Star-galaxy classification in the Dark Energy Survey Y1 dataset

I. Sevilla-Noarbe[1]∗, B. Hoyle[2,3], M.J. Marchã[4], M.T. Soumagnac[5], K. Bechtol[6],
A. Drlica-Wagner[7], F. Abdalla[4,8], J. Aleksić[9], C. Avestruz[10], E. Balbinot[11],
M. Banerji[12,13], E. Bertin[14,15], C. Bonnett[9], R. Brunner[16], M. Carrasco-Kind[17], A. Choi[18],
T. Giannantonio[12,13,2], E. Kim[17], O. Lahav[4], B. Moraes[4], B. Nord[7], A.J. Ross[18],
E.S. Rykoff[19,20], B. Santiago[21,22], E. Sheldon[23], K. Wei[10,24], W. Wester[7], B. Yanny[7],
T. Abbott[25], S. Allam[7], D. Brooks[4], A. Carnero-Rosell[22,26], J. Carretero[9], C. Cunha[19],
L. da Costa[22,26], C. Davis[19], J. de Vicente[1], S. Desai[27], P. Doel[4], E. Fernandez[9],
B. Flaugher[7], J. Frieman[7,10], J. Garcia-Bellido[28], E. Gaztanaga[29,30], D. Gruen[19,20],
R. Gruendl[16,17], J. Gschwend[22,26], G. Gutierrez[7], D.L. Hollowood[31], K. Honscheid[18,32],
D. James[33], T. Jeltema[31], D. Kirk[4], E. Krause[34,35], K. Kuehn[36], T. S. Li[7,10], M. Lima[37,22],
M. A. G. Maia[22,26], M. March[38], R. G. McMahon[4,8], F. Menanteau[16,17], R. Miquel[39,9],
R. L. C. Ogando[22,26], A. A. Plazas[35], E. Sanchez[1], V. Scarpine[7], R. Schindler[20],
M. Schubnell[40], M. Smith[41], R. C. Smith[25], M. Soares-Santos[42], F. Sobreira[43,22],
E. Suchyta[44], M. E. C. Swanson[17], G. Tarle[40], D. Thomas[45], D. L. Tucker[7], A. R. Walker[25]

(The DES Collaboration)

Author affiliations are listed at the end of this paper.

ABSTRACT

We perform a comparison of different approaches to star-galaxy classification using the broad-band photometric data from Year 1 of the Dark Energy Survey. This is done by performing a wide range of tests with and without external 'truth' information, which can be ported to other similar datasets. We make a broad evaluation of the performance of the classifiers in two science cases with DES data that are most affected by this systematic effect: large-scale structure and Milky Way studies. In general, even though the default morphological classifiers used for DES Y1 cosmology studies are sufficient to maintain a low level of systematic contamination from stellar mis-classification, contamination can be reduced to the O(1%) level by using multi-epoch and infrared information from external datasets. For Milky Way studies the stellar sample can be augmented by ∼ 20% for a given flux limit. Reference catalogues used in this work are available at `http://des.ncsa.illinois.edu/releases/y1a1`.

Key words: Techniques: photometric – Methods: statistical – Methods: data analysis

1 INTRODUCTION

Accurate classification of astrophysical sources is essential for interpreting photometric surveys. Specifically, separating foreground stars from background galaxies is important for many astronomical research topics, from Galactic science to cosmology. Conventional morphological classification techniques separate point sources (mostly stars) from resolved sources (galaxies) using selections in magnitude-radius space or similar variables (MacGillivray et al. 1976; Kron 1980; Heydon-Dumbleton et al. 1989; Yee 1991).

∗E-mail: ignacio.sevilla@ciemat.es

For bright sources, morphology has proven to be a sufficient metric for classification. In this regime, for weak lensing applications, a very pure, but also abundant, star sample is vital for deriving the correct point spread function in the images which is used to later infer cosmic shear (Soumagnac et al. 2015; Jarvis et al. 2016; Zuntz et al. 2017). At fainter magnitudes unresolved galaxies will begin to contaminate catalogues of point-like sources and noisy measurements of stars will contaminate the galaxy sample. Blended sources become an issue as well, because distant and/or faint sources start to merge into single detected objects with spurious shapes. Misclassification of stars and galaxies at faint magnitudes can introduce spurious correlations in galaxy surveys (Ross et al. 2011) and will hamper the study of stellar distributions (Drlica-Wagner et al. 2015).

The advent of CCD detectors provided larger, more reliable data sets which became an obvious target for machine learning classification algorithms (e.g., Odewahn et al. 1992; Bertin & Arnouts 1996; Sevilla-Noarbe & Etayo-Sotos 2015; Machado et al. 2016; Kim & Brunner 2017). In addition, many large, multi-band imaging surveys use morphology (such as for SDSS, Stoughton et al. 2002) and/or have incorporated colour information into their classifiers (see Ball et al. 2006 for SDSS as well, Hildebrandt et al. 2012 for CFHTLS or Saglia et al. 2012 for Pan-STARRS). Adopting a Bayesian approach to incorporate fits to stellar and galaxy templates has been shown to be a promising avenue (Fadely et al. 2012), as well as the use of infrared data to complement the optical band observations (Małek et al. 2013; Kovács & Szapudi 2015; Banerji et al. 2015).

In this paper we test different strategies for classifying objects as point-like or extended sources in the Dark Energy Survey (DES) Year 1 data (Y1). We subsequently analyze the impact in two broad science cases, and possible developments to improve object classification in future analyses of this data. Throughout this paper, 'extended' will be used as a synonym for 'galaxy' whereas 'point-like' includes both stars and quasi-stellar objects (QSOs) on first approximation and we will collectively call them 'stars' in this work. For the case studies considered here and the general catalogue, the contamination of QSOs in the large-scale stellar and galactic catalogues is not deemed important. However, a good star-QSO separation is needed for quasar science, as studied in detail in Tie et al. (2017) for DES data.

After a description of the dataset in Section 2 and the classifiers we are considering here in Section 3, we compare the classifiers in calibration fields (Section 4) and then analyze the response in the complete Y1 dataset for a few selected ones (Section 5). Then we study the impact on large-scale structure and Milky Way studies (Section 6). Finally, Section 7 presents the conclusions and discusses possible additional developments.

2   DARK ENERGY SURVEY DATASETS

The DES consists of a 5000 square-degree "wide" survey using the *grizY* photometric bands to AB 10σ magnitude limits of (24.6, 24.4, 23.7, 22.7, 21.5) respectively for 2 arc-second apertures, together with a ~27 square degree supernovae survey observed in the *griz* bands with an approximately weekly cadence. In February 2018, the project completed the original five planned observing seasons (Years 1 through 5, Y1-Y5). Additional science-quality data was collected during an earlier Science Verification (SV) season. The core goal of DES is a multi-probe study of dark energy at different cosmological epochs using the same DECam instrument (Flaugher et al. 2015) and DES Data Management (DESDM) pipeline (Mor-

ganson et al. 2018), as showcased with its first results in DES Collaboration et al. (2017). However, the richness of this dataset allows astronomers and cosmologists to go beyond this initial objective (DES Collaboration et al. 2016).

For this study, we use the subset of highest quality data from DES SV[1] and Y1[2](Drlica-Wagner et al. 2018) comprising the "Gold" catalogue. We note the following features that are relevant for the present study:

- The object catalogues are obtained applying `SExtractor` (Bertin & Arnouts 1996) to coadded images with typically 2 to 4 overlapping exposures in each band in the case of Y1 or ~ 10 for SV data, with object detection performed on a combined *riz* image.
- `SExtractor` magnitudes have been calibrated through a global calibration module (Tucker et al. 2007) and subsequently adjusted through a fit to the stellar locus (High et al. 2009) anchored to the *i* band[3]. This procedure also corrects for Galactic extinction. In general, `MAG_AUTO` is used for photometry (for binning purposes and as inputs for the template based method described below), as it behaves more robustly for these coadded catalogues. `MAG_MODEL`, `MAG_DETMODEL`[4] and `MAG_PSF` are used as inputs for the machine learning methods as well. Shape measurements in this code include `FLUX_RADIUS`, `CLASS_STAR` and `SPREAD_MODEL`, some of which will be specifically studied here.
- In addition, a multi-object, multi-epoch fitting pipeline (`MOF`) has been run on the single-epoch image counterparts for each coadd catalogue detection to obtain more precise photometric measurements for the objects. It simultaneously fits a Gaussian mixture model to the individual images, also modelling light from nearby neighbours for each object (more details in Drlica-Wagner et al. 2018). The main flux measurements used for the methods described here are the fluxes using this composite Gaussian mixture model (`CM_MAG`) and the PSF magnitudes derived from the same `MOF` pipeline (`PSF_MAG`). `CM_T` is a size estimator from the code before PSF convolution, which will be studied here in detail.
- All objects are required to be in areas for which there is at least one exposure in each of the *griz* bands.

We define two distinct regions in which we will perform our tests:

(i) A **calibration field**: defined by those areas that overlap external datasets that we can use to train, validate and test our methods. These are the supernova (SN) fields from the DES SN survey, which overlap specific spectroscopic surveys and miscellaneous Hubble Space Telescope (HST) datasets; and the area of the survey overlapping the Sloan Digital Sky Survey (SDSS; York et al. 2000) Stripe 82 region (Frieman et al. 2008). In addition, the COSMOS field[5] has been imaged with DECam, providing a very useful dataset given the richness of multi-band imaging and spectroscopy available. Table 1 summarises the numbers of objects matched to various external datasets (details in Section 4.3). Some of these fields have a large number of DES exposures, due to their application for SN searches,

---

so special coadds were made from ∼ 4 exposures in each band in order to resemble the Y1 depth. The selection of these exposures was made so that their coaddition would provide similar characteristics in terms of sky brightness and seeing as the wide survey coadds (Drlica-Wagner et al. 2018; Neilsen et al. 2016). This procedure is not needed in forthcoming releases as the wide survey extends to cover the supernova regions.

(ii) An **application field**: the remaining area of the DES footprint for which suitable external datasets for training are not presently available. This includes the so-called 'SPT' region due to the overlap with the South Pole Telescope[6] (Ruhl et al. 2004) observations, in which we can make some quality assessment as well, though limited by the lack of external references.

3 DESCRIPTION OF THE OBJECT CLASSIFIERS

Table 2 summarises the methods explored in this paper to perform object classification. These include a variety of algorithms using machine learning methods (training on morphological and/or colour information), pixel-level flux measurements and template-fitting. For the sake of clarity and conciseness, not all algorithms are subjected to every test in this paper, but usually a selection is made in each case. Additional details and references are given below:

3.1 CLASS_STAR

This is the standard `SExtractor` star-galaxy classifier, providing a neural network real number output (a 'stellarity' index from 0 to 1) based on the training on a large simulation of galaxy and star images on CCDs.

Input data: For every object, eight isophotal areas above the background are measured, plus the value of the intensity at the peak pixel in the object and the value of the FWHM for the image.

Method: It uses a backpropagation model (Werbos 1982) for learning, based on simulations that include a wide range of PSF profiles and sizes, though they are optimised to work best on intermediate magnitude ranges (in the DES magnitude scale) of $V \sim 18 - 22$ due to the types of galaxies simulated and relative star-galaxy mixture.

3.2 SPREAD_MODEL

This quantity is a linear discriminant-based algorithm available with the `SExtractor` package. The `SPREAD_MODEL` estimator was originally developed as a star-galaxy classifier for the DESDM pipeline, and has also been used in other surveys (e.g., Desai et al. 2012; Bouy et al. 2013).

Input data: the image data at pixel level is used for each detected object in `SExtractor`.

Method: `SPREAD_MODEL` indicates which of the best fitting local PSF model $\vec{\phi}$ (representing a point source) or a slightly more extended model $\vec{G}$ (representing a galaxy) better matches the image data. $\vec{G}$ is obtained by convolving the local PSF model with a circular exponential model with scale length = 1/16 FWHM (Full-Width at Half-Maximum). `SPREAD_MODEL` is normalised to allow comparing sources with different PSFs throughout the field:

$$\text{SPREAD\_MODEL} = \frac{\vec{G}^T \mathbf{W} \vec{p}}{\vec{\phi}^T \mathbf{W} \vec{p}} - \frac{\vec{G}^T \mathbf{W} \vec{\phi}}{\vec{\phi}^T \mathbf{W} \vec{\phi}}, \tag{1}$$

where $\vec{p}$ is the image vector centered on the source[7]. $\mathbf{W}$ is a weight matrix constant along the diagonal except for bad pixels where the weight is 0. By construction, `SPREAD_MODEL` is close to zero for point sources, positive for extended sources (galaxies), and negative for detections smaller than the PSF, such as cosmic ray hits. The RMS error on `SPREAD_MODEL` is estimated by propagating the uncertainties on individual pixel values:

$$\begin{aligned} \text{SPREADERR\_MODEL} = \quad & \frac{1}{(\vec{\phi}^T \mathbf{W} \vec{p})^2} \Big(\vec{G}^T \mathbf{V} \vec{G} \, (\vec{\phi}^T \mathbf{W} \vec{p})^2 \\ & + \vec{\phi}^T \mathbf{V} \vec{\phi} \, (\vec{G}^T \mathbf{W} \vec{p})^2 \\ & - 2\vec{G}^T \mathbf{V} \vec{\phi} \, (\vec{G}^T \mathbf{W} \vec{p} \, \vec{\phi}^T \mathbf{W} \vec{p}) \Big)^{1/2} \end{aligned} \tag{2}$$

where $\mathbf{V}$ is the noise covariance matrix, which is assumed to be diagonal.

An example of a classifier derived from `SPREAD_MODEL` is the default classification scheme (`MODEST_CLASS`) used in the Y1 Gold catalogue, which includes the following criteria:

galaxies ⟺

> `SPREAD_MODEL_I+`
> $(5/3) \times$ `SPREADERR_MODEL_I` > 0.005
> `AND NOT` (3)
> (`|WAVG_SPREAD_MODEL_I|` < 0.002
> `AND`
> `MAG_AUTO_I` < 21.5)

stars ⟺

> `|SPREAD_MODEL_I+` (4)
> $(5/3) \times$ `SPREADERR_MODEL_I|` < 0.002

where `WAVG_SPREAD_MODEL` has been computed from a weighted average of the `SPREAD_MODEL` values of single-epoch shapes corresponding to that coadd object. These provide a better separation (DES Collaboration 2018) with respect to the standard `SPREAD_MODEL` on coadd images, albeit with a limited depth reach, as not all coadd objects have single epoch detections from which a weighted averaged can be computed (a faint object could be detected *only* in the coadded image and not in the individual epochs contributing to the image). The weights come from the weight map of the Data Management processing outputs and the band chosen is the *i* band where the images have a higher signal to noise, and have also demonstrated best performance in detailed simulations. Objects which do not fall into the categories expressed by Equations 3 and 4 are grouped into either a 'fringe' category between

---

[6] https://pole.uchicago.edu/

[7] This definition of `SPREAD_MODEL` differs from the one given in previous papers (Desai et al. 2012; Bouy et al. 2013), which was incorrect. In practice both estimators give very similar results.

Table 1. External datasets used in this work. SDSS-stripe 82 data shows two numbers according to simultaneous 2MASS and WISE matches, and VHS matches. More details are provided in Appendix B.

Catalogue	Type	Usage in this work	Nb. matched objects	Reference
ACS-COSMOS	Space optical imaging	Truth table	116017	Leauthaud et al. (2007)
Hubble-SC	Space optical imaging	Truth table	12927	Whitmore et al. (2016)
SDSS-stripe 82	Ground optical spectroscopy	Truth table	18984/46700	Albareti et al. (2017)
VVDS	Ground optical spectroscopy	Truth table	4442	Le Fèvre et al. (2013)
WISE	Space NIR imaging	Complementary data	18984	Wright et al. (2010)
2MASS	Ground NIR imaging	Complementary data	18984	Skrutskie et al. (2006)
VHS	Ground NIR imaging	Complementary data	46700	McMahon et al. (2013)

Table 2. Summary of classification methods. Type of data denotes whether measurements or direct pixel data are used, and in the first case if it is based on morphological and/or flux measurements. The specific algorithmical approach is named on the third column.

Name	Type of data used	Algorithm
CLASS_STAR	Isophotal level measurements, morphological	Neural Network
SPREAD_MODEL	Pixel-level	Normalised Linear Discriminant
CM_T	Measurements on fitted shape, morphological	Second moments of Gaussian mixture fit (object)
MCAL_RATIO	Measurements on fitted shape, morphological	2nd moments of Gaussian mixture fit (noisified object and PSF)
ADA_PROB	Most discriminating features from a combination of simple functions used over all catalogue columns.	Boosted Decision Trees
GALSIFT_PROB	All catalogue columns (PCA)	Random Forests
SVM	MAG_AUTO,FLUX_RADIUS,SPREAD_MODEL, flux and morphology	Support Vector Machine
CONCENTRATION	Catalogue information, morphological	Direct subtraction of magnitudes measured with model and PSF
W1-J, J-K	Catalogue information, fluxes	Colour cut
HB_PROB	Catalogue information, fluxes	Template fitting of spectral energy distributions

both or an 'artifact' category (approximately 5% of the catalogue considered here).

3.3 CM_T

CM_T is an intrinsic size estimator for the object from the image fitting provided by the MOF pipeline.

Input data: The fitted Gaussian mixture model using the shapes across the images composing the coadd detection.

Method: The MOF code estimates the shapes and fluxes of objects detected in the coadd catalogues, using a mixture of Gaussians[8,9] to simulate the PSF light profile and then convolve them with assumed bulge and disk models (fitted independently for each object, finding the best linear combination) likewise approximated using Gaussian mixtures (Hogg & Lang 2013). This is done by fitting across several images of the same object in multiple epochs and bands and then subtracting the flux of neighbours accurately. Concretely, CM_T is defined as:

$$CM\_T = \langle x^2 \rangle + \langle y^2 \rangle \tag{5}$$

where x and y denote the distance from the object's centre determined by the model fit. The value $\langle x^2 \rangle + \langle y^2 \rangle$ can be obtained analytically from the individual component Gaussians. The PSF is convolved with the fitted model to obtain these pre-PSF values. An associated uncertainty is computed as well, and our best performing classifier, as tested[10] in the COSMOS field, is based on the quantity

CM_T + 2 × CM_T_ERR. Typical values are in the range between -0.5 and 0.5.

3.4 MCAL_RATIO

This measurement is derived from the size estimates obtained by the *metacalibration* technique, developed for shear measurement in weak lensing studies (Sheldon & Huff 2017), in which the single epoch objects are artificially sheared to quantify the response of such an effect in the image.

Input data: The object size and PSF model size obtained using this technique.

Method: This approach uses the same ngmix code as MOF above. However this measurement is much noisier as the metacalibration technique (Huff & Mandelbaum 2017) adds extra noise as part of the correlated noise correction. This is part of the procedure to correct for selection effects in shear inference, as detailed in Sheldon & Huff (2017). The discriminating quantity used is:

$$MCAL\_RATIO = \frac{T_{mcal}}{T_{PSF}} \tag{6}$$

where T_{mcal} and T_{PSF} are sizes of the object or PSF respectively as defined in Equation 5. In this case the size is obtained from a single Gaussian fit, so the suffix CM (composite model) is not used. Values are not constrained, but typical ranges explored for star-galaxy separation are between 0 and 1.

---

[8] https://github.com/esheldon/ngmix
[9] https://github.com/esheldon/ngmixer
[10] Technically, a different, *validation* set would be required to tune this classifier in terms of the quantity multiplying CM_T_ERR, to avoid bias a

towards a specific value, though in practice the differences are small between different choices.

3.5 ADA_PROB

This is the name given to a machine learning framework using the `scikit-learn` package (Pedregosa et al. 2011).

Input data: This method uses feature generation (using various simple mathematical functions of various catalogue variables), and feature pre-selection (selecting the most informative variables).

Method: The selected quantities are fed into several machine learning algorithms (including AdaBoost) which are drawn from `scikit-learn` with an additional probability recalibration step. The details of the framework are described in detail in Appendix A. Two variants have been used of this approach, using either `SExtractor` quantities `ADA_PROB` or `MOF` quantities `ADA_PROB_MOF`.

3.6 GALSIFT_PROB

A probabilistic estimate based on machine learning approach over principal components, as used in the 'Multi_class' algorithm in Soumagnac et al. (2015).

Input data: A principal component analysis (PCA) over the catalogue quantities is performed to outline the correlations between the object parameters and extract the most relevant information. We perform a calculation of the Fisher discriminant (Fisher 1936) for each of the new parameters to quantify their aptitude to separate between the classes.

$$\mathcal{F}_i = \frac{(\overline{X_{G,i}} - \overline{X_{S,i}})^2}{\sigma_{G,i}^2 + \sigma_{S,i}^2} \qquad (7)$$

G and S corresponding to the galaxy and star classes respectively.

Method: We select the parameters with the highest Fisher discriminant (hence the highest 'separation power' of the classes) and use them as input to a machine learning classification algorithm. Whereas in Soumagnac et al. (2015) the authors used `ANNz` (Collister & Lahav 2004), in this application we have replaced it by a Random Forest classification algorithm implemented as part of the `scikit-learn` package for Python (Pedregosa et al. 2011). The output is a probability of the object being a star or a galaxy. In this case, we have used a classifier based only on `MOF` quantities, `GALSIFT_PROB_MOF`.

3.7 SVM

Following Wei et al. (in prep), the support vector machine (`SVM`) is a single-band, purely morphological and magnitude based classifier.

Input data: The input features used by the SVM are `MAG_AUTO_I`, `FLUX_RADIUS_I`, and `SPREAD_MODEL_I`.

Method: `SVM` is a supervised machine learning algorithm that constructs a separating hyperplane in any arbitrary n-dimensional space that maximises the margins of objects to the hyperplane. To make the `SVM` robust across various data sets with intrinsic variations in observation conditions, the algorithm performs linear transformations on the three input features to remove the means and make the standard deviations across all objects to be one. This preprocessing procedure also allows all three features to have equal levels of feature importance. This prevents any features with particularly large

numerical values from dominating the `SVM` classification decision. The `SVM` uses a Gaussian radial basis function (rbf) kernel, where the hyperparameters, $\gamma = 0.01$ and $C = 46.4$, are selected while training the SVM through an exhaustive cross-validated grid search. The SVM outputs distances of objects to the hyperplane, where a high positive (negative) value corresponds to a high confidence star (galaxy) classification.

3.8 CONCENTRATION

A parameter similar to what was used as a star-galaxy classifier for SDSS (Abazajian et al. 2004).

Input data: The PSF and model magnitudes for each object.

Method: In the case of DES, this translates to the use of the difference between the `MOF` PSF magnitude and a bulge + disk, or composite, model magnitude computed by the `MOF` pipeline:

$$\mathrm{CONCENTRATION} = \mathrm{PSF\_MAG\_I} - \mathrm{CM\_MAG\_I} \qquad (8)$$

3.9 W1-J, J-K infrared bands

In the Stripe 82 region, we will compare with the information provided by the Vista Hemisphere Survey DR3 (McMahon et al. 2013) as proposed in Banerji et al. (2015) up to the available depth. We will also estimate the classification power of a cut in the infrared bands from WISE (Wright et al. 2010), 2MASS (Skrutskie et al. 2006), as described in Kovács & Szapudi (2015).

Input data: Magnitudes W1 (WISE), J (2MASS, VHS) and K (VHS).

Method: Colour cuts in W1-J and J-K.

3.10 HB_PROB

Additionally, we implemented a Hierarchical Bayesian method (`HB_PROB`) developed and explored by Fadely et al. (2012); Kim et al. (2015) with CFHTLS data. The lack of u-band in our case severely impacted the performance of this method, so it was not pursued further in our analysis.

Table 3 shows the specific selection methods used with respect to a varying threshold t for each of the algorithms used in this work.

4 PERFORMANCE ON CALIBRATION FIELDS

In this section, we will look first at the metrics used to compare classifiers using the calibration fields, describe the datasets (including training and validation) and finally analyze the results.

4.1 Receiver Operating Characteristic (ROC) curves

We compare the performance of the different classification techniques using the calibration fields, by calculating Receiver Operating Characteristic (ROC; Fawcett 2006; Bradley 1997) curves which compare the *True Positive Rate (TPR)* of galaxy or star detection, given a specific threshold for the classifier, versus the *False Positive Rate (FPR)*, as defined by:

Table 3. Selection methods.

Name	Selection method for galaxies using threshold t
CLASS_STAR	$CLASS\_STAR < t$
SPREAD_MODEL	$SPREAD\_MODEL + 1.67 * SPREADERR\_MODEL > t$
CM_T	$CM\_T + 2 * CM\_T\_ERR > t$
MCAL_RATIO	$MCAL\_RATIO > t$
ADA_PROB	$ADA\_PROB > t$
GALSIFT_PROB	$GALSIFT\_PROB > t$
SVM	$SVM\_PROB > t$
CONCENTRATION	$PSF\_MAG\_I - CM\_MAG\_I > t$
WISE J-K	$(J - K - 0.6)/(MAG\_AUTO\_G - MAG\_AUTO\_I) > t$

$$TPR = \frac{TP}{TP + FN} \qquad (9)$$

$$FPR = \frac{FP}{FP + TN} \qquad (10)$$

where TP are correctly identified galaxies, given a cut for a specific classifier; FN are incorrectly classified galaxies as stars; FP are incorrectly classified stars as galaxies and TN correctly identified stars (in the assumption of using 'truth' for galaxy type). See Table 4 for a reference on these concepts. Therefore, the ROC curve is confined by construction to an area spanning from 0 to 1 in FPR and TPR. As we vary the threshold t for classification for a given classifier (Table 3), a curve will be drawn across the area from (0,0) to (1,1). A completely random 'classifier' would show as a diagonal line.

In particular, the AUC (area under the ROC curve) has been classically used as a threshold-independent metric to compare the performance of classifiers, as well as being relatively insensitive to the specific positive to negative composition (as long as sufficient statistics are available). The closer the AUC gets to unity, the better the discriminating power of the classifier associated with that particular curve. Again, a random classifier would show an AUC value of 0.5.

There are, however, some caveats to be aware of, namely the possibility of misleading results when ROC curves cross each other (Hand 2009) and that misclassification costs can be different according to the scientific case, and this is not reflected in ROC curves. We address this by extending the range of metrics used for different classifiers, in order to have a broader view of the performance for our particular needs.

4.2 Purity and completeness

In astronomy, we are interested in evaluating the performance of classifiers in terms of their impact on measurable on parameters of interest. It is common to find the requirements for a survey defined in terms of *purity* and *completeness*. In Soumagnac et al. (2015), for example, the authors formulate the scientific requirements for weak lensing and large-scale structure studies in terms of these two observables.

'Purity' is a measurement of the contamination of a sample by misclassified objects, which can also be called *precision* or *positive predictive value (PPV)*:

$$PPV = \frac{TP}{TP + FP} \qquad (11)$$

'Completeness' (also known as, *recall*) is another name for the

TPR defined in Equation 9. A good approach to easily compare the performances of several classifiers is to use the precision-recall (PR) curve, where both quantities can be visualised simultaneously.

4.3 Training and testing fields

The dataset on which we train the machine learning (ML) codes is the weak lensing catalogue from HST ACS in the COSMOS field(Leauthaud et al. 2007), as this provides a largely unbiased measurement of all extended and point-like sources from DES (albeit the star-galaxy mixture is affected by the specific position in the sky with respect to the Galactic plane). In particular, the MU_CLASS parameter is used for this reference, defined in the peak surface brightness - MAG_AUTO space, which in space-based imaging shows very distinct loci with respect to the same objects viewed through the atmosphere. This has been used previously in star-galaxy separation assessments in, e.g., Crocce et al. (2016) and Aihara et al. (2018).

This training set, after a 1" positional match with DES sources, contains $\sim$ 114k extended and $\sim$ 12k point-like sources. The COSMOS dataset will also be used for some tests only with the non-ML codes in order to avoid biased conclusions based on their training in that same area.

Even in the case in which we use unbiased, imaging data, the particular position on the sky of the field will condition the relative mixture of stars and galaxies in a prominent way. Therefore we add some extra imaging data extracted from the Hubble Source Catalog[11] (Hubble-SC) (Whitmore et al. 2016) where it overlaps the DES survey. Most of it is either too inhomogeneous or targets specific objects (nearby, large galaxies or globular clusters), but a few deep fields can be matched with some of the SN fields from DES. In this case we use the Hubble-SC catalogues' concentration index with a cut of 1.2 which seems optimal in the concentration-magnitude plane.

Spectroscopy is also a valuable resource to provide a one-to-one truth table for our classifications. However, the spectroscopic targeting and measurement efficiency is not complete in a statistical sense relative to the DES catalogue, as certain types of sources were given higher priority and some types are more difficult to classify spectroscopically, therefore the testing of purity/completeness can be strongly biased. The photometric properties of the stars and galaxies selected can also be highly skewed to particular types that introduce additional biases. This limits the usefulness of any purity metric we try to derive from these fields. For this reason, the spectroscopic datasets have been limited to those that provide a relatively unbiased sample by construction, which includes the VVDS-DEEP and VVDS-CDFS (Le Fèvre et al. 2013) data releases. The SDSS

---

[11] `https://archive.stsci.edu/hst/hsc/`

Table 4. Definitions of different figures of merit for classifiers, according to the outcome of the classification using a 'truth' reference (also termed 'confusion matrix'). The term 'positive' can refer to 'galaxy' or 'star' classes depending on the use case. The metrics examined in this work are emphasised in bold. 'Purity' can be used as a synonym for the Positive Predictive Value, PPV, whereas 'completeness' can be interchanged with the True Positive Rate, TPR.

		Prediction			
		Positive	Negative		
Truth	Positive	True Positive (TP)	False Negative (FN)	**True positive rate** **(TPR) = TP/(TP+FN)**	False negative rate (FNR) = FN/(TP+FN)
	Negative	False Positive (FP)	True Negative (TN)	**False positive rate** **(FPR) = FP/(FP+TN)**	True negative rate (TNR) = TN/(FP+TN)
		Positive predictive value **(PPV) = TP/(TP+FP)**	False omission rate (FOR) = FN/(FN+TN)		
		False discovery rate (FDR) = FP/(TP+FP)	Negative predictive value (NPV) = TN/(FN+TN)		

DR13 (Albareti et al. 2017) updated spectro-photometric sample over Stripe 82 is also used due to the relative variety of spectra available, and the possibility to test our classification methods against 'true' spectroscopic typing. We use redshifts (a cut in $z < 0.001$) as the method to identify stars. However, we also consider a selection based on SDSS spectroscopic CLASS obtaining similar conclusions. For the VVDS data, we require the redshifts to be 'reliable' according to the classification in Le Fèvre et al. (2013), that is, values of 2,3,4 or 9 in their redshift quality estimate.

Both the COSMOS catalogues and the ones recovered from the Hubble-SC have been cross-tested against spectroscopic catalogues (VIMOS-Ultra Deep Survey DR1 (Tasca et al. 2017), zCOSMOS DR3 (Lilly et al. 2009), and VVDS-CDFS (Le Fèvre et al. 2013)) to check the robustness of their morphological classifications against a 'true' type based on their spectra. In both cases, around 5% of spectroscopically classified stars are misclassified as galaxies when using these space-imaging based measurements, whereas around 2% of spectroscopically classified galaxies are misclassified as stars. This misclassification happens at faint magnitudes (F814W from ACS-HST > 24 for COSMOS, F814W > 23 for the other Hubble fields used), denoting possible compact galaxies that are unresolvable by HST, errors in the spectroscopic measurement or matching. These corrections are not considered for the purity estimates derived here as they belong to fainter fluxes than the truth tables used in our tests.

See Table 1 and Appendix B for details on the reference data in different fields including the database queries used to create these datasets.

4.4 Results

4.4.1 Using HST imaging

We compare here the results for the classifiers used on the COSMOS field (excluding the ML codes that were trained on this field) and the supernova fields for which we have found publicly available deep HST data from the Hubble-SC.

• The results for the ROC comparison are shown in Figure 1 for the COSMOS field and Figure 2 for the SN fields with Hubble-SC data. The AUC of the respective curves are tabulated in Table 5.

From these plots it can be readily seen that among the morphological classifiers, the algorithms based on a linear discriminant over coadded images, SPREAD_MODEL, and intrinsic size on MOF estimates, CM_T, are the best performing ones. It is also seen that the ML classifiers (in Figures 2 and 3) do perform better, even considering a different field with respect to training as in the case of the Hubble-SC test. It is noteworthy to point out that most of the differences

showcased in Figures 1 and 2 become more evident when we restrict ourselves to faint objects ($i > 22$). The SPREAD_MODEL-based cut does a good job at avoiding stellar contamination but suffers from decreased galaxy completeness. This is a result of the galaxy locus merging with the stellar locus in the magnitude-SPREAD_MODEL space where noisier measurements will increase the effect even further. CM_T fares better in this respect, but a conservative cut will provide a more pure galaxy sample using SPREAD_MODEL. On the other hand, the metacalibration size ratio does not perform as well as the other morphological classifiers, though this measurement is noisier than the direct assessment of sizes and shapes from the MOF pipeline.

• Figure 2 shows that ML classifiers are able to take advantage of ancillary information for very faint objects where shape measurements are uncertain. Results with SVM in the SN fields show that a ML approach based exclusively on morphological and magnitude information can provide some advantage over simple cuts on morphological variables. SVM is shown to be robust outside of its training field, however other machine learning algorithms provide an extra edge in performance as shown by the higher AUC values. This is due to forgoing the additional information encoded in the rest of the variables available in the catalogue. However, this approach could provide a middle-ground solution to the issues one might encounter when incorporating colour-based information, which can incorporate interesting physics we would not like to be entangled with our star-galaxy sample selection (see Section 6). Further developments of this approach is explored in Wei et al. (in prep).

The comparison between the COSMOS and Hubble-SC fields reveals that the CM_T classification is more robust as we switch between fields. SPREAD_MODEL and CLASS_STAR, which are derived from coadded PSFs are more vulnerable to the contribution of bad exposures and PSF inhomogeneities in the coadded image. It is worthwhile noting here that preliminary tests on Y3 data (DES Collaboration 2018) using Hyper Suprime Camera deep data (Aihara et al. 2018) reinforce this idea, which will be explored further in a future publication, therefore favoring in general the use of a multi-epoch classifier (such as CM_T based on the MOF pipeline). Both the COSMOS field dataset and SN field coadds have a much smaller dithering than the wide-field exposures. This might artificially bias classifications based on the coadded PSF to somewhat better performances than actually present in the wide-field data.

• Figures 3 and 4 show the precision-recall metric, for galaxies and stars respectively (COSMOS plots not shown for conciseness, but provide similar conclusions).

These plots provide a similar conclusion as the ROC curves, though in terms of more useful quantities with respect to scientific requirements such as the recall (i.e. completeness) and precision (i.e.

purity). Again, the `CM_T` morphological classifier and the ML codes provide the best results, and this manifests even more strongly for selecting a star sample (these results motivate the choice for stellar classification based on multi-epoch pipelines in Shipp et al. 2018). It is noteworthy to add that the ML classifiers using `MOF` quantities do not add much more than a straight cut in `CM_T` itself, due to the large information content included in this classifier with regards to star-galaxy classification. On the other hand, the ML classifiers based on `SExtractor` quantities are able to extract more value from the different outputs of this code, with respect to a simple `SPREAD_MODEL` cut.

• In Figures 5 and 6, we can appreciate the dependence of the completeness with the magnitude as we go to the fainter end in the sample, in the galactic and stellar case respectively.

Unlike in the previous plots, in this case a choice of threshold has to be made. We have decided to pick cuts in the variables in question in order to have a similar galaxy purity (99%) in each magnitude bin, so we can compare completeness appropriately, and similarly for stars (80%). We chose the COSMOS field which has good statistics to faint magnitudes, though this disallows using the ML codes in the comparison. This example shows a case where classifiers such as the concentration estimation from the `MOF` pipeline, not necessarily favored at first sight from the integral under the ROC curve, works better in this regime due to its good selection of very pure samples. The ROC curve only informs about *overall* classifier performance (i.e. considering all possible thresholds), and different classifiers have to be tested for the specific science case at hand.

For stars, a similar behaviour is seen for `CM_T`, `CONCENTRATION` and `SPREAD_MODEL`. `CLASS_STAR` for instance suffers from a poor completeness near the faint end, as a high thresholding cut in this case removes most of the objects, which in the neural network tend to cluster towards intermediate values when the object classification is uncertain. `MCAL_RATIO` incorporates noisier measurements and additional cuts to the sample that make it less complete when providing a classified sample.

• In addition, in Figure 7 a similar comparison is shown as a function of a realization of the photometric redshift from the probability distribution function obtained from the algorithm BPZ (Benítez 2000), this time also adding the ML classifiers (again over the SN fields with Hubble-SC). A similar conclusion is drawn from these plots; `MOF` fitting methods and ML classifiers perform best, as indicated by the ROC curves. Note the stability of the purity of the galaxy sample with respect to photo-z, suggesting that a photo-z selected sample would not be biased by the star-galaxy separation classifiers analyzed here (however, see Section 6.1 for an important caveat to this conclusion).

4.4.2 *Using ground-based spectroscopy*

Turning now to tests on the overlapping spectroscopic data, we show ROC plots to demonstrate the consistency with the results from the previous section and add a comparison with external infrared information.

Figure 8 shows the ROC for the VVDS test and Figure 9 shows the ROC for the Stripe 82 test. The former does not add much to the conclusions mentioned above, but provides an assurance that conclusions are consistent with a different class of 'truth' typing. We also add here a test on the SN fields computing the ROC curves and their areas, versus the signal to noise of the detected objects, to demonstrate it behaves as expected as well, including the ML codes (see Figure 10).
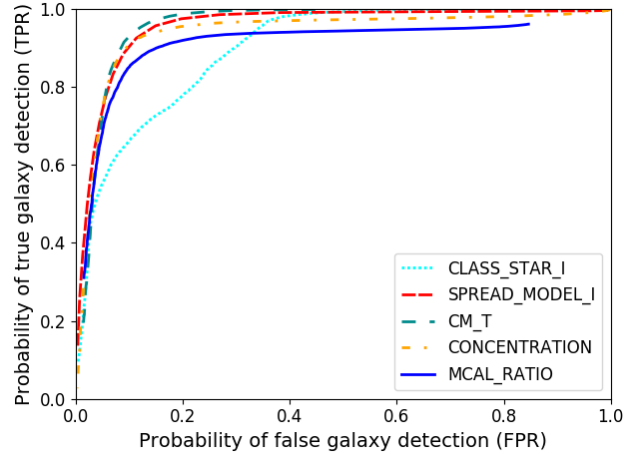
Figure 1. ROC plot for classifiers tested on the COSMOS field. Only non-ML codes are shown, as the machine-learning ones were trained in this dataset. Magnitude range is given by `MAG_AUTO_I` = (17,24). The `SPREAD_MODEL`-based cut is similar to `MODEST_CLASS` used in Y1 analyses. The ROC curve is obtained by varying the threshold at which the classification divides the galaxy and star sample.
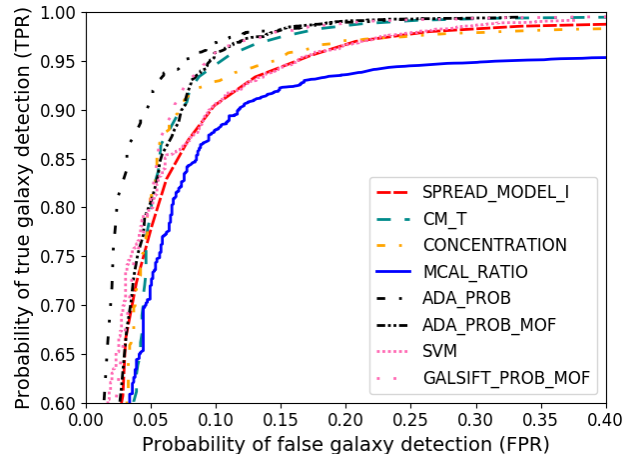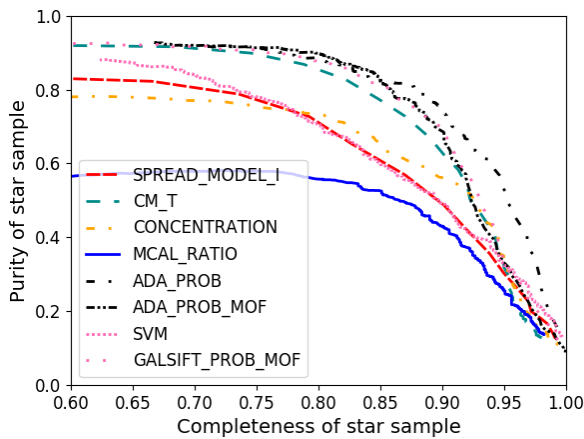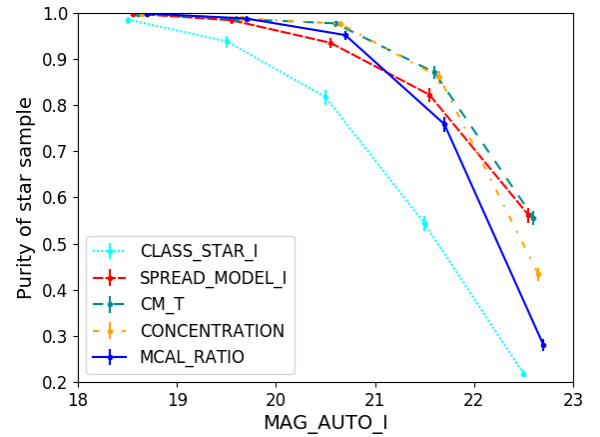

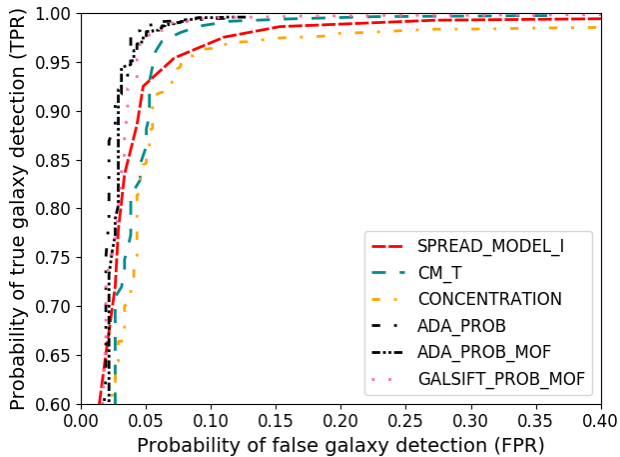**Figure 2.** ROC plot for classifiers tested on the SN fields over the Hubble-SC catalogue. Magnitude range is given by `MAG_AUTO_I` = (17,24). The `SPREAD_MODEL`-based cut is similar to `MODEST_CLASS` used in Y1 analyses. The ROC curve is obtained by varying the threshold at which the classification divides the galaxy and star sample.

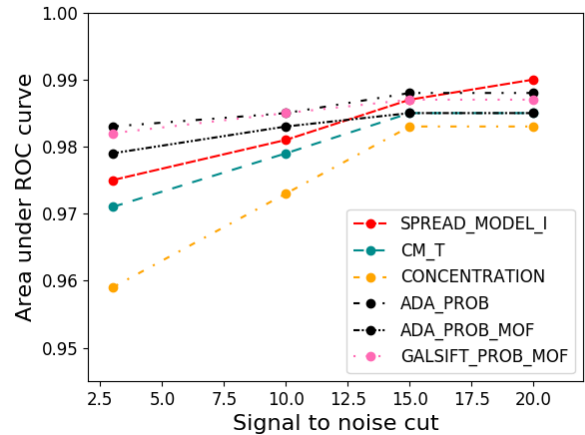The stripe 82 dataset is shallower and therefore does not allow for a clear distinction between the performance of most of the algorithms described here. The comparison with the combination with external infrared colour cuts on the other hand, shows an important increase in performance, specifically when attempting to select a very pure stellar sample, as already advanced in Baldry et al. (2010) and Banerji et al. (2015). It is important to note here again that the nature of the test is different with respect to the ones based on space imaging. In this case we are using spectroscopic redshifts to determine the nature of the object (galactic or extra-galactic) and not its extendedness. What we see here is that infrared information will select out the stars from the galaxy and QSO (which are point-like

Table 5. Area under the ROC curves for different classifiers. Dashes indicate tests that have not been run for that specific code and dataset combination.

Name	COSMOS, imaging	SN fields, imaging	SN fields, spectroscopy	stripe 82, spectroscopy
CLASS_STAR	0.898	0.885	0.950	0.976
SPREAD_MODEL	0.954	0.956	0.975	0.962
CM_T (MOF)	0.957	0.959	0.971	0.972
CONCENTRATION (MOF)	0.938	0.953	0.950	0.967
MCAL_RATIO	0.910	0.924	–	–
VHS J-K vs G-I	–	–	–	0.993
ADA_PROB	–	0.978	0.983	0.967
ADA_PROB (MOF)	–	0.967	0.980	0.967
GALSIFT_PROB (MOF)	–	0.969	0.981	0.962
SVM	–	0.962	–	–
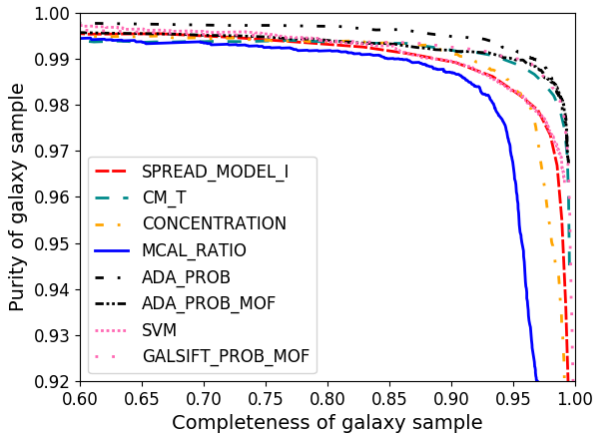
Figure 3. Precision-Recall (or completeness-purity) plot for classifiers tested on the SN fields over the Hubble-SC catalogue, using **galaxies** as truth. Magnitude range is given by MAG_AUTO_I = (17,24). The SPREAD_MODEL-based cut is similar to the MODEST_CLASS used in DES Y1 analyses.
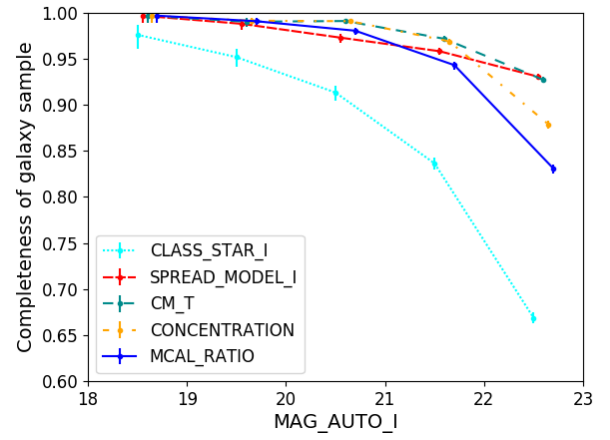
Figure 5. Completeness of a **galaxy** sample as a function of magnitude for classifiers tested on the COSMOS field, for a fixed galaxy purity of 99%.

Figure 4. Precision-Recall (or completeness-purity) plot for classifiers tested on the SN fields over the Hubble-SC catalogue, using **stars** as truth. Magnitude range is given by MAG_AUTO_I = (17,24). The SPREAD_MODEL-based cut is similar to the MODEST_CLASS used in DES Y1 analyses.

Figure 6. Completeness of **stellar** sample as a function of magnitude for classifiers tested on the COSMOS field, for a fixed 80% purity.

generally) population. We have also attempted to add W1-J version from 2MASS and WISE (as suggested in Kovács & Szapudi (2015)) but the matches proved too shallow to be of any interest for these samples.

Unfortunately, the current VHS data does not cover the full
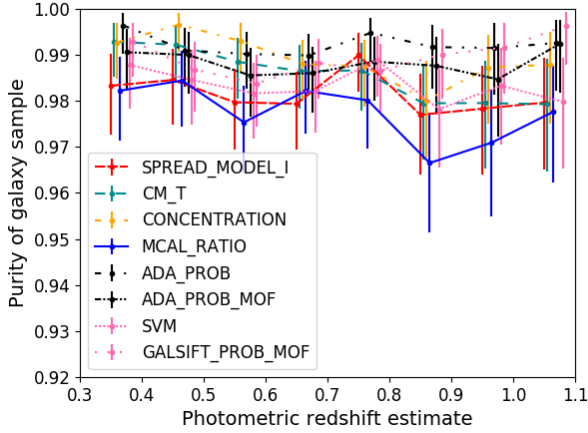
Figure 7. Purity of the galaxy sample as a function of photo-z for classifiers tested on Hubble-SC matches over the SN fields field, for a fixed 90% completeness. We use a random MonteCarlo sampling of the probability distribution function of redshift predicted by BPZ for that particular object as an estimate of its photo-z.

Figure 9. ROC plot for classifiers tested on the Stripe 82 region overlapping SDSS and VHS data. Magnitude range is given by `MAG_AUTO_I` = (17,21). Note the logarithmic scale in the x-axis in this instance. The ROC curve is obtained by varying the threshold at which the classification divides the galaxy and star sample.

Figure 8. ROC plot for classifiers tested on the SN fields over the VVDS catalogues. Magnitude range is given by `MAG_AUTO_I` = (17,24). The ROC curve is obtained by varying the threshold at which the classification divides the galaxy and star sample.

Figure 10. Area under the curve measured for the same classifiers as Figure 8, for different signal-to-noise thresholds, using the `MAGERR_AUTO` quantity.

5 PERFORMANCE ON APPLICATION FIELD

It has been shown by Fadely et al. (2012) that machine learning techniques in star-galaxy classification will perform better if a representative training dataset is found. We have studied the impact of this effect by testing ML algorithms over different fields other than the training set in Section 4. However all these additional areas are quite constrained either in depth or area, when compared to the complete DES volume.

In this section, we extend the scope of the performance tests in classification to have a broader picture, by making the following checks on the application field (see Section 2):

(i) General distribution of the classifier-flux space to qualitatively analyze the algorithms' outputs.

(ii) Number count distributions of stars against a well-tested sim-

breadth and depth of the survey and a careful combined catalogue with adequate matching is needed (overcoming the less precise infrared astrometry) beyond what was done here for comparison purposes. Cross-matching with bright sources will be explored in more detail with DES Y3 data with the goals of enhancing star selection for creating PSF models and reference catalogues for large scale structure. A combination of classifiers, as done for instance in Kim et al. (2015) or Molino et al. (2014), seems to be an appropriate option in this case and even more so if matched-aperture photometry of VHS data can be performed survey-wide for DES (Banerji et al. 2015). This would also have important applications for photometric redshift determination (Banerji et al. 2008).

ulation, both as a function of magnitude and as function of galactic latitude.

(iii) Galaxy versus star density profiles in search of correlations, using different proxies for the true stellar distribution.

(iv) Density of classified galaxies as a function of proximity to the Large Magellanic Cloud.

(v) Consistency of classified stars with the expected stellar locus (Covey et al. 2007).

Except where noted, the sample sizes for each of these cases are approximately 1 million objects, limited by the size of tested region, magnitude range or photo-z binning.

5.1 Classifier outputs

A first step towards understanding the quality of classification for different algorithms in the application field of DES is to study the outputs as a function of magnitude and the number counts of classified objects.

In Figure 11 several density plots showcase how objects distribute in the classifier-magnitude space. These distributions are based on a 1% sample of the Y1 Gold catalogue. Direct morphological outputs from the DESDM pipeline CLASS_STAR, SPREAD_MODEL and CM_T) show two loci that merge in the faint end. CLASS_STAR outputs merge into a region of 50% probability by construction of its base neural network. This uncertainty region appears at shallower magnitudes than other classifiers as shown previously, due to the characteristics of the simulations used for its training. However, a classifier using a feature importance selection[12] manifests a more 'clear-cut' classification of objects, with a large predominance of galaxies at the faint end, as expected. This can be attributed to the fact that there is a large predominance of galaxies over stars in raw numbers (a very imbalanced dataset) at faint magnitudes, so the algorithms will 'learn' that the most probable classification for a given object in this range is a galaxy.

5.2 Number counts of classified stars

On the other hand, if we limit our study to the point in which Y1 data are fairly complete over a large area ($r \sim 22.5$), we can assess for instance the similarity of the stellar distribution in magnitude versus a detailed simulation such as Galaxia (Sharma et al. 2011), which has been tested against Gaia DR1 data (Gaia Collaboration (2016), Koposov private communication). This is shown in Figure 12 for a few selected classifiers, spanning a varied range of those mentioned in Section 3, in the DES r band. Thresholds were used to provide a similar number of stars as MODEST_CLASS, the default DES Y1 Gold star-galaxy classifier based on SPREAD_MODEL. Up to $r \sim 21$, the behaviour for most of them with respect to the simulation is similar. Two machine learning classifiers based on MOF quantities show a significant lack of brigh objects ($r < 19$) due to failures from the Y1 version of the MOF pipeline in fitting stars in this regime[13]. This has been identified as failures of the galaxy fits for which MOF was designed when applied to moderately bright stars. A consistent overestimation of stars by Galaxia with respect to DES stars is apparent for all classifiers, as was seen in Li et al. (2016). On the other hand, other simulations such as the ones described in Robin

---

[12] A preselection of the input variables which provide the most predictive power for the task at hand, e.g. star-galaxy separation.
[13] Y3 Gold MOF photometry has solved this issue.

et al. (2003) and Girardi et al. (2005) show discrepancies of this size as well at this latitude and longitude. This disappears at the faint end, as compact galaxies start to leak into the stellar sample. After that, a completeness drop kicks in as we enter the survey's magnitude limit. At the faint end, CLASS_STAR shows a drop in completeness sooner than the other classifiers. The nature of this classifier, which provides an intermediate value of probability for 'uncertain' sources, is such that a fixed threshold cut tends to 'lose' stars at the faint end, if we adjust all classifiers to the same number of stars.

5.3 Stellar density as a function of Galactic latitude

As a complementary measure of goodness of stellar identification, we compare the number of stars as a function of Galactic latitude (Figure 13). We limit the comparison to the range in which any possible issues deriving from the current MOF processing are avoided (see Section 5.2). A slight deficit is seen nonetheless as was verified before, but the comparison of all these different approaches are qualitatively in the same range, without any preferred or outstanding behaviour from any of the classifiers tested here.

5.4 Galaxy vs stellar density

As mentioned in Section 4.2, we do not have a large-scale 'truth' table available that we could use as reference to check the precision of our classification on an object-by-object basis. However, several studies of large-scale structure (e.g., Ross et al. 2011) have devised an estimate of the purity of the galaxy sample, for a given classification scheme, by measuring correlations of classified galaxy density versus some reliable measurement of the relative stellar distribution (using a very pure cut for stars, a model, or an external catalogue). This is done via the pixelisation of the field using the HEALPix software (Górski et al. 2005) and fitting a linear relation between the galaxy overdensity as a function of stellar density in said pixels. For this study we used a pixelisation parameter NSIDE=512, which corresponds to a pixel size of approximately 0.01 square degrees.

In Figure 14 we show a comparison of the galaxy density as a function of stellar density for several classifiers, tested on the application field for the galaxy sample with the magnitude cuts shown in Table 6. Errors for each point are computed using the jackknife method (Efron & Stein 1981), whereas the ones in the table correspond to the estimated error from the fit.

The galaxy density over samples of increasing stellar density would theoretically increase with a linear relationship, if stellar contamination was the only effect that a dense star field would introduce. However, as seen already in Ross et al. (2011) (in their Figure 3), moderately bright stars can also induce an 'occultation' effect which makes detection around them more difficult. This effect is more predominant for fainter sources. This will create an inverse, possibly non-linear, relationship between galaxy density and stellar density. The overall effect is to create a proportionality relationship at low to moderate stellar densities, which may or may not change in slope and even decrease, depending on the separation power of the classifier, as galaxies get removed from the catalog due to the presence of foreground bright stars. For our purposes here, i.e., to understand the star-galaxy separation power for different classifiers, we use the intercept value of the linear fit to the first part of the plot, in order to estimate the purity of the galaxy sample. We adjusted the cuts for the classifiers to provide a similar number of detected 'galaxies' (i.e. a similar completeness) as MODEST_CLASS, in order
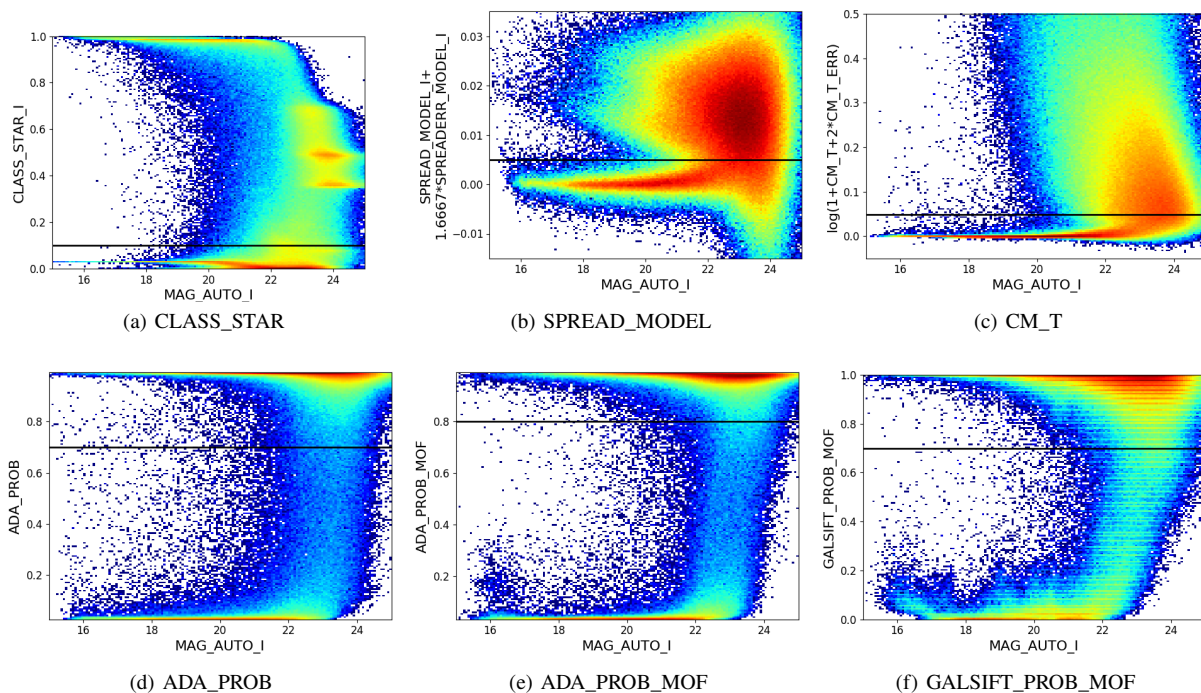
Figure 11. Object classification heatmaps as a function of magnitude for different classifiers. The black line represents the cut for which a 99% galaxy purity is obtained in the Hubble-SC sample in the i=(17,24) magnitude range. With the exception of CLASS_STAR, all classifiers assign higher values to extended sources.
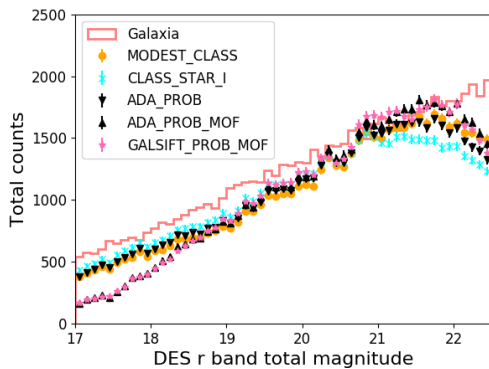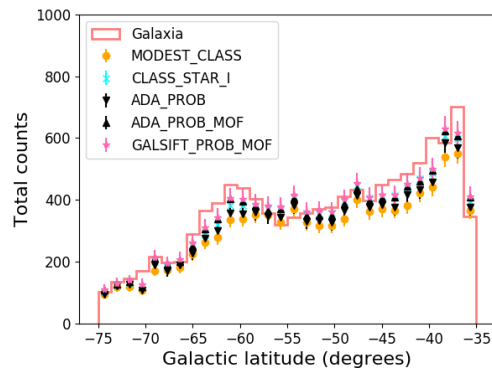

Figure 12. Counts for stars as classified by different algorithms compared to a `Galaxia` simulation (Sharma et al. 2011) using DES photometry, in the patch of the Y1 DES footprint with 45<RA<50, -45<DEC<-50.

Figure 13. Counts for stars as classified by different algorithms compared to a Galaxia simulation (Sharma et al. 2011) for the application field (SPT region of the DES-Y1 footprint) for the magnitude range $r = (19,21.5)$.

to get a better handle on how purity compares on the same grounds, similarly to what we did on Section 4.4.1.

We note that using the application sample in bulk shows no strong contamination component for the SPREAD_MODEL- or MOF-based quantities or for the machine learning approaches using magnitude and colour information. Slightly better performance is found using MOF quantities and the ADABOOST code, especially for fainter objects. This is explained by the more accurate shape measurement of the MOF code and by how additional information is captured by ADA_PROB_MOF.

One of the components of these calculations is the choice of a star map to establish the density relationships. We have derived a ∼ 1% systematic uncertainty in the estimation of the impurity derived

from comparing brighter and fainter stellar samples (Figure 15). The 2MASS and Tycho-2 (Skrutskie et al. 2006; Høg et al. 2000) stellar maps are included for completeness, but their magnitude range does not track accurately the range of brightness we need to account for Milky Way distribution in DES. Gaia's DR2 corresponds to the data described in Brown et al. (2018).

5.5 Galaxy ratio near the Large Magellanic Cloud

Using the same pixelisation as above, we also approach the comparison of different classifiers using a figure of merit based on the identified galaxy density in each of these pixels, as compared to the one found at a certain distance to the centre of the Large Magellanic

Table 6. Contamination for different classification methods for the galaxy vs stellar density tests. Threshold cuts were selected to adjust to the same number of detected galaxies as provided by `MODEST_CLASS`.

Sample	MODEST_CLASS	CLASS_STAR	ADA_PROB	ADA_PROB_MOF	GALSIFT_PROB_MOF	CM_T
i< 22	2.7 ± 0.4%	2.1 ± 0.5%	2.2 ± 0.4%	2.2 ± 0.4%	2.3 ± 0.4%	2.3 ± 0.4%
i< 23	3.2 ± 0.4%	4.6 ± 0.2%	2.4 ± 0.4%	2.1 ± 0.4%	2.8 ± 0.3%	2.4 ± 0.4%
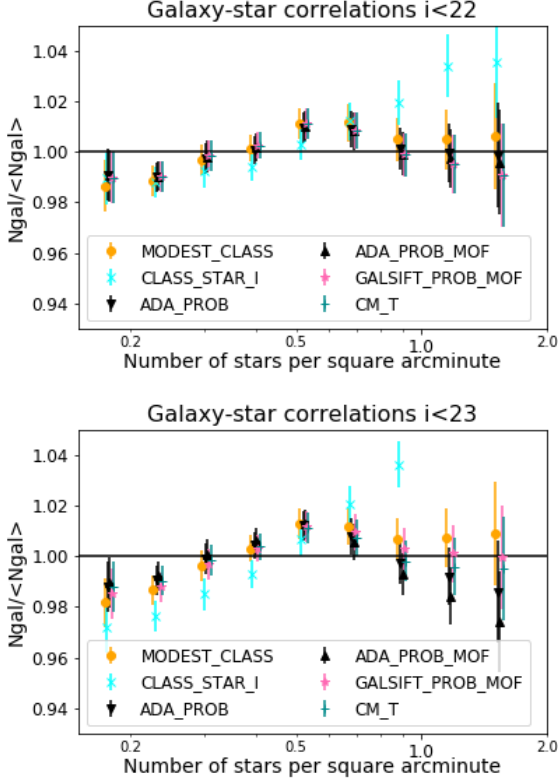
Figure 14. Galaxy vs star density plot for several classifiers, for $i < 22$ (*top*) and $i < 23$ (*bottom*). Star density is traced by an external map of 'secure' moderately bright stars.
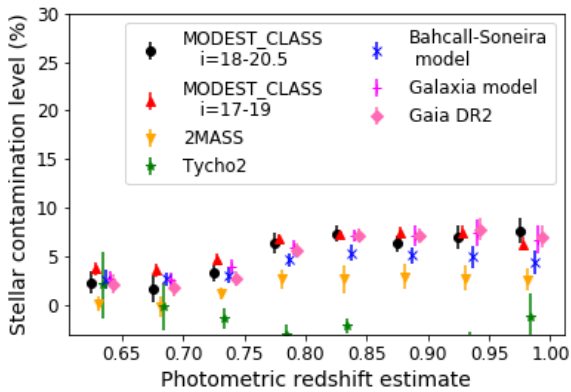
Figure 15. Star contamination levels for different stellar maps. A ~ 1% systematic uncertainty is derived by comparing the `MODEST_CLASS` moderate to bright stars, is estimated from this plot. Tycho and 2MASS stars are added for comparison, but their magnitude ranges (much brighter than the stellar sample considered as contaminants) do not make them good candidates for deriving this uncertainty.
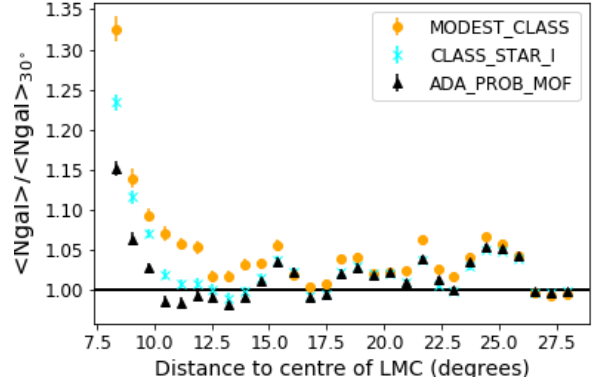
Figure 16. Galaxy ratio (with respect to galaxy density at 30 degrees from LMC) for as a function of angular distance from the LMC centre.

Cloud (LMC), set at $(\alpha, \delta) = $ (5h23m34.5s, −69°45'11"). This value is normalised to one at 30 degrees from the centre of the LMC (Figure 16). Here we use a flux limited sample with $i < 23$. In this case, we can see a clear advantage in using a classifier with multiple input attributes (including colour), possibly helped by the fact that in a crowded field such as the peripheries of the LMC, morphology starts to have a smaller discriminating power. On the other hand, the LMC has a bluer population, but this doesn't seem to offset the ML classification significantly, though this aspect is worth studying further in a future work.

Using a metric such as this at a given fixed distance of the LMC could be useful as a figure of merit. In this case 10 degrees seems convenient but we must remark that this could be due to the odd geometry available around the LMC, so other photometric surveys might find other ranges for comparison more valuable.

5.6 Stellar locus of classified stars

Finally, we tested the consistency of the stellar locus derived in $r - i$ vs. $g - r$ colour space to a similar fit to stars in the COSMOS field. The stellar locus was fit by a fifth-order polynomial, as shown in Figure 17, similarly to what is realised in Covey et al. (2007). The same fit curve from Figure 17 is shown again versus several classifiers in Figure 18. In general a good agreement is seen except for the faintest end, where classified stars seem to deviate from the expected stellar locus for `CLASS_STAR`.

6 DISCUSSION: IMPLICATIONS FOR LARGE-SCALE STRUCTURE AND MILKY WAY STUDIES

In the previous section we explored a variety of tests both with and without truth information assessing the relative performance of a wide range of star-galaxy classifiers in DES Y1 data. We now turn to the impact of making different selections on scientific analyses
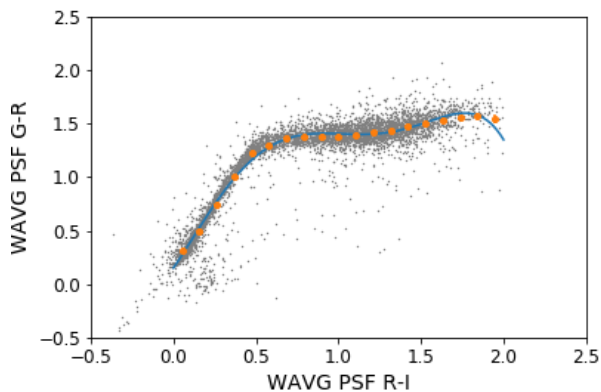
Figure 17. Fit to stellar locus using a fifth degree polynomial
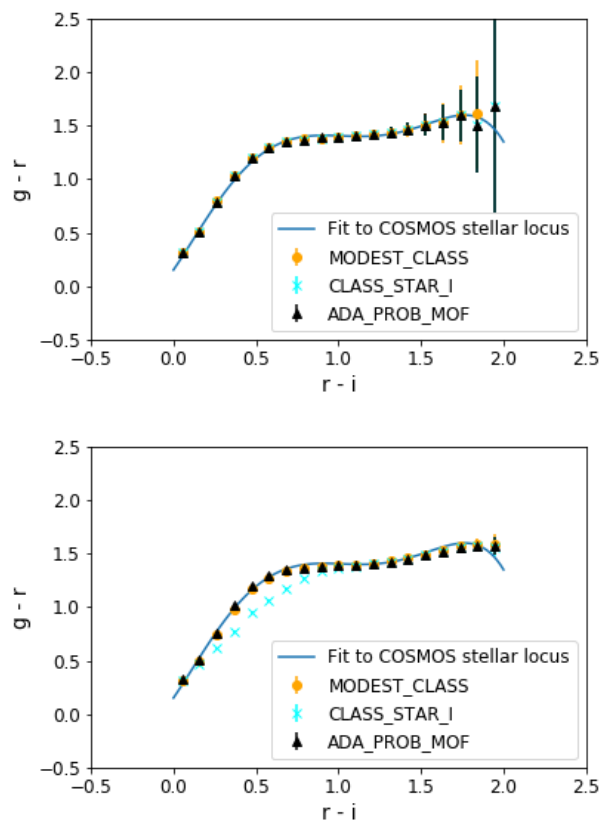
Figure 18. Stellar locus for star samples from various classifiers, for a bright sample ($i < 21$, *top*) and a fainter one ($i < 24$, *bottom*).

of interest to astronomers and cosmologists. Though it is beyond the scope of this work to define specific choices for any arbitrary study, in this section we sketch out the general implications of the results shown here for two broad ranging topics of interest, namely the large-scale structure (LSS) of galaxies and Milky Way analyses within DES. With regards to weak lensing shear catalogues, Zuntz et al. (2017) have shown that star-galaxy contamination is at most a second-order contaminant when either MODEST_CLASS or MCAL_RATIO are used for the DES Y1 cosmology analyses. For

a thorough discussion on LSS and weak lensing requirements for star-galaxy separation, Soumagnac et al. (2015) provides an in-depth review.

6.1 Large-Scale Structure

The impact of stellar contamination on studies of clustering amplitude has been well studied for several years now (Ross et al. 2011; Crocce et al. 2016, e.g.,) with an impact of the order of $(1 - I)^2$ in the angular correlation function $\omega(\theta)$ if we assume an unclustered component that contaminates the galaxy population with impurity fraction I. A large contamination can severely dilute the signal (reducing the significance of the BAO peak as shown by Carnero et al. (2012)), or even create a large-scale component if unaccounted for, thus mimicking an effect such as primordial non-Gaussianities (Giannantonio & Percival 2014). However, in the range $I \sim O(2\%)$, the accuracy by which we determine I becomes much more relevant, as this is the systematic that will dominate in the determination of the uncertainty in galaxy bias measurements and multiple probe analyses.

Figure 14 implies that the choice of classifier does not matter too much for cosmology analyses in the broadest sense. However, going into a more realistic sample for large scale structure studies, using a selection for red galaxies that have better estimated photo-z and galaxy bias (Crocce et al. 2017) for BAO analysis for example, some evident differences appear for the highest redshifts (where due to their colours, many faint stars are mis-classified into those bins of photo-z). This is the main photo-z region of interest for BAO for DES. Also between the classifiers, which become more evident when the flux cut is driven to fainter magnitudes as shown before. See Figure 19.

These results show that a realistic LSS sample, is more severely affected by stellar contamination, driving the impurity levels up to $5 - 6\%$ in some redshift bins. This is seen more clearly in Figure 20 where photo-zs are shown for the true stars in the fields overlapping the COSMOS region for a general selection and an LSS-like, red galaxy selection. One way to drive down this impurity therefore is to either apply more stringent constraints to the star-galaxy thresholds, sacrificing a percentage of true galaxies along the way. For the case of MODEST_CLASS and ADA_PROB_MOF, we can push down to 2% by removing ~9% and ~4% galaxies respectively. Though a ML approach seems more convenient in this case, the use of colour and magnitude information may lead to potential correlations between object classification and photo-z determination that must be investigated in more detail. As for the uncertainty of determining I using the density plots, Figure 15 shows that using fainter stellar maps to derive the impurity via this method generates a different contamination rate. This can be due to tracing of different components of the Galaxy, but for maps built upon possibly contaminated data it could well be that the star maps themselves are not ideal (e.g. the bright MODEST_CLASS stars could have a small component from misclassified compact galaxies). An improvement in understanding the underlying Galactic stellar structure through simulations or an adequate culling of the reference stellar maps to improve agreement would reduce this limitation in the determination of the impurity level, I.

6.2 Milky Way

In the case of Milky Way studies, in broad terms we are interested in obtaining a more complete and pure stellar sample, down to faint
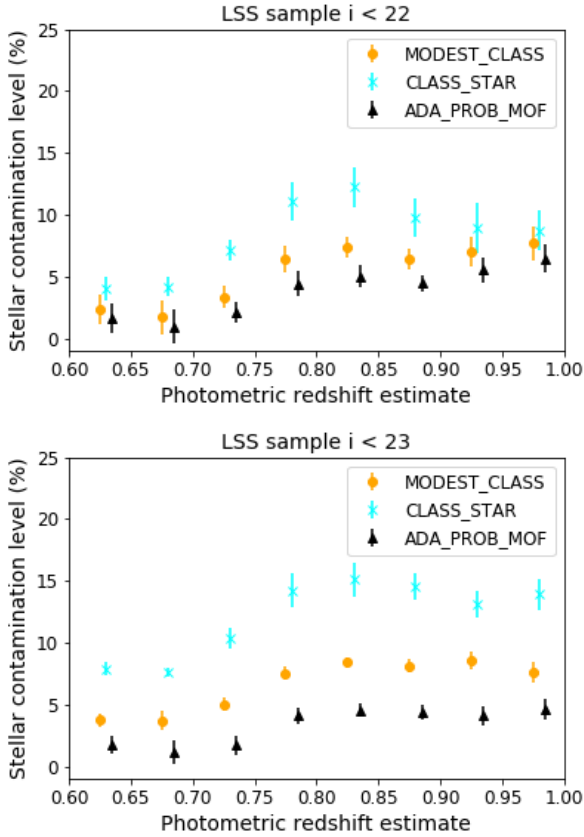
Figure 19. Stellar contamination level as a function of redshift for a bright sample (*top*, i< 22) and a faint sample (*bottom*, i< 23), derived with the method described in Section 5.4 for different samples classified by photometric redshift.
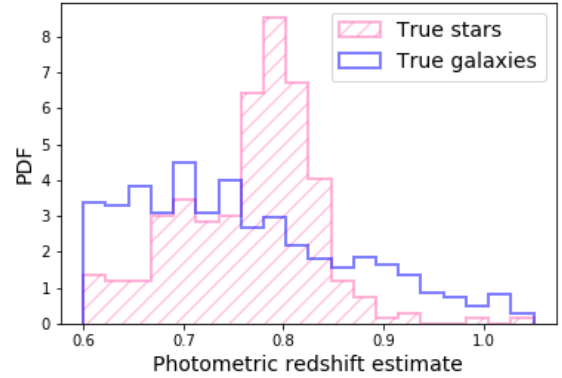
Figure 20. Normalised distribution of BPZ redshifts for a typical red galaxy sample that would be used for LSS studies, over a region with known identification of stars and galaxies through Hubble Space Telescope imaging.
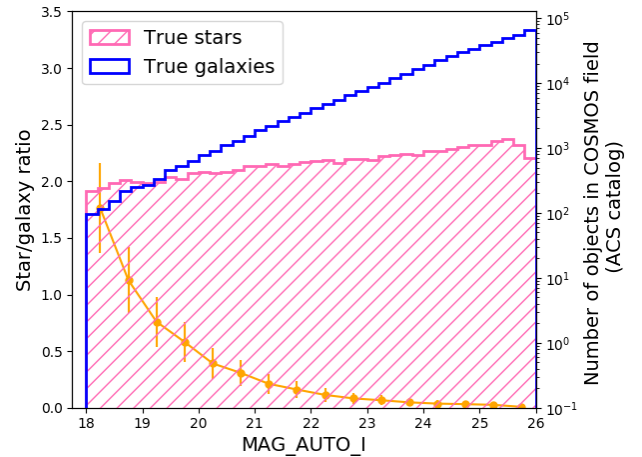

Figure 21. Star-galaxy ratio in differential `MAG_AUTO` bins, taken from the COSMOS ACS catalogue. Point-sources are overwhelmed by extended sources in the faint end.

magnitudes. Studies such as those in Fadely et al. (2012), show that currently this can become a major systematic effect in deriving the Galaxy structure. Additionally, misclassified galaxies become a limiting factor for discovering faint resolved stellar overdensities (Willman 2010; Bechtol et al. 2015; Drlica-Wagner et al. 2015; Pieres et al. 2017, e.g.,). This problem is evidenced returning to the COSMOS ACS catalogue used in Section 4, which can be used to understand the ratio of stars to galaxies up to a very faint limit (shown in Figure 21).

In this sense, the results in Pieres et al. (2017) or Shipp et al. (2018), for example, show that the very good results can be obtained based on a multi-epoch based classifier such as the weighted averaged `SPREAD_MODEL` quantity or the `MOF` pipeline.

The use of machine learning codes in this case is limited by the fact that if we want to study the distribution of specific types of stars, or search for Milky Way neighbours with a particular range of colours and magnitudes, we have to be very careful with introducing biases or complex selection functions in our application sample, much like what happens with photometric redshifts for the LSS case.

What the results of the current study show (e.g. Figure 4) is that the `MOF` technique has the potential of being the best candidate for selecting stellar candidates from its very tight morphological stellar locus and its capacity of reaching deeper into the separation of extended and point-like sources, by increasing by ∼ 20% the amount of stars in the sample for a given purity and magnitude cut

versus a 'classical' `SPREAD_MODEL` cut (in this plot, at 0.8 purity we go from 0.70 to 0.84 completeness). However, additional fine-tuning of the algorithm is needed to reach a good completeness in the bright end, where the model fit is not especially attuned to fits of stellar shapes. This is an open line of development in the algorithm in DES.

7 CONCLUSIONS

In this paper, we have compiled a wide variety of tests over a diverse array of star-galaxy classifiers for the DES Y1 dataset. These tests can be ported or used as examples for any other photometric dataset. The classifiers range from well-tested algorithms in the literature, to new developments using morphological information and/or flux information, using priors for stars/galaxies or training sets for machine learning codes based on space imaging information from the Hubble Space Telescope. We have studied their relative performance both using accurate truth information from spectro-

scopic and space imaging external datasets, and devised tests over the broad DES Y1 footprint that do not require this information. In the light of these results, we have analyzed the impact of using these algorithms on two broad science cases of interest to users of the DES data, namely, large-scale structure analyses and Milky Way studies. Star-galaxy classification remains as a non-dominant but important systematic source of error for cosmology, and very critical for Milky Way structure measurements and discoveries. These are the specific items that were highlighted in this work:

• Machine learning methods perform very well on calibration fields tests (Figures 2 to 4 and Table 5). In the application field the results are slightly better than for non-ML classification, especially in the faint end (Figure 19). Optical colour based classifiers however could potentially introduce biases in sample selection.

• Although CLASS_STAR has been used in the past to good effect, its lack of performance in the faint end (see e.g. Figures 1 and 12) leads us to recommend alternative classification methods such as SExtractor's SPREAD_MODEL or a multi-epoch fit to the shape. In this sense, using multi-epoch, multi-object fitting instead of directly using coadded information is the preferred option for object classification in optical wavelengths (as shown in Section 4).

• As has been demonstrated in the past, the addition of infrared data is very valuable, albeit limited currently by the depth and extension of such surveys (Section 4.4.2).

• Photometric redshift binning will affect stellar contamination of specific galaxy samples (Figure 20).

7.1 Expected improvements for Y3 and beyond

Considering these results, we have identified very clear future directions to expand and improve star-galaxy classification in forthcoming DES science analyses (Y3 and beyond).

• Improvement of the MOF quantities to better fit stellar shapes and prevention of fitting failures.

• Understanding the impact of using colour information on specific science cases (photo-z, stellar type selections) to ascertain whether or not the usage of this information in ML codes hampers their utility for star-galaxy separation in extragalactic and Milky Way studies respectively, in exchange of an additional 2-5% in purity depending on the case.

• The combination of information as done in Kim et al. (2015) from different approaches, especially adding external infrared colours, could greatly benefit the performance of some classifiers. Once an adequate template set is studied for the DES data, trying to overcome the impact of the lack of u-band information, template-based codes could be considered as well to complement this impact study. In addition, this would provide a truly probabilistic output that could be employed in statistical studies of large-scale structure, removing the need of having to eliminate a subsample of galaxies according to an arbitrary threshold.

• Besides VHS data, the addition of Gaia's DR2 information (Brown et al. 2018) will provide a robust and broad complement to these tests at magnitudes $r < 21$.

7.2 Ideas for further study

Finally, we call attention to other approaches and tests that we have not specifically investigated here which could be relevant for future studies:

• Adding available u-band and specially infrared band information using matched-aperture photometry as part of the algorithms used here.

• With respect to a template-fitting approach, the characteristics of this dataset (lack of u-band or infrared information), severely limit its usability. But expanding the dataset, jointly with an accurate understanding of the template range to be used can be considered as a promising approach if these requirements are met, to be used in a joint probabilistic method.

• Including very detailed image-based simulations for training, such as Balrog (Suchyta et al. 2016) or UFIG (Chang et al. 2015), to understand the failure modes of different classifiers.

• Adding seeing as part of the features of the machine learning classifiers, as well as for characterization of the performance of the different approaches.

• Usage of the object position in the sky can also provide an additional lever for a probabilistic approach, as a prior to be added to the overall posterior estimation. This should be approached with care for certain analysis (e.g. Milky Way structure).

• PSF homogenization will improve the SExtractor estimates as shown in Desai et al. (2012). However, using MOF-based photometry is a more promising alternative that avoids some of the problems associated with homogeneization.

• Convolutional Neural Networks (e.g., Kim & Brunner 2017) can be applied directly to the images to provide a new and complementary approach to ML applied at catalogue-level. Image-level analyses may benefit by using information from multiple (>10) bands (e.g., Cabayol et al. (2018)).

The data used in this paper are provided at `http://des.ncsa.illinois.edu/releases/y1a1`.

REFERENCES

Abazajian K., et al., 2004, AJ, 128, 502
Aihara H., et al., 2018, PASJ, 70, S8
Albareti F. D., et al., 2017, ApJS, 233, 25
Baldry I. K., et al., 2010, MNRAS, 404, 86
Ball N. M., Brunner R. J., Myers A. D., Tcheng D., 2006, ApJ, 650, 497
Banerji M., Abdalla F. B., Lahav O., Lin H., 2008, MNRAS, 386, 1219
Banerji M., et al., 2015, MNRAS, 446, 2523
Bechtol K., et al., 2015, ApJ, 807, 50
Benítez N., 2000, ApJ, 536, 571
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bouy H., Bertin E., Moraux E., Cuillandre J.-C., Bouvier J., Barrado D., Solano E., Bayo A., 2013, A&A, 554, A101
Bradley A. P., 1997, Pattern Recognition, 30, 1145
Brown A. G. A., et al., 2018, A&A, 616, A1
Burke D. L., et al., 2018, AJ, 155
Cabayol L., et al., 2018, preprint, p. arXiv:1806.08545 (arXiv:1806.08545)
Carnero A., Sánchez E., Crocce M., Cabré A., Gaztañaga E., 2012, MNRAS, 419, 1689
Carretero J., et al., 2017, PoS, EPS-HEP2017, 488
Chang C., et al., 2015, ApJ, 801, 73
Collister A. A., Lahav O., 2004, PASP, 116, 345
Covey K. R., et al., 2007, AJ, 134, 2398
Crocce M., et al., 2016, MNRAS, 455, 4301
Crocce M., et al., 2017, preprint, (arXiv:1712.06211)
Cross N. J. G., et al., 2012, A&A, 548, A119
DES Collaboration 2018, preprint, (arXiv:1801.03181)
DES Collaboration et al., 2016, MNRAS, 460, 1270
DES Collaboration et al., 2017, preprint, (arXiv:1708.01530)
Desai S., et al., 2012, ApJ, 757, 83
Drlica-Wagner A., et al., 2015, ApJ, 813, 109
Drlica-Wagner A., et al., 2018, ApJS, 235
Efron B., Stein C., 1981, Ann. Statist., 9, 586

Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15
Fawcett T., 2006, Pattern Recognition Letters, 27, 861
Fisher R. A., 1936, Annals of Eugenics, 7, 179
Flaugher B., et al., 2015, AJ, 150, 150
Frieman J. A., et al., 2008, AJ, 135, 338
Gaia Collaboration 2016, A&A, 595, A2
Giannantonio T., Percival W. J., 2014, MNRAS, 441, L16
Girardi L., Groenewegen M. A. T., Hatziminaoglou E., da Costa L., 2005, A&A, 436, 895
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Hambly N. C., et al., 2008, MNRAS, 384, 637
Hand D. J., 2009, Machine Learning, 77, 103
Heydon-Dumbleton N. H., Collins C. A., MacGillivray H. T., 1989, Monthly Notices of the Royal Astronomical Society, 238, 379
High F. W., Stubbs C. W., Rest A., Stalder B., Challis P., 2009, AJ, 138, 110
Hildebrandt H., et al., 2012, MNRAS, 421, 2355
Høg E., et al., 2000, A&A, 355, L27
Hogg D. W., Lang D., 2013, Publications of the Astronomical Society of the Pacific, 125, 719
Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275
Huff E., Mandelbaum R., 2017, preprint, p. arXiv:1702.02600 (arXiv:1702.02600)
Irwin M. J., et al., 2004, in Quinn P. J., Bridger A., eds, Proc. SPIEVol. 5493, Optimizing Scientific Return for Astronomy through Information Technologies. pp 411–422, doi:10.1117/12.551449
Jarvis M., et al., 2016, MNRAS, 460, 2245
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Kim E. J., Brunner R. J., Carrasco Kind M., 2015, MNRAS, 453, 507
Kovács A., Szapudi I., 2015, MNRAS, 448, 1305
Kron R. G., 1980, ApJS, 43, 305
Le Fèvre O., et al., 2013, A&A, 559, A14
Leauthaud A., et al., 2007, ApJS, 172, 219
Li T. S., et al., 2016, ApJ, 817, 135
Lilly S. J., et al., 2009, ApJS, 184, 218
MacGillivray H. T., Martin R., Pratt N. M., Reddish V. C., Seddon H., Alexander L. W. G., Walker G. S., Williams P. R., 1976, MNRAS, 176, 265
Machado E., et al., 2016, in 2016 International Joint Conference on Neural Networks (IJCNN). pp 123–130, doi:10.1109/IJCNN.2016.7727189
Małek K., et al., 2013, A&A, 557, A16
McMahon R. G., Banerji M., Gonzalez E., Koposov S. E., Bejar V. J., Lodieu N., Rebolo R., VHS Collaboration 2013, The Messenger, 154, 35
Molino A., et al., 2014, MNRAS, 441, 2891
Morganson E., et al., 2018, Publications of the Astronomical Society of the Pacific, 130, 074501
Neilsen E., Bernstein G., Gruendl R., Kent S., 2016, Fermilab report, FERMILAB-TM-2610-AE-CD
Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A., 1992, AJ, 103, 318
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Pieres A., et al., 2017, MNRAS, 468, 1349
Robin A. C., Reylé C., Derrière S., Picaud S., 2003, A&A, 409, 523
Ross A. J., et al., 2011, MNRAS, 417, 1350
Ruhl J., et al., 2004, in Bradford C. M., et al., eds, Proc. SPIEVol. 5498, Z-Spec: a broadband millimeter-wave grating spectrometer: design, construction, and first cryogenic measurements. pp 11–29 (arXiv:astro-ph/0411122), doi:10.1117/12.552473
Saglia R. P., et al., 2012, ApJ, 746, 128
Sevilla-Noarbe I., Etayo-Sotos P., 2015, Astronomy and Computing, 11, 64
Sharma S., Bland-Hawthorn J., Johnston K. V., Binney J., 2011, ApJ, 730, 3
Sheldon E. S., Huff E. M., 2017, ApJ, 841, 24
Shipp N., et al., 2018, ApJ, 862, 114
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Soumagnac M. T., et al., 2015, MNRAS, 450, 666
Stoughton C., et al., 2002, AJ, 123, 485
Suchyta E., et al., 2016, MNRAS, 457, 786
Tasca L. A. M., et al., 2017, A&A, 600, A110
Tie S. S., et al., 2017, AJ, 153, 107
Tucker D. L., et al., 2007, in Sterken C., ed., Astronomical Society of the Pacific Conference Series Vol. 364, The Future of Photometric, Spectrophotometric and Polarimetric Standardization. p. 187 (arXiv:astro-ph/0611137)
Werbos P. J., 1982, in Drenick R. F., Kozin F., eds, System Modeling and Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 762–770
Whitmore B. C., et al., 2016, AJ, 151, 134
Willman B., 2010, Advances in Astronomy, 2010, 285454
Wright E. L., et al., 2010, AJ, 140, 1868
Yee H. K. C., 1991, PASP, 103, 396
York D. G., et al., 2000, AJ, 120, 1579
Zuntz J., et al., 2017, preprint, (arXiv:1708.01533)

APPENDIX A: `ADA_PROB` TECHNICAL DETAILS

This appendix describes the details of one of the machine learning frameworks called `ADA_PROB`.

The framework first selects an exhaustive list of photometric properties, or features, and generates linear combinations of these features to produced new features. This may include unphysical combinations, such as magnitudes and radii being combined. We also generate features 'intelligently', by using the current state of the art. For the problem of star-galaxy separation for DES, this means including both a binary `MODEST_CLASS` class value, and a continuous `MODEST_CLASS` variable for both stars and galaxies.

Next, the enormous feature list is sorted by rank, using the value of the mutual information[14], which is a non-linear correlation coefficient, between the selected feature and the target class. Finally the top 150 features are selected to form the inputs to the machine learning algorithms.

The framework then explores many machine learning algorithms, each of which are trained with random variations of each of their own hyper-parameters. The framework explores a plethora of algorithms, drawn from the sci-kit-learn (Pedregosa et al. 2011) package. These include AdaBoost, which often performs well, and also Random Forests, Extra Randomised Trees, Quadratic Discriminant Analysis and the K-Nearest Neighbours Classifier.

The performance of each selected algorithm and set of hyper-parameters is quantified by measuring the average F_1 score on 30 held out samples during 30 fold cross validation. The F_1 score is the geometric mean between the precision and the recall, and 30 fold cross validation is akin to making 30 jackknife samples of the data, training on all but the held out sample, and then making predictions on that held out sample, and then repeating. The held out jackknife results, or "class weights", for each training object are retained for future classification calibration.

The winning algorithm and hyper-parameter set is then retrained on the full training sample. The training procedure is deemed to have been completed once at least 50 systems have been explored and when the F_1 score has not been improved upon after 20 iterations. In our empirical experience, we find this to be a generally stable point at which one can stop the exploration of the different algorithms, hyper-parameters, and move on to the final stage of the framework.

This final stage then uses isotonic regression to calibrate the

[14] https://en.wikipedia.org/wiki/Mutual_information

held out class weights of the training data. This enforces the statistical properties of the class weights to more closely resemble a probability. This rescaling is performed by comparing the total number of those objects within a class weight bin, with the fraction of objects to have the true class value. This comparison leads to a rescaling of class weights to class probabilities which we note are conditional on the training data.

The winning machine learning algorithm, which happened to be AdaBoost in this case, is then used to make class weight predictions on both the test sample and the science samples, and their output class weights are scaled using the previously learned rescaling, to make them more closely resemble probabilities.

We can also perform a feature importance analysis (see, e.g., Hoyle et al. 2015) which suggests that the features with the most predictive power are indeed those derived from MOD-EST, with other ranking features being `WAVG_SPREAD_MODEL_R` and `MAGERR_MODEL_I`.

APPENDIX B: EXTERNAL DATASETS

B1 Access to external catalogues used in this work

The catalogues used in Section 4 and listed on Table 1 can be obtained from the following website: `http://des.ncsa.illinois.edu/releases/y1a1`.

B2 Queries used to extract the datasets

Query to the SDSS CASJOBS interface (used as imaging truth table for some tests) to obtain 2MASS, WISE matches with SDSS data, to match with DES data on same area.

```
SELECT
  s.ra, s.dec, s.dered_r,
  w.w1mpro as w1, w.j_m_2mass as j, s.z,
  s.class
INTO
  mydb.stripe82_wise_2mass_z_match
FROM
  wise_xmatch as xm
JOIN
  specPhoto as s on xm.sdss_objid = s.objid
JOIN
  wise_allsky as w on xm.wise_cntr = w.cntr
WHERE
  ((s.dered_g < 23.0) or (s.dered_r < 23.0)
  or (s.dered_i < 23.0)) and
  ((s.ra > 0 and s.ra < 5 and s.dec > -2.5 and
  s.dec < 3.5) or  (s.ra > 315 and s.dec > -3
  and s.dec < 3)) and s.zWarning = 0
  and s.zErr < 0.001
```

Query to the SDSS CASJOBS interface (used as imaging truth table for some tests) to match with DES and VHS data on same area.

```
SELECT
  s.ra, s.dec, s.dered_r, s.z, s.class
INTO
  mydb.stripe82_z_dr13
FROM
  specPhoto as s
WHERE
```

```
  ((s.ra > 0 and s.ra < 5 and
  s.dec > -2.5 and s.dec < 3.5) or
    (s.ra > 315 and s.dec > -3 and
    s.dec < 3)) and s.zWarning = 0
  and s.zErr < 0.001
```

Query to the Hubble Source Catalog CASJOBS interface (used as imaging truth table for some tests):

```
SELECT
    p.MatchRA, p.MatchDEC, p.MatchID as hscv2_id,
    p.CI, p.CI_Sigma, m.A_F814W, m.A_F814W_Sigma
INTO
    hsc_source_catalog
FROM
    SumPropMagAutoCat p
JOIN
    SumMagAutoCat m ON p.MatchID = m.MatchID
WHERE
    m.A_F814W > 0 and m.A_F814W_Sigma is not null
    and p.numimages > 2
```

Query to the VISTA Science Archive, using the VHSDR3 database.

```
SELECT     ra, dec, jpetromag, jpetromagerr,
jmksext, jmksexterr
FROM
    vhsSource
WHERE
    jerrbits = 0 and kserrbits = 0 and
    (priOrSec=0 OR priOrSec=frameSetID) and
    dec between -2 and 2 and (ra > 315 or ra < 5)
```

Query to Gaia's DR2, using the CosmoHub (Carretero et al. 2017) interface.

```
SELECT 'ra', 'dec', 'phot_g_mean_mag', 'l', 'b',
'phot_g_mean_flux_over_error',
'astrometric_primary_flag'
FROM gaia_dr2
WHERE
(('ra' > 305) or ('ra' < 90)) and ('dec' > -61)
and ('dec' < -35) and phot_g_mean_mag > 18.5
```

AFFILIATIONS

[1] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[2] Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, 85748 Garching, Germany
[3] Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, 81679 München, Germany
[4] Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK
[5] Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Rehovot 76100, Israel
[6] LSST, 933 North Cherry Avenue, Tucson, AZ 85721, USA
[7] Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA
[8] Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa
[9] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra

(Barcelona), Spain

[10]Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA

[11] Department of Physics, University of Surrey, Guildford GU2 7XH, UK

[12] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

[13] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

[14] CNRS, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France

[15] Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France

[16] Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801, USA

[17] National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA

[18] Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA

[19] Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA

[20] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

[21]Instituto de Física, UFRGS, Caixa Postal 15051, Porto Alegre, RS - 91501-970, Brazil

[22]Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil

[23]Brookhaven National Laboratory, Bldg 510, Upton, NY 11973, USA

[24]Department of Physics, University of Chicago, Chicago, Illinois 60637, USA [25]Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, Casilla 603, La Serena, Chile

[26]Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil

[27]Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India

[28]Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain

[29]Institut d'Estudis Espacials de Catalunya (IEEC), 08193 Barcelona, Spain

[30]Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain

[31]Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA

[32]Department of Physics, The Ohio State University, Columbus, OH 43210, USA

[33]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA

[34]Department of Astronomy/Steward Observatory, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA

[35]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

[36]Australian Astronomical Observatory, North Ryde, NSW 2113, Australia

[37]Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP, 05314-970, Brazil

[38]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

[39]Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain

[40]Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA

[41]School of Physics and Astronomy, University of Southampton, Southampton, SO17 1BJ, UK

[42]Brandeis University, Physics Department, 415 South Street, Waltham MA 02453, USA

[43]Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, 13083-859, Campinas, SP, Brazil

[44]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[45]Institute of Cosmology & Gravitation, University of Portsmouth, Portsmouth, PO1 3FX, UK