min_samples_split = 10, n_estimators = 300. The only hyperparameters that vary across the best-performing models are max_depth (None, 50, or 100) and max_features (sqrt or log2).

The performance metrics reported in Table 2 show that excluding the *y* band results in comparable, slight better performance than removing the *u* band. Figure A.3 provides additional

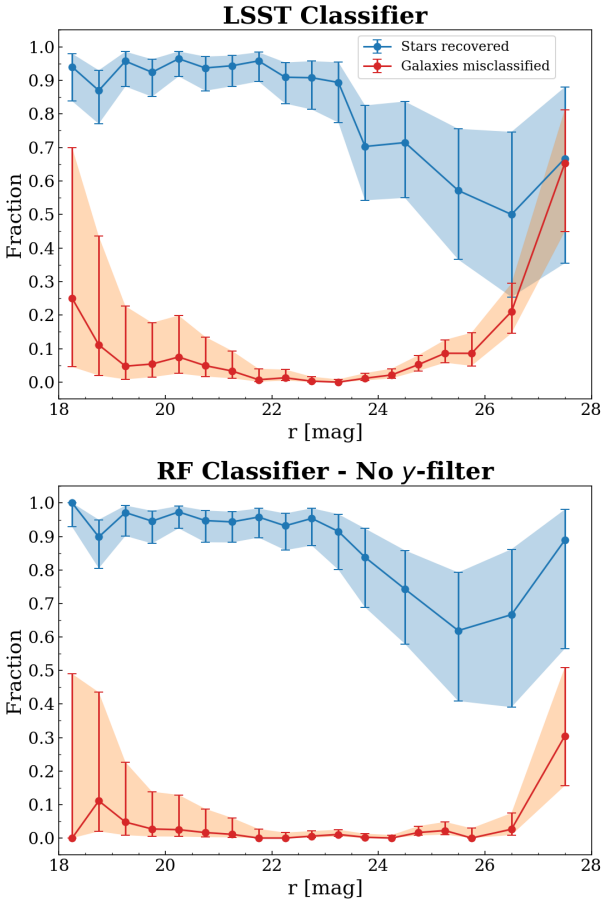

Fig. A.2. Same as Figure 4, but computed on the validation sample for the experiment in which the y band is excluded from the input feature set.

insight into the behaviour of the classifier as a function of magnitude. Overall, the fraction of misclassified galaxies shows a slight improvement compared to the case where the u-filter was excluded from the feature set, except for the faintest bin, where the contamination level reaches ∼30%. This experiment reinforces the conclusion that having access to the full six-band LSST photometry is more important for reliable star–galaxy classification than including the REFEXTENDEDNESS parameter at the expense of one filter, when high stellar purity is required. In particular, losing even a single band degrades performance more severely than omitting the morphological parameter, augmenting the risk that compact clusters or group galaxies are misidentified as UFDs.

Appendix A.3: Impact of photometric uncertainties in the absence of morphological features

In our last experiment, the classifier is trained using LSST multiband photometry and the corresponding photometric uncertainties, while explicitly excluding the REFEXTENDEDNESS parameter from the feature set. The goal of this test is to assess whether photometric uncertainties alone can compensate for the lack of morphological information, particularly in the faint regime where morphology becomes unreliable.

The hyperparameter optimization (see Sect. 3.1) identified six distinct configurations that achieved the maximum F1 score of 90.1%. All optimal configurations share the same core set of hyperparameters, namely: bootstrap = False, criterion = entropy,

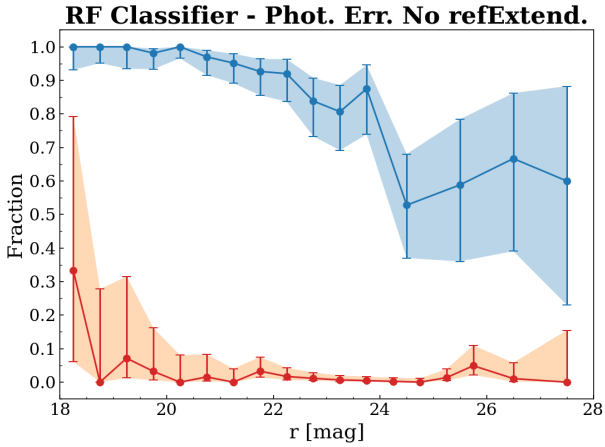

Fig. A.3. Same as Figure 4, but computed on the validation sample for the experiment in which the REFEXTENDEDNESS parameter is excluded and photometric uncertainties are included among the input features.

min_samples_leaf = 1, min_samples_split = 2, n_estimators = 300. The only hyperparameters that vary among these best-performing configurations are max_depth (None, 50, or 100) and max_features (sqrt or log2).

The performance metrics reported in Table 2 show that, even in the absence of the morphological parameter, the inclusion of photometric uncertainties yields results only slightly inferior to those obtained with the reference input feature set. For instance, the accuracy reaches 97.3%, compared to 97.8% for the reference configuration. Notably, relative to the experiment in which REFEXTENDEDNESS was excluded and photometric uncertainties were not included, the metrics associated with the stellar class improve substantially. In particular, the stellar recall increases from 91.7% to 92.9%, approaching the value of 94.3% obtained with the reference feature set.

Figure A.3[11] reveals that the fraction of misclassified galaxies even shows a slight improvement at the faintest magnitudes compared to the reference featuere set (left central panel of Fig. 4). This behaviour indicates that the morphological parameter does not carry critical information for the correct classification of compact galaxies at faint magnitudes.

When comparing the two configurations that exclude the morphological parameter, the inclusion of photometric uncertainties leads to a marginally more stable recovery of stars at intermediate and faint magnitudes. Although the differences are modest and largely consistent within the uncertainties, this trend suggests that photometric errors help the classifier to better accommodate stars that are scattered away from their intrinsic, narrow colour-colour loci. In this context, the reduced stellar recall observed at faint magnitudes in the right central panel of Fig. 4 appears to be driven primarily by increasing photometric scatter rather than by the absence of morphological information, which can be slightly mitigated by incorporating photometric uncertainties into the input features.

Finally, we note that the photometric precision and depth achieved in DP1 is lower than that expected for the final LSST survey (see Sect. 1). Consequently, future LSST data, benefiting from a deeper and more precise photometry, are likely to yield an even higher stellar recall at faint magnitudes, further mitigating the limitations observed in the present analysis.

[11] Note that we did not display the figure for the LSST internal classifier in this case as it would be identical to that displayed in the leftmost panel of Fig. 4.

Table B.1. Performance metrics for the XGBoost classifier on the validation sample for the reference feature set.

Features	Class	Precision %	Recall %	F1 score %	N objects
Reference Set	Galaxies	98.1	98.7	98.4	3109
	Stars	96.6	95.0	95.8	1199
Accuracy = 97.7%					

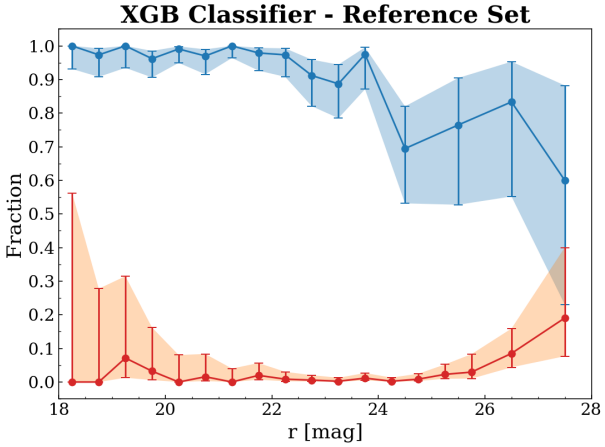

Fig. B.1. Same as Figure 4, but computed with the XGBoost classifier.

Appendix B: Comparison with XGBoost

In this section, we describe an additional experiment using the XGBoost algorithm (Chen & Guestrin 2016; Chen et al. 2026). For a direct comparison with the Random Forest, we restricted this test to the reference feature set (i.e. all LSST colour combinations together with the REFEXTENDEDNESS parameter). We optimized the hyperparameters following the same procedure described in Sect. 3.1. The explored grid included:

- n_estimators = 100, 300, 500, 1000;
- max_depth = 3, 5, 7;
- min_child_weight = 1, 5;
- learning_rate = 0.01, 0.05, 0.1;
- subsample = 0.8, 1.0;
- colsample_bytree = 0.8, 1.0;
- gamma = 0, 0.1, 0.3.

This corresponds to 864 distinct hyperparameter configurations. The best-performing model corresponds to the configuration: colsample_bytree = 1.0, gamma = 0, learning_rate = 0.05, max_depth = 7, min_child_weight = 1, n_estimators = 300, subsample = 0.8.

The resulting performance metrics on the validation sample are reported in Table B.1. The overall accuracy is 97.7%, namely 0.1% lower than that obtained with Random Forest. For galaxies, XGBoost yields a slightly higher precision (+0.3%) but a marginally lower recall (-0.4%). For stars, the recall increases by 0.7%, but the precision decreases by 1.0%, implying a slightly higher contamination of galaxies within the stellar sample. For both classes, the F1 score is lower by 0.1% compared to Random Forest. Figure B.1 shows the magnitude-dependent behaviour. At faint magnitudes ($r \gtrsim 26$ mag), XGBoost exhibits a slightly earlier rise in galaxy contamination compared to Random Forest, reaching approximately 20% in the last magnitude bin. The stellar recovery fraction also decreases to about 60% in the faintest bin.

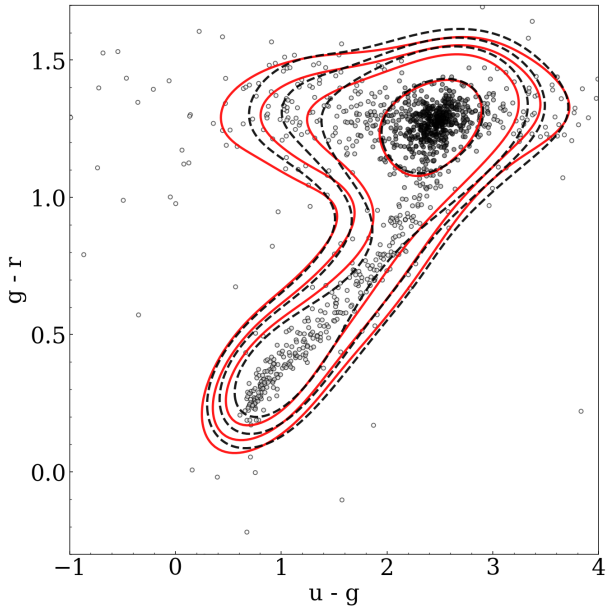

Fig. C.1. $(u-g)$ versus $(g-r)$ colour-colour diagram comparing objects classified as stars by the Random Forest model with bona-fide stars in our catalog. Coloured points show the predicted stellar sample in the validation set. Black dashed contours represent the isodensity distribution of the predicted stars, while red solid contours correspond to the bona-fide stellar sample from the full catalogue.

Overall, XGBoost delivers performance fully consistent with Random Forest, with marginally lower global metrics and slightly higher contamination at the faint end. Given our primary goal of minimizing galaxy contamination in deep stellar catalogues, we adopted Random Forest as the reference classifier throughout this work. We note, however, that XGBoost is significantly more computationally efficient: the hyperparameter search was approximately 45 times faster than for Random Forest. For larger datasets or more extensive hyperparameter grids, XGBoost may therefore represent a competitive alternative.

Appendix C: Comparison of predicted and true stellar locus

To further assess the robustness of the Random Forest, we compared the colour-colour distribution of objects classified as stars with the locus defined by our sample of bona-fide stars. Figure C.1 shows the $(u-g)$ versus $(g-r)$ diagram. The coloured points represent objects classified as stars by the Random Forest model in the validation sample, while the black dashed contours indicate their isodensity distribution. The red solid contours correspond to the isodensity distribution of bona-fide stars from the full catalogue. The predicted stellar sample closely follows the well-defined stellar locus over the entire colour range. In particular, the high-density core of the classified stars overlaps remarkably well with that of the true stars, demonstrating that the classifier preserves the intrinsic structure of the stellar population in colour-colour space. At lower densities, the agreement remains very good. Only a slight difference is visible in the bluest $(u-g)$ wing, where the predicted sample appears marginally less extended than the true one.