

Star-galaxy classification in the Dark Energy Survey Y1 dataset

I. Sevilla-Noarbe^{1*}, B. Hoyle^{2,3}, M.J. Marchã⁴, M.T. Soumagnac⁵, K. Bechtol⁶, A. Drlica-Wagner⁷, F. Abdalla^{4,8}, J. Aleksić⁹, C. Avestruz¹⁰, E. Balbinot¹¹, M. Banerji^{12,13}, E. Bertin^{14,15}, C. Bonnett⁹, R. Brunner¹⁶, M. Carrasco-Kind¹⁷, A. Choi¹⁸, T. Giannantonio^{12,13,2}, E. Kim¹⁷, O. Lahav⁴, B. Moraes⁴, B. Nord⁷, A.J. Ross¹⁸, E.S. Rykoff^{19,20}, B. Santiago^{21,22}, E. Sheldon²³, K. Wei^{10,24}, W. Wester⁷, B. Yanny⁷, T. Abbott²⁵, S. Allam⁷, D. Brooks⁴, A. Carnero-Rosell^{22,26}, J. Carretero⁹, C. Cunha¹⁹, L. da Costa^{22,26}, C. Davis¹⁹, J. de Vicente¹, S. Desai²⁷, P. Doel⁴, E. Fernandez⁹, B. Flaugher⁷, J. Frieman^{7,10}, J. Garcia-Bellido²⁸, E. Gaztanaga^{29,30}, D. Gruen^{19,20}, R. Gruendl^{16,17}, J. Gschwend^{22,26}, G. Gutierrez⁷, D.L. Hollowood³¹, K. Honscheid^{18,32}, D. James³³, T. Jeltema³¹, D. Kirk⁴, E. Krause^{34,35}, K. Kuehn³⁶, T. S. Li^{7,10}, M. Lima^{37,22}, M. A. G. Maia^{22,26}, M. March³⁸, R. G. McMahon^{4,8}, F. Menanteau^{16,17}, R. Miquel^{39,9}, R. L. C. Ogando^{22,26}, A. A. Plazas³⁵, E. Sanchez¹, V. Scarpine⁷, R. Schindler²⁰, M. Schubnell⁴⁰, M. Smith⁴¹, R. C. Smith²⁵, M. Soares-Santos⁴², F. Sobreira^{43,22}, E. Suchyta⁴⁴, M. E. C. Swanson¹⁷, G. Tarle⁴⁰, D. Thomas⁴⁵, D. L. Tucker⁷, A. R. Walker²⁵

(The DES Collaboration)

Author affiliations are listed at the end of this paper.

Accepted XXX. Received YYY; in original form ZZZ

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 catalogs used in this work will be made available upon publication. .

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). 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 catalogs 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. Mis-classification 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 have incorporated color information into their classifiers (see Ball et al. 2006 for SDSS, 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 (Malek 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 catalog, the contamination of QSOs in the large-scale stellar and galactic catalogs 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 (Morganson 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 SV¹ and Y1 (Drlica-Wagner et al. 2018) comprising the “Gold” catalog. We note the following features that are relevant for the present study:

- The object catalogs 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* band². 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 catalogs. **MAG_MODEL**, **MAG_DETMODEL**³ and **MAG_PSF** are used as inputs for the machine learning methods as well.
- In addition, a multi-object, multi-epoch fitting pipeline (MOF) has been run on the single-epoch image counterparts for each coadd catalog detection to obtain more precise photometric measurements for the objects, simultaneously fitting the individual images and modeling light from nearby neighbors for each object (more details in Drlica-Wagner et al. 2018).
- 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

¹ <https://des.ncsa.illinois.edu/releases/sva1>

² This calibration approach was eventually superseded in Y3 data products with the Forward Global Photometric Calibration approach described in Burke et al. (2018)

³ In this case the exponential model used in **SExtractor** is fitted on the detection image and scaled in the measurement images of each band.

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 field⁴ has been imaged with DECam, providing a very useful dataset given the richness of multi-band imaging and spectroscopy available. Table 1 summarizes 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 Telescope⁵ (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 summarizes 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 color 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:

- **CLASS_STAR** (Bertin & Arnouts 1996) 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. In particular, it uses a backpropagation model (Werbos 1982) for learning, and bases its training on eight isophotal areas above the background, plus the value of the intensity at the peak pixel in the object and the value of the FWHM for the image. The simulations include a wide range of PSF profiles and sizes, though they are optimized 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.

- **SPREAD_MODEL** 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). **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 normalized to allow comparing sources with different PSFs throughout the field:

$$\text{SPREAD_MODEL} = \frac{\vec{G}^T \mathbf{W} \vec{p}}{\vec{\phi}^T \mathbf{W} \vec{p}} - \frac{\vec{G}^T \mathbf{W} \vec{\phi}}{\vec{\phi}^T \mathbf{W} \vec{\phi}}, \quad (1)$$

where \vec{p} is the image vector centered on the source⁶. \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:

$$\begin{aligned} \text{SPREADERR_MODEL} &= \frac{1}{(\vec{\phi}^T \mathbf{W} \vec{p})^2} \left(\vec{G}^T \mathbf{V} \vec{G} (\vec{\phi}^T \mathbf{W} \vec{p})^2 \right. \\ &\quad \left. + \vec{\phi}^T \mathbf{V} \vec{\phi} (\vec{G}^T \mathbf{W} \vec{p})^2 \right. \\ &\quad \left. - 2 \vec{G}^T \mathbf{V} \vec{\phi} (\vec{G}^T \mathbf{W} \vec{p} \vec{\phi}^T \mathbf{W} \vec{p}) \right)^{1/2} \quad (2) \end{aligned}$$

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 catalog, which includes the following criteria:

$$\begin{aligned} \text{galaxies} &\iff \\ &\text{SPREAD_MODEL_I+} \\ &(5/3) \times \text{SPREADERR_MODEL_I} > 0.005 \\ &\text{AND NOT} \\ &(|\text{WAVG_SPREAD_MODEL_I}| < 0.002 \\ &\text{AND} \\ &\text{MAG_AUTO_I} < 21.5) \end{aligned} \quad (3)$$

$$\begin{aligned} \text{stars} &\iff \\ &|\text{SPREAD_MODEL_I+} \\ &(5/3) \times \text{SPREADERR_MODEL_I}| < 0.002 \end{aligned} \quad (4)$$

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

⁴ <http://cosmos.astro.caltech.edu/>

⁵ <https://pole.uchicago.edu/>

⁶ 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. More details are provided in Appendix B.

Catalog	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	18985/53345	Albaret et al. (2017)
VVDS	Ground optical spectroscopy	Truth table	4442	Le Fèvre et al. (2013)
WISE	Space NIR imaging	Complementary data	18985	Wright et al. (2010)
2MASS	Ground NIR imaging	Complementary data	18985	Skrutskie et al. (2006)
VHS	Ground NIR imaging	Complementary data	53345	McMahon et al. (2013)

Table 2. Summary of classification methods.

Name	Type of data used	Description
CLASS_STAR	Morphological	Neural Network on isophotal measurements
SPREAD_MODEL	Morphological	Normalized Linear Discriminant
CM_T	Morphological	Multi epoch fitting to shape
MCAL_RATIO	Morphological	Metacalibration ratio of object size over PSF size
ADA_PROB	Most discriminating features	Boosted Decision Trees
GALSIFT_PROB	All catalog columns (PCA)	Random Forests
SVM	Morphological	Support Vector Machine on a limited column set
CONCENTRATION	Morphological	Light concentration
W1-J, J-K	Fluxes	Infrared color discrimination
HB_PROB	Fluxes	Template fitting of spectra

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 both or an ‘artifact’ category (approximately 5% of the catalog considered here).

- **CM_T** is an intrinsic size estimator for the object from the image fitting provided by the MOF pipeline. This fitting tool estimates the shapes and fluxes of objects detected in the coadd catalogs, using a mixture of Gaussians^{7,8} 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 neighbors accurately. Concretely, **CM_T** is really a ‘size squared’ defined as:

$$\text{CM_T} = \langle x^2 \rangle + \langle y^2 \rangle \quad (5)$$

where x and y denote the distance from the object’s center. 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 tested⁹ in the COSMOS field, is based on the quantity $\text{CM_T} + 2 \times \text{CM_T_ERR}$. Typical values are in the range between -0.5 and 0.5.

- A **CONCENTRATION** parameter similar to what was used as a star-galaxy classifier for SDSS (Abazajian et al. 2004). 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 ($\text{PSF_MAG_I} - \text{CM_MAG_I}$).

- **MCAL_RATIO** is derived from the measurements of the object size and PSF model size obtained using the *metacalibration* technique, developed for shear measurement in weak lensing studies (Sheldon & Huff 2017). This uses the same **ngmix** code as MOF above. This measurement is much noisier as the metacalibration technique 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).

$$\text{MCAL_RATIO} = \frac{T_{mcal}}{T_{PSF}} \quad (6)$$

where T_{mcal} and T_{PSF} are sizes of the object or PSF respectively as defined in 5. Values are not constrained, but typical ranges explored for star-galaxy separation are between 0 and 1.

- **ADA_PROB** is the name given to a machine learning framework which combines feature generation, and feature pre-selection with machine learning algorithms (including AdaBoost) drawn from the **scikit-learn** package (Pedregosa et al. 2011), and a probability calibration. 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**.

- **GALSIFT_PROB** is the probabilistic estimate provided by **Galsift** (‘Multi_class’ in Soumagnac et al. (2015)). It is a multi-parameter classifier that consists of three steps:

- Step 1 - A principal component analysis (PCA) to outline the correlations between the object parameters and extract the most relevant information.
- Step 2 - Calculation of the Fisher discriminant (Fisher 1936) for each of the new parameters to quantify

⁷ <https://github.com/esheldon/ngmix>

⁸ <https://github.com/esheldon/ngmixer>

⁹ Technically, a different, *validation* set would be required to tune this classifier in terms of the quantity multiplying **CM_T_ERR**, to avoid bias towards a specific value, though in practice the differences are small between different choices.

their aptitude to separate between the classes.

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

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

– Step 3 - Selection of the parameters with the highest Fisher discriminant (hence the highest ‘separation power’ of the classes) and using 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`.

- **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 **W1-J** from WISE (Wright et al. 2010), 2MASS (Skrutskie et al. 2006), as described in Kovács & Szapudi (2015).

- **SVM**, support vector machine, is a supervised machine learning algorithm that constructs a separating hyperplane in any arbitrary n -dimensional space that maximizes the margins of objects to the hyperplane. Following Wei et al. (in prep), the **SVM** is a single-band, purely morphological and magnitude based classifier. The only input features used by the **SVM** are `MAG_AUTO_I`, `FLUX_RADIUS_I`, and `SPREAD_MODEL_I`. 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. 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.

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:

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

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*):

$$PPV = \frac{TP}{TP + FP} \quad (10)$$

‘Completeness’ (also known as, *recall*) is another name for the TPR defined in Equation 8. A good approach to

Table 3. Selection methods.

Name	Selection method for galaxies using threshold t
CLASS_STAR	$CLASS_STAR < t$
SPREAD_MODEL	$SPREAD_MODEL + 1.67 * SPREADERR_MODEL > t$
CMT	$CMT + 2 * CMT_ERR > t$
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	$(J - K - 0.6)/(MAG_AUTO_G - MAG_AUTO_I) > t$

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 emphasized in bold.

		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)		

easily compare the performances of several classifiers is to use the precision-recall (PR) curve, where both quantities can be visualized simultaneously.

4.3 Training and testing fields

The dataset on which we train the machine learning (ML) codes is the weak lensing catalog 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 ~ 114k extended and ~ 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 Catalog¹⁰ (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 catalogs’ 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 catalog, 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. We also constrain ourselves to showing the ROC curves which are more insensitive to these possible biases. The SDSS 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.

Both the COSMOS catalogs and the ones recovered from the Hubble-SC have been cross-tested against spectroscopic catalogs (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, whereas around

¹⁰ <https://archive.stsci.edu/hst/hsc/>

2% of spectroscopically classified galaxies are misclassified as stars. This happens at F814W magnitudes from the ACS instrument greater than 24 (for COSMOS) and 23 (for the miscellaneous Hubble-SC field), respectively, for each catalog, denoting compact galaxies that are unresolvable by HST. 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 catalog. However, this approach could provide a middle-ground solution to the issues one might encounter when incorporating color-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 behavior 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 addi-

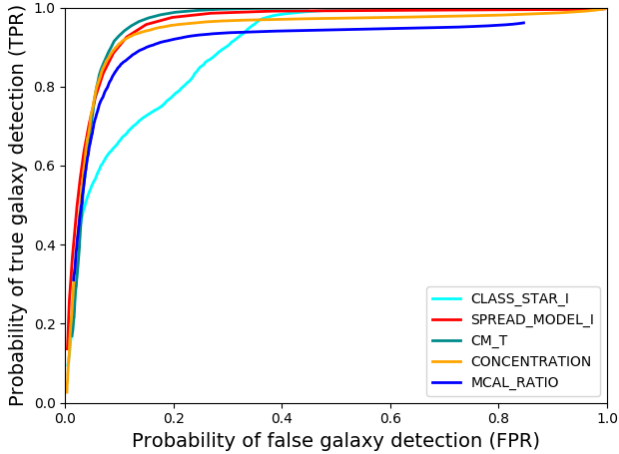


Figure 1. ROC plot for classifiers tested on the COSMOS field. Only non-ML codes are shown, as they are trained in this dataset. Magnitude range is given by $MAG_AUTO_I = (17,24)$. The SPREAD_MODEL-based cut is similar to MODEST_CLASS used in Y1 analyses. Galaxies are assumed to be the ‘truth’ for the purposes of the ROC definition.

tional 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).

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 color cuts on the other hand, shows an important increase in performance, specifically when attempting to select a very pure stellar sample,

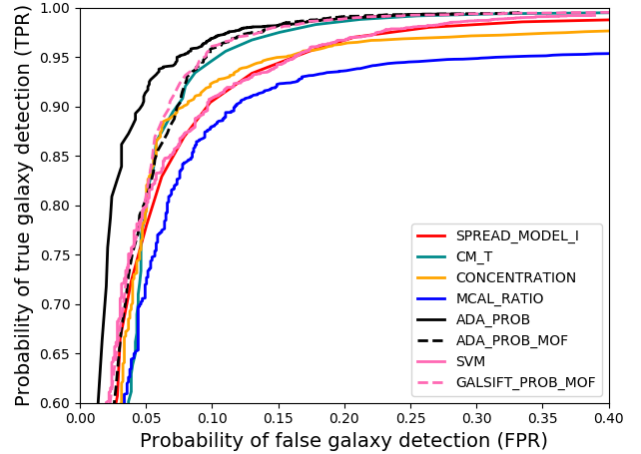


Figure 2. ROC plot for classifiers tested on the SN fields over the Hubble-SC catalog. Magnitude range is given by $MAG_AUTO_I = (17,24)$. The SPREAD_MODEL-based cut is similar to MODEST_CLASS used in Y1 analyses. Galaxies are assumed to be the ‘truth’ for the purposes of the ROC definition.

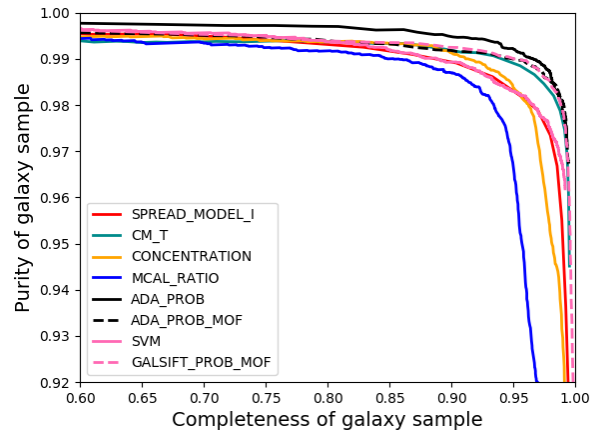


Figure 3. Precision-Recall (or completeness-purity) plot for classifiers tested on the SN fields over the Hubble-SC catalog, using galaxies as truth. Magnitude range is given by $MAG_AUTO_I = (17,24)$. The SPREAD_MODEL-based cut is similar to the MODEST_CLASS used in DES Y1 analyses.

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 extragalactic) 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 generally) population which is truly what we intend. 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

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	0.969
SPREAD_MODEL	0.954	0.956	0.975	0.952
CM_T (MOF)	0.957	0.959	0.971	0.965
CONCENTRATION (MOF)	0.938	0.950	0.959	0.937
MCAL_RATIO	0.910	0.924	–	–
VHS J-K vs G-I	–	–	–	0.990
ADA_PROB	–	0.978	0.983	0.955
ADA_PROB (MOF)	–	0.965	0.980	0.956
GALSIFT_PROB (MOF)	–	0.966	0.981	0.958
SVM	–	0.962	–	–

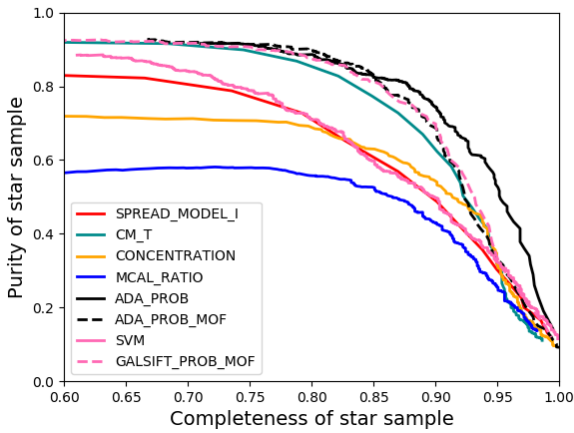


Figure 4. Precision-Recall (or completeness-purity) plot for classifiers tested on the SN fields over the Hubble-SC catalog, using stars as truth. Magnitude range is given by $\text{MAG_AUTO_I} = (17, 24)$. The SPREAD_MODEL-based cut is similar to the MODEST_CLASS used in DES Y1 analyses.

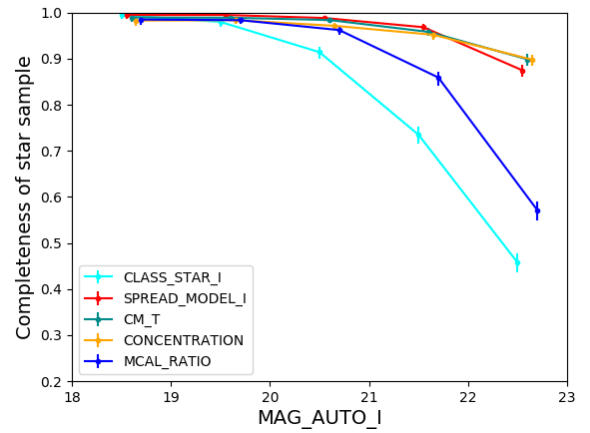


Figure 6. Completeness of **stellar** sample as a function of magnitude for classifiers tested on the COSMOS field, for a fixed 80% purity.

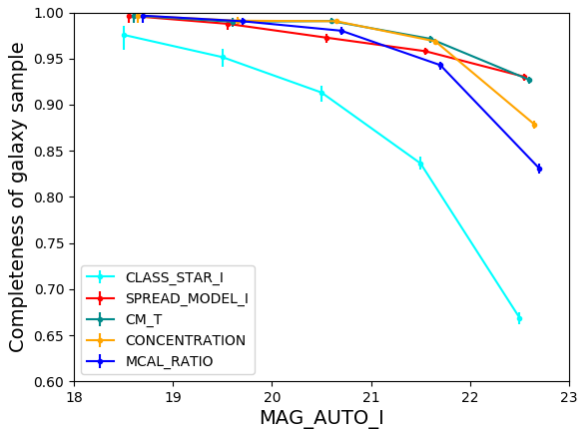


Figure 5. Completeness of a **galaxy** sample as a function of magnitude for classifiers tested on the COSMOS field, for a fixed galaxy purity of 99%.

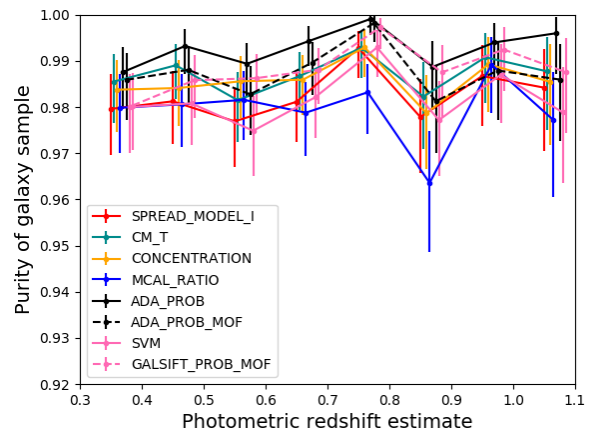


Figure 7. Purity of the galaxy sample as a function of photo- z for classifiers tested on Hubble-SC matches over the SN fields field, for a fixed 90% completeness. We use a random MonteCarlo sampling of the probability distribution function of redshift predicted by BPZ for that particular object as an estimate of its photo- z .

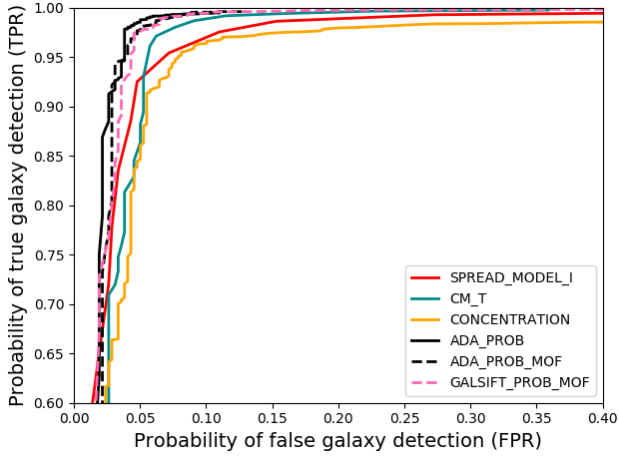


Figure 8. ROC plot for classifiers tested on the SN fields over the VVDS catalogs. Magnitude range is given by $\text{MAG_AUTO_I} = (17,24)$. Galaxies are assumed to be the ‘truth’ for the purposes of the ROC definition.

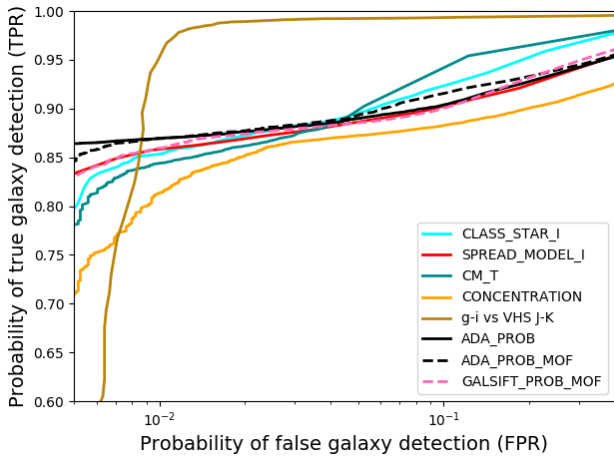


Figure 9. ROC plot for classifiers tested on the Stripe 82 region overlapping SDSS and VHS data. Magnitude range is given by $\text{MAG_AUTO_I} = (17,21)$. Note the logarithmic scale in the x-axis in this instance. Galaxies are assumed to be the ‘truth’ for the purposes of the ROC definition.

full breadth and depth of the survey and a careful combined catalog 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 catalogs 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 forced 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).

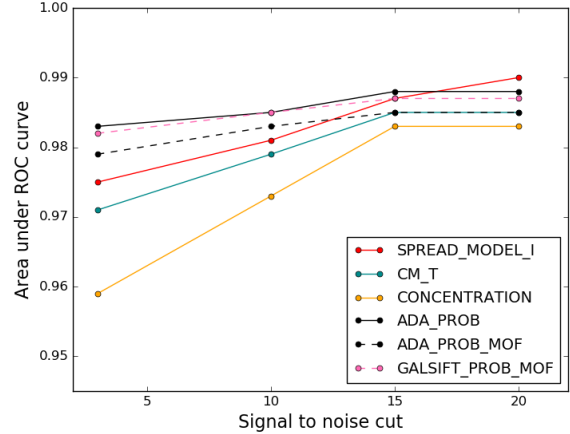


Figure 10. Area under the curve measured for the same classifiers as Figure 8, for different signal-to-noise thresholds.

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 simulation, 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 catalog. 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 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), Kuposov 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 behavior for most of them with respect to the simulation is similar. Two machine learning classifiers based on MOF quantities show a significant lack of bright objects ($r < 19$) due to failures from the Y1 version of the MOF pipeline in fitting stars in this regime¹¹. 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 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

¹¹ Y3 Gold MOF photometry has solved this issue.

same range, without any preferred or outstanding behavior 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 catalog). This is done via the pixelization 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 pixelization parameter `NSIDE=512`, which corresponds to a pixel size of approximately 0.01 square degrees.

In Figure 14 and Figure 15 we show a comparison of several classifiers, tested on the application field for the galaxy sample with the magnitude cuts shown in Table 6. The intercept value of the linear fits can be used to estimate the purity of the galaxy sample (actually, a combination of the galaxy sample purity and the stellar obscuration effect, by which stars of moderate brightness block the fainter galaxies around them). 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 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 color 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 $\sim 1\%$ systematic uncertainty in the estimation of the impurity derived from comparing brighter and fainter stellar samples (Figure 16). 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 Gaia Collaboration et al. (2018).

5.5 Galaxy ratio near the Large Magellanic Cloud

Using the same pixelization 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 center of the Large Magellanic Cloud (LMC), set at $(\alpha, \delta) = (5^{\text{h}}23^{\text{m}}34.5^{\text{s}}, -69^{\circ}45'11'')$. This value is normalized to one at 30 degrees from the center of the LMC (Figure 17). Here

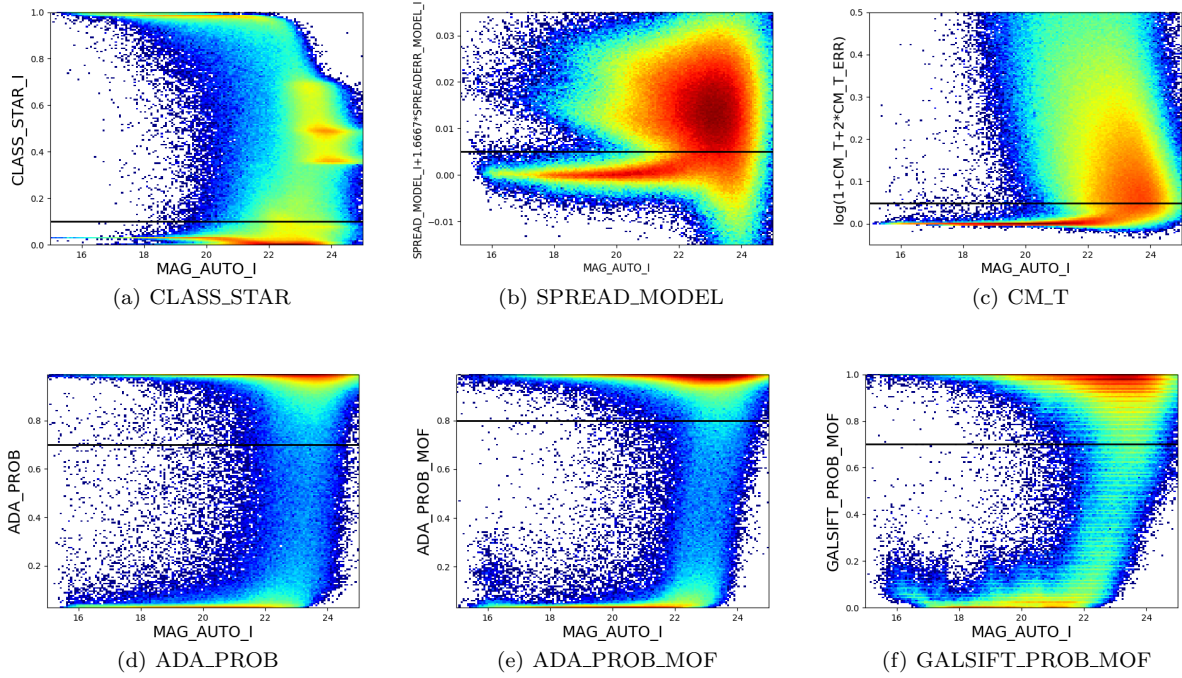


Figure 11. Object classification heatmaps as a function of magnitude for different classifiers. The black line represents the cut for which a 99% galaxy purity is obtained in the Hubble-SC sample in the $i=(17,24)$ magnitude range. With the exception of CLASS_STAR, all classifiers assign higher values to extended sources.

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
$i < 22$	$1.4 \pm 0.6\%$	$2.8 \pm 0.2\%$	$1.0 \pm 0.5\%$	$0.9 \pm 0.5\%$	$0.8 \pm 0.6\%$	$0.8 \pm 0.6\%$
$i < 23$	$1.7 \pm 0.7\%$	$5.1 \pm 2.8\%$	$0.6 \pm 0.7\%$	$0.2 \pm 0.7\%$	$1.2 \pm 0.6\%$	$0.8 \pm 0.6\%$

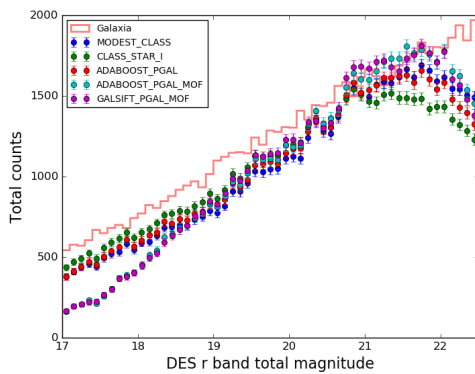


Figure 12. Counts for stars as classified by different algorithms compared to a Galaxia simulation (Sharma et al. 2011) using DES photometry, in the patch of the Y1 DES footprint with $45 < RA < 50$, $-45 < DEC < -50$.

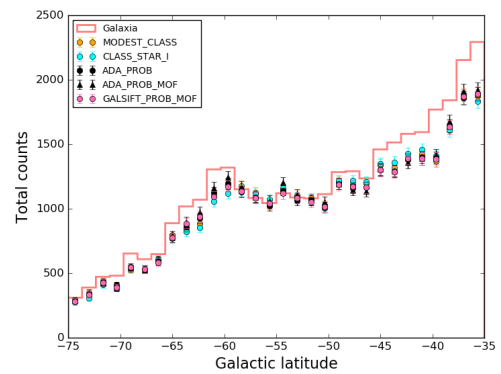


Figure 13. Counts for stars as classified by different algorithms compared to a Galaxia simulation (Sharma et al. 2011) for the application field of the DES-Y1 footprint for the magnitude range $r = (19,21.5)$.

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 color), 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

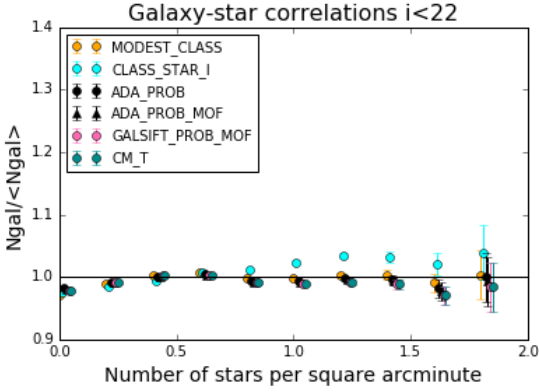


Figure 14. Galaxy vs star density plot for several classifiers, $i < 22$.

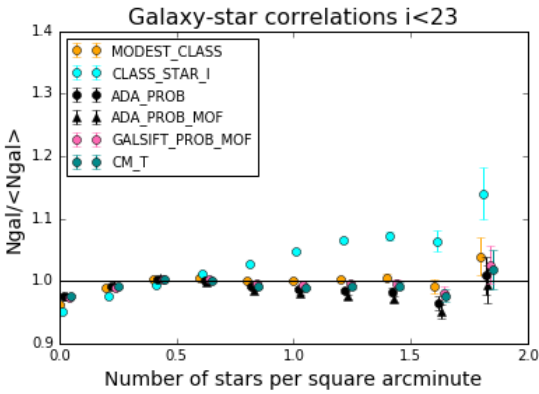


Figure 15. Galaxy vs star density plot for several classifiers, $i < 23$.

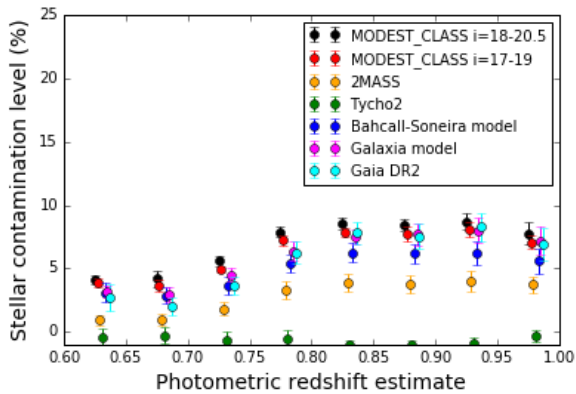


Figure 16. Star contamination levels for different stellar maps. A $\sim 1\%$ systematic uncertainty is derived by comparing the MOD-EST_CLASS moderate to bright stars, is estimated from this plot. Tycho and 2MASS stars are added for comparison, but their magnitude ranges (much brighter than the stellar sample considered as contaminants) do not make them good candidates for deriving this uncertainty.

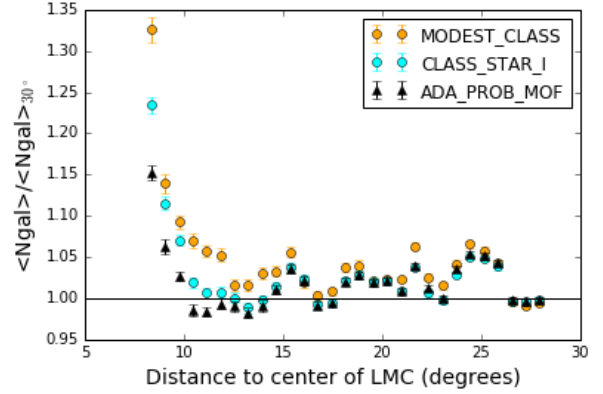


Figure 17. Galaxy ratio (with respect to galaxy density at 30 degrees from LMC) for as a function of angular distance from the LMC center.

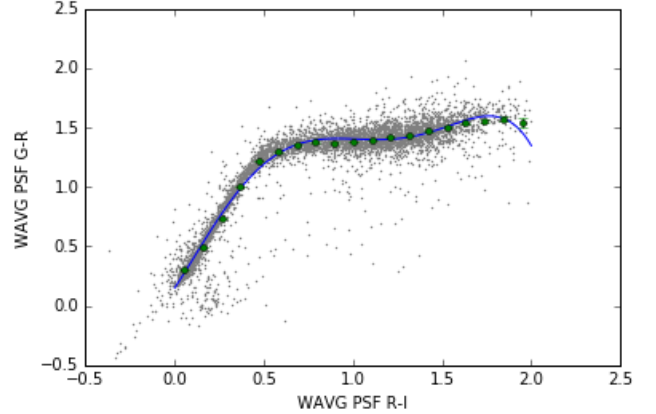


Figure 18. Fit to stellar locus using a fifth degree polynomial

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. $g - r$ color 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 18, similarly to what is realized in Covey et al. (2007). The same fit curve in Figure 18 is shown again versus several classifiers in Figures 19 and 20. 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

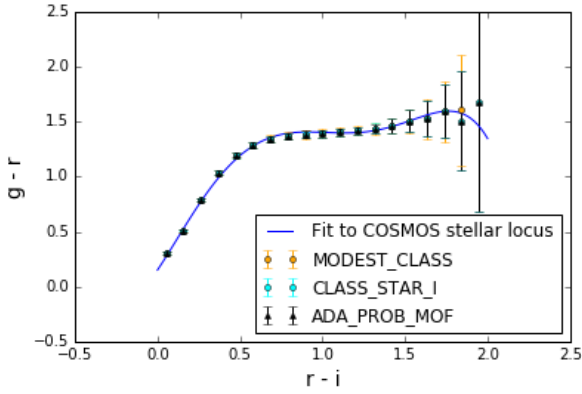


Figure 19. Stellar locus for star samples from various classifiers ($i < 21$)

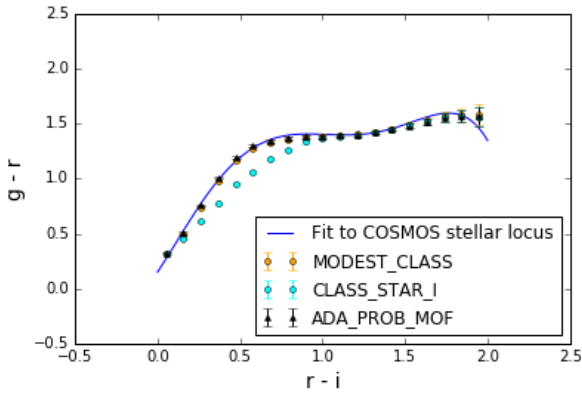


Figure 20. Stellar locus for star samples from various classifiers ($i < 24$)

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 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 catalogs, 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

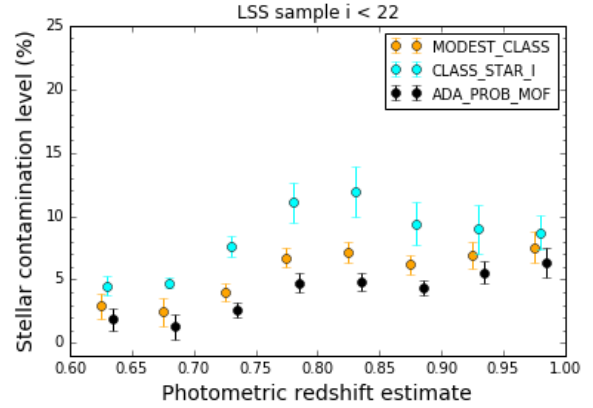


Figure 21. Stellar contamination level as a function of redshift: $i < 22$

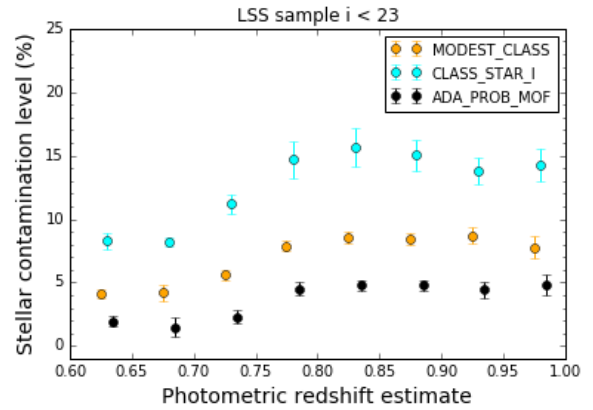


Figure 22. Stellar contamination level as a function of redshift: $i < 23$

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.

Figures 14 and 15 imply 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 colors, 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 21 and Figure 22.

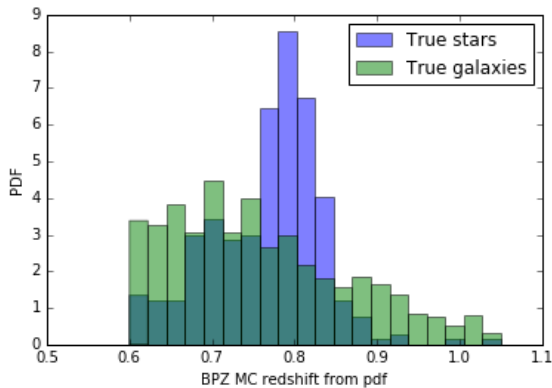


Figure 23. Normalized distribution of BPZ redshifts for a typical red galaxy sample that would be used for LSS studies, over a region with known identification of stars and galaxies through Hubble Space Telescope imaging.

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 23 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 color 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 16 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 magnitudes. Studies such as those in Fadelly 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

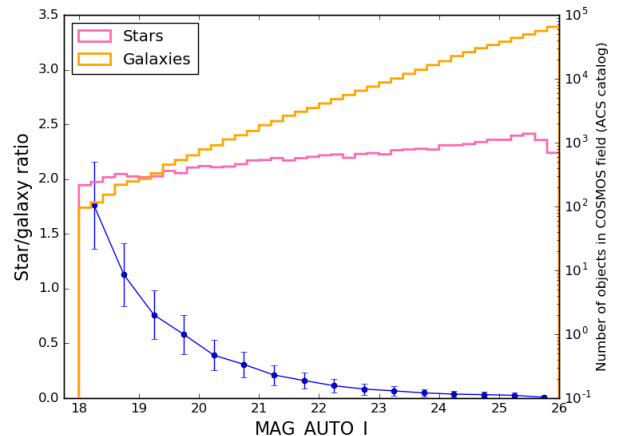


Figure 24. Star-galaxy ratio in differential `MAG_AUTO` bins, taken from the COSMOS ACS catalog. Point-sources are overwhelmed by extended sources in the faint end.

the COSMOS ACS catalog 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 24).

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 neighbors with a particular range of colors 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 spectroscopic 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 22). Optical color 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 23).

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 color 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 colors, 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 (Gaia Collaboration et al. 2016, 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 infrared band information as part of the ML 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 homogenization.

- 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 catalog-level. Image-level analyses may benefit by using information from multiple (>10) bands (e.g., Cabayol et al. in prep.).

Reference catalogs used in this work will be made available upon publication. .

ACKNOWLEDGEMENTS

ISN would like to thank Ž.Ivezić and A.Robin for useful discussions and insights; R.González-Gérboles for help in carrying out the HB tests; S.Koposov for providing useful insights into the expected stellar distributions; F.Ostrovsky for expert opinion on star-QSO classification and possible impact and A.Kovács for suggestions on using infrared datasets. BN, KW would like to thank Camille Avestruz for useful advice in developing the SVM code. E.Balbinot acknowledges financial support from the European Research Council (StG-335936).

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle

Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MINECO under grants AYA2015-71825, ESP2015-66861, FPA2015-68048, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO), through project number CE110001020, and the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) e-Universe (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

This research uses data from SDSS-III. Funding for SDSS-III has been provided by the Alfred P. Sloan Founda-

tion, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is <http://www.sdss3.org/>. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

This research uses data from the VIMOS VLT Deep Survey, obtained from the VVDS database operated by Césam, Laboratoire d'Astrophysique de Marseille, France.

This research uses data based on observations made with the NASA/ESA Hubble Space Telescope, and obtained from the Hubble Legacy Archive, which is a collaboration between the Space Telescope Science Institute (STScI/NASA), the Space Telescope European Coordinating Facility (ST-ECF/ESAC/ESA) and the Canadian Astronomy Data Centre (CADAC/NRC/CSA).

This research uses data based on zCOSMOS observations carried out using the Very Large Telescope at the ESO Paranal Observatory under Programme ID: LP175.A-0839.

The VISTA Data Flow System pipeline processing and science archive are described in Irwin et al. (2004), Hambly et al. (2008) and Cross et al. (2012).

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.

CosmoHub has been developed by the Port d'Informació Científica (PIC), maintained through a collaboration of the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT). The work was partially funded by the "Plan Estatal de Investigación Científica y Técnica y de Innovación" program of the Spanish government.

REFERENCES

- Abazajian K., et al., 2004, *AJ*, 128, 502
- Aihara H., et al., 2018, *PASJ*, 70, S8
- Albaret 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
- Burke D. L., et al., 2018, *AJ*, 155
- Carnero A., Sánchez E., Croce 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
- Croce M., et al., 2016, *MNRAS*, 455, 4301
- Croce M., et al., 2017, preprint, ([arXiv:1712.06211](https://arxiv.org/abs/1712.06211))
- Cross N. J. G., et al., 2012, *A&A*, 548, A119
- DES Collaboration 2018, preprint, ([arXiv:1801.03181](https://arxiv.org/abs/1801.03181))
- DES Collaboration et al., 2016, *MNRAS*, 460, 1270
- DES Collaboration et al., 2017, preprint, ([arXiv:1708.01530](https://arxiv.org/abs/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, *The Astrophysical Journal Supplement Series*, 235
- 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
- Gaia Collaboration et al., 2016, *A&A*, 595, A1
- Gaia Collaboration Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Bailer-Jones C. A. L., 2018, preprint, ([arXiv:1804.09365](https://arxiv.org/abs/1804.09365))
- 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
- Irwin M. J., et al., 2004, in Quinn P. J., Bridger A., eds, *Proc. SPIE Vol. 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, *Monthly Notices of the Royal Astronomical Society*, 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., Kuposov 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, preprint, ([arXiv:1801.03177](https://arxiv.org/abs/1801.03177))
- Neilsen E., Bernstein G., Gruendl R., Kent S., 2016, Fermilab report, FERMILAB-TM-2610-AE-CD
- Odehahn 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. SPIE Vol. 5498, Z-Spec: a broadband millimeter-wave grating spectrometer: design, construction, and first cryogenic measurements*. pp 11–29 ([arXiv:astro-ph/0411122](https://arxiv.org/abs/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, preprint, ([arXiv:1801.03097](https://arxiv.org/abs/1801.03097))
- Skrutskie M. F., et al., 2006, *AJ*, 131, 1163
- Soumagnac M. T., et al., 2015, *MNRAS*, 450, 666
- 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](https://arxiv.org/abs/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](https://arxiv.org/abs/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 produce 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 information¹², 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 Randomized Trees, Quadratic Discriminant Analysis and the K-Nearest Neighbors 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 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 MODEST, with other ranking features being WAVG_SPREAD_MODEL_R and MAGERR_MODEL_I.

APPENDIX B: EXTERNAL DATASETS

B1 Access to external catalogs used in this work

Reference catalogs used in this work will be made available upon publication. .

¹² https://en.wikipedia.org/wiki/Mutual_information

B2 Queries used to extract the datasets

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

```
SELECT
  s.ra, s.dec, s.dered_r, s.class,
  w.w1mpro as w1, w.j_m_2mass as j,
  s.z, s.zErr
INTO
  mydb.stripe82_wise_2mass_z_match
FROM
  wise_xmatch as xm
JOIN
  specPhotoAll as s on xm.sdss_objid = s.objid
JOIN
  wise_allsky as w on xm.wise_centr = w.centr
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 zWarning = 0
```

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 VHS DR3 database.

```
SELECT      ra,dec,jpetromag,jpetromagerr,jmksext,jmksexterr
FROM
  vhsSource
WHERE
  jerrbits = 0 and ksextbits = 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_
FROM gaia_dr2
WHERE
  (('ra' > 305) or ('ra' < 90)) and ('dec' > -61) and ('dec' < -
```

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 Física Teórica UAM/CSIC, Universidad Autónoma 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

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.