A&A 645, A87 (2021)
https://doi.org/10.1051/0004-6361/202038986
© ESO 2021

Astronomy & Astrophysics

The miniJPAS survey: star-galaxy classification using machine learning[⋆,⋆⋆]

P. O. Baqui[1,⋆⋆⋆], V. Marra[1,2,3,4,⋆⋆⋆], L. Casarini[5], R. Angulo[6,7], L. A. Díaz-García[8], C. Hernández-Monteagudo[9,10,11], P. A. A. Lopes[12], C. López-Sanjuan[9], D. Muniesa[13], V. M. Placco[14], M. Quartin[12,15], C. Queiroz[16], D. Sobral[17], E. Solano[18], E. Tempel[19], J. Varela[9], J. M. Vílchez[20], R. Abramo[16], J. Alcaniz[21], N. Benitez[20], S. Bonoli[13,6,7], S. Carneiro[22], A. J. Cenarro[9], D. Cristóbal-Hornillos[13], A. L. de Amorim[23], C. M. de Oliveira[24], R. Dupke[21,25,26], A. Ederoclite[24], R. M. González Delgado[20], A. Marín-Franch[9], M. Moles[13], H. Vázquez Ramió[9], L. Sodré[24], and K. Taylor[27]

(Affiliations can be found after the references)

Received 21 July 2020 / Accepted 6 November 2020

ABSTRACT

Context. Future astrophysical surveys such as J-PAS will produce very large datasets, the so-called "big data", which will require the deployment of accurate and efficient machine-learning (ML) methods. In this work, we analyze the miniJPAS survey, which observed about ∼1 deg^2 of the AEGIS field with 56 narrow-band filters and 4 *ugri* broad-band filters. The miniJPAS primary catalog contains approximately 64 000 objects in the *r* detection band (mag$_{AB}$ $\lesssim$ 24), with forced-photometry in all other filters.

Aims. We discuss the classification of miniJPAS sources into extended (galaxies) and point-like (e.g., stars) objects, which is a step required for the subsequent scientific analyses. We aim at developing an ML classifier that is complementary to traditional tools that are based on explicit modeling. In particular, our goal is to release a value-added catalog with our best classification.

Methods. In order to train and test our classifiers, we cross-matched the miniJPAS dataset with SDSS and HSC-SSP data, whose classification is trustworthy within the intervals $15 \le r \le 20$ and $18.5 \le r \le 23.5$, respectively. We trained and tested six different ML algorithms on the two cross-matched catalogs: K-nearest neighbors, decision trees, random forest (RF), artificial neural networks, extremely randomized trees (ERT), and an ensemble classifier. This last is a hybrid algorithm that combines artificial neural networks and RF with the J-PAS stellar and galactic loci classifier. As input for the ML algorithms we used the magnitudes from the 60 filters together with their errors, with and without the morphological parameters. We also used the mean point spread function in the *r* detection band for each pointing.

Results. We find that the RF and ERT algorithms perform best in all scenarios. When the full magnitude range of $15 \le r \le 23.5$ is analyzed, we find an area under the curve AUC = 0.957 with RF when photometric information alone is used, and AUC = 0.986 with ERT when photometric and morphological information is used together. When morphological parameters are used, the full width at half maximum is the most important feature. When photometric information is used alone, we observe that broad bands are not necessarily more important than narrow bands, and errors (the width of the distribution) are as important as the measurements (central value of the distribution). In other words, it is apparently important to fully characterize the measurement.

Conclusions. ML algorithms can compete with traditional star and galaxy classifiers; they outperform the latter at fainter magnitudes ($r \gtrsim 21$). We use our best classifiers, with and without morphology, in order to produce a value-added catalog.

Key words. methods: data analysis – catalogs – galaxies: statistics – stars: statistics

1. Introduction

An important step in the analysis of data from wide-field surveys is the classification of sources into stars and galaxies. Although challenging, this separation is crucial for many areas of cosmology and astrophysics. Different classification methods have been proposed in the literature, each having their respective advantages and disadvantages. One of the most frequently used methods is based on morphological separation, where parameters related to the object structure and photometry are used (Bertin & Arnouts 1996; Henrion et al. 2011; Molino et al. 2014; Díaz-García et al. 2019; López-Sanjuan et al. 2019). These methods assume that stars appear as point sources and galaxies as extended sources. This has been shown to be consistent with previous spectroscopic observations (Le Fevre et al. 1995; Dawson et al. 2013; Newman et al. 2013). However, at fainter magnitudes, the differences between these point-like and extended structures decrease and the method becomes unreliable. In what follows, by "stars" we mean point-like objects that are not galaxies, that is, both stars and quasars[1].

Future photometric surveys such as the Javalambre-Physics of the Accelerating Universe Astrophysical Survey (J-PAS, Benitez et al. 2014)[2] and the *Vera Rubin* Observatory Legacy Survey of Space and Time (LSST, Marshall et al. 2017)[3] will detect a large number of objects and will need to manage data

[⋆] Full Table 2 is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/645/A87

[⋆⋆] The catalog is available at http://j-pas.org/datareleases via the ADQL table minijpas.StarGalClass. The ML models are available at github.com/J-PAS-collaboration/StarGalClass-MachineLearning.

[⋆⋆⋆] These authors contributed equally to this work.

[1] Also very compact galaxies such as Green Peas fall into the category of point-like objects (Cardamone et al. 2009; Amorín et al. 2010).

[2] www.j-pas.org

[3] www.lsst.org

produced at an unprecedented rate. The LSST in particular will reach a rate of petabytes of data per year (Garofalo et al. 2016). This wealth of data demands very efficient numerical methods but also gives us the opportunity to deploy machine-learning (ML) algorithms, which, trained on large astronomical datasets, have the potential to outperform traditional methods based on explicit programming, if biases due to potentially unrepresentative training sets are kept under control.

ML has been widely applied in the context of cosmology and astrophysics, see Ishak (2017). A non-exhaustive list of applications is the photometric classification of supernovae (Lochner et al. 2016, Charnock & Moss 2017, Vargas dos Santos et al. 2020), gravitational wave analysis (Biswas et al. 2013; Carrillo et al. 2015), photometric redshift (Bilicki et al. 2018; Cavuoti et al. 2015), galaxy morphology (Gauci et al. 2010; Banerji et al. 2010), and determination of atmospheric parameters for stellar sources (Whitten et al. 2019).

ML applications to star-galaxy separation have been successfully performed on many surveys. Vasconcellos et al. (2011), for example, used various tree methods to classify Sloan Digital Sky Survey (SDSS, Alam et al. 2015) sources. Kim et al. (2015) used classifiers that mixed supervised and unsupervised ML methods with Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS) data. Recently, convolutional neural networks (CNN) have been adopted: using images as input, they achieve an area under the curve (AUC) > 0.99 for CFHTLenS and SDSS data (Kim & Brunner 2017). For more ML applications in the context of star and galaxy classification, see Costa-Duarte et al. (2019), Sevilla-Noarbe et al. (2018), Cabayol et al. (2019), Fadely et al. (2012) and Odewahn et al. (2004).

Our goal here is to classify the objects detected by Pathfinder miniJPAS (Bonoli et al. 2020), which observed ∼1 deg^2 of the AEGIS field with the 56 narrow-band J-PAS filters and the 4 *ugri* broad-band filters for a total of approximately 64 000 objects (mag$_{AB} \lesssim 24$). The ML algorithms that we consider in this work are supervised, and for the learning process, need an external trustworthy classification. We adopted SDSS and Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP, Aihara et al. 2019) data. We compared different ML models to each other and to the two classifiers adopted by the JPAS survey: the `CLASS_STAR` provided by SExtractor (Bertin & Arnouts 1996), and the stellar and galactic loci classifier (SGLC) introduced in López-Sanjuan et al. (2019).

This paper is organized as follows. In Sect. 2 we briefly describe J-PAS and miniJPAS, and we review the classifiers adopted in miniJPAS. In Sect. 3 we present the ML algorithms used in this work, and in Sect. 4 we define the metrics that we use to assess the performance of the classifiers. Our results are presented in Sects. 5 and 5.3, and our conclusions in Sect. 6.

2. J-PAS and miniJPAS

J-PAS is a ground-based imaging survey that will observe 8500 deg^2 of the sky with the technique of quasi-spectroscopy: by observing with 56 narrow-band filters and 4 *ugr*(*i*) broad-band filters[4], it will produce a low-resolution ($R \sim 60$) spectrum (*J*-spectrum hereafter) for every pixel (for the filter specifications, see Bonoli et al. 2020). It features a dedicated 2.5 m telescope with an excellent étendue, equipped with a 1.2 Gigapixel camera with a very large field of view of 4.2 deg^2. The observatory is on the mountain range called Sierra de Javalambre (Spain), at an altitude of approximately 2000 meters. This is an especially dark region with a very good median seeing of 0.7″ (Cenarro et al. 2010). Therefore, J-PAS sits between photometric and spectroscopic surveys, fruitfully combining the advantages of the former (speed and low cost) with those of the latter (spectra). In particular, its excellent photo-*z* performance will enable an accurate study of the large-scale structure of the universe using the galaxy and quasar catalogs produced by J-PAS (Bonoli et al. 2020).

Between May and September 2018, the 2.5 m J-PAS telescope with its filter set was equipped with the Pathfinder camera, which was used to test the telescope performance and execute the first scientific operations. The camera features a 9k × 9k CCD, with a 0.3 deg^2 field of view and 0.225 arcsec pixel size. This led to the miniJPAS survey, which covered a total of ∼1 deg^2 of the AEGIS field[5], reaching the target depth planned for J-PAS (mag$_{AB}$, 5σ in a 3″ aperture, between 21.5 and 22.5 for the narrow-band filters and up to 24 for the broad-band filters). miniJPAS consists of the four fields or pointings AEGIS1-4, each of which has an approximately 0.25 deg^2 field of view. The miniJPAS primary catalog contains 64 293 objects in the *r* detection band, with forced-photometry in all other filters. See Bonoli et al. (2020) for the presentation paper. The miniJPAS Public Data Release was presented to the public in December 2019[6].

2.1. Cross-matched catalogs

The goal of this paper is to develop an ML model that can accurately classify the objects detected by Pathfinder miniJPAS. We consider supervised ML algorithms, therefore we require for the learning process a trustworthy classification by some other survey that has a sufficiently high overlap with miniJPAS. We used SDSS[7] and HSC-SSP[8] data, whose classification is expected to be trustworthy within the intervals $15 \leq r \leq 20$ and $18.5 \leq r \leq 23.5$, respectively. As said earlier, by "stars" we mean point-like objects that are not galaxies, that is, both stars and quasars. We assume that the classification by SDSS and HSC-SSP is trustworthy within this definition (Alam et al. 2015; Aihara et al. 2019).

We found 1810 sources in common with SDSS, 691 galaxies and 1119 stars, and 11 089 sources in common with HSC-SSP, 9398 galaxies and 1691 stars. We refer to Fig. 1 for the *r*-band distributions of stars and galaxies and to Fig. 2 for the redshift distribution of galaxies.

2.1.1. SDSS classification

The SDSS is a photometric and spectroscopic survey conducted at the Apache Point Observatory (New Mexico, USA) with a 2.5 m primary mirror. We used the SDSS DR12 photometric catalog `minijpas.xmatch_sdss_dr12`[9]. Stars are defined according to an extendedness (difference between the CModel and point spread function, PSF, magnitudes) smaller than 0.145[10].

In order to test the photometric calibration by SDSS, we cross-matched the latter with the catalog from the Advance

[4] miniJPAS also features the *i* band, while J-PAS is not expected to have it.

[5] See Davis et al. (2007) for information on the All-wavelength Extended Groth strip International Survey (AEGIS).

[6] `http://j-pas.org/datareleases/minijpas_public_data_release_pdr201912`

[7] `sdss.org/dr12/`

[8] `hsc-release.mtk.nao.ac.jp/doc/`

[9] For details, see `archive.cefca.es/catalogues/minijpas-pdr201912`

[10] `www.sdss.org/dr12/algorithms/classify/#photo_class`

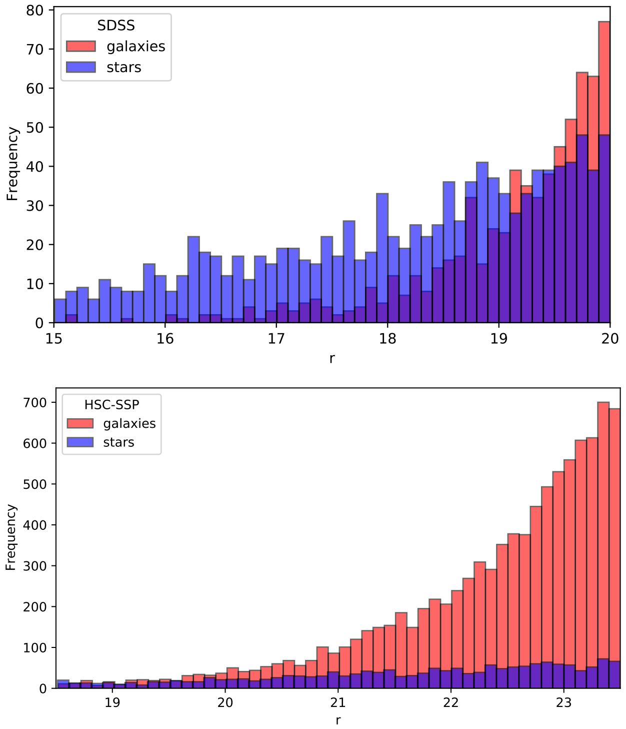

Fig. 1. Histograms of the r-band magnitudes of the sources of the miniJPAS catalog cross-matched with the SDSS (*top*) and HSC-SSP (*bottom*) catalogs. Classification by SDSS and HSC-SSP, respectively. Galaxies are shown in red, stars in semi-transparent blue.

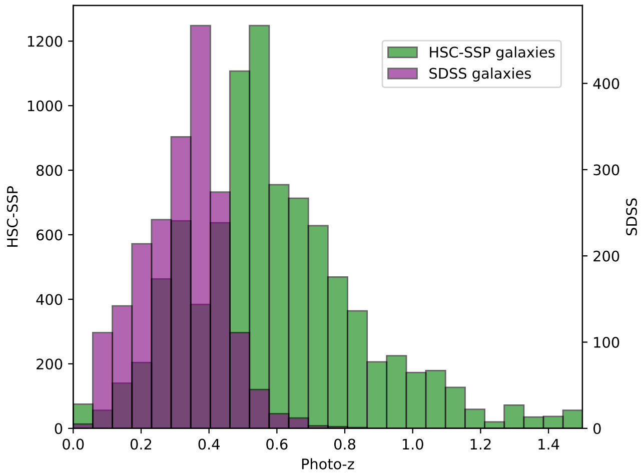

Fig. 2. Histograms of the photometric redshifts of the galaxies of the miniJPAS catalog crossmatched with the SDSS (semi-transparent purple) and HSC-SSP (green) catalogs.

Large Homogeneous Area Medium Band Redshift Astronomical (ALHAMBRA) survey (Moles et al. 2008)[11]. We obtained 1055 sources after imposing mask and saturation flags. As discussed in Molino et al. (2014), ALHAMBRA provides a trustworthy classification in the magnitude range $15 \leq r \leq 21$.

Figure 3 (top) shows that ALHAMBRA covers the relevant magnitude range and agrees well with the SDSS until $r = 20$ (bottom). Within $15 \leq r \leq 20$, the percentages of false negatives and false positives are 0.2% and 1.9%, respectively (positive refers to the object being a galaxy). For the value- added catalog, we will use SDSS in the more limited range $15 \leq r \leq 18.5$ so that the percentages of false negatives and false positives are 0% and 0.7%, respectively (using $p_{\rm cut} = 0.5$, see Sect. 4.1).

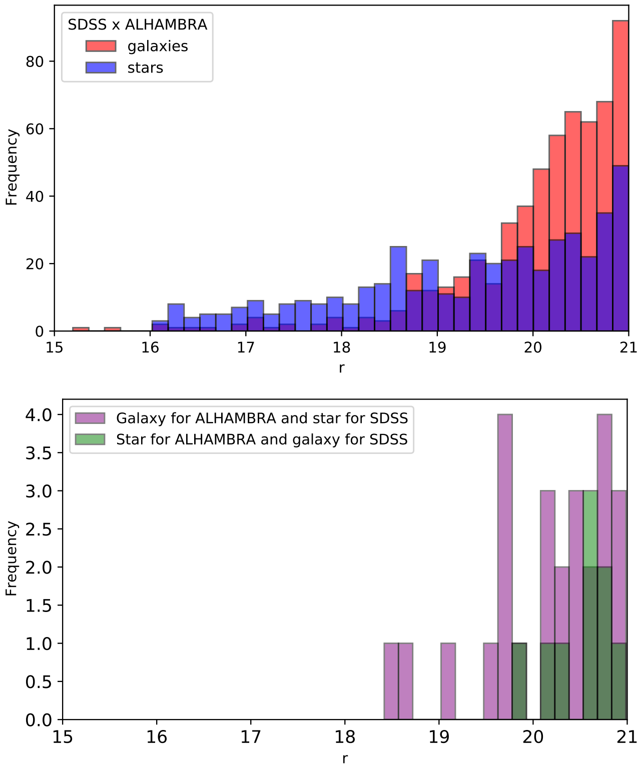

Fig. 3. *Top*: histograms of the r-band magnitudes of the objects resulting from the cross match between the SDSS catalog used in this paper and ALHAMBRA. Galaxies are shown in red, and stars in semi-transparent blue. *Bottom*: disagreement between SDSS and ALHAMBRA as a function of r magnitude. Sources classified as galaxies by ALHAMBRA and as stars by SDSS are in purple, and the reverse is plotted in semi-transparent green.

2.1.2. HSC-SSP classification

The HSC-SSP is a photometric survey with a 8.2 m primary mirror located in Hawaii, USA. We cross-matched the miniJPAS data with the wide field from the Public Data Release 2. Stars are defined according to an extendedness smaller than 0.015[12]. We used the following data-quality constraints: `isprimary = True`, `r_extendedness_flag!=1` and `r_inputcount_value>=4` for HSC-SSP, and `flag=0` and `mask=0` for miniJPAS. The cross match was performed with the `TOPCAT`[13] software with a tolerance of 1 arcsec.

In order to test the photometric calibration by HSC-SSP, we cross-matched the latter with the spectroscopic catalogs from the DEEP2 Galaxy Redshift Survey (1992 sources Matthews et al. 2013). We did not use this spectroscopic catalog to check the photometric SDSS calibration because it does not cover the required magnitude range.

Figure 4 (top) shows that DEEP2 covers the relevant magnitude range and agrees well with HSC-SSP (bottom). For the range $18.5 \leq r \leq 23.5$, the percentages of false negatives and false positives are 1.9% and 0%, respectively.

[11] svo2.cab.inta-csic.es/vocats/alhambra

[12] hsc-release.mtk.nao.ac.jp/doc/index.php/stargalaxy-separation-2/

[13] www.star.bris.ac.uk/ mbt/topcat/

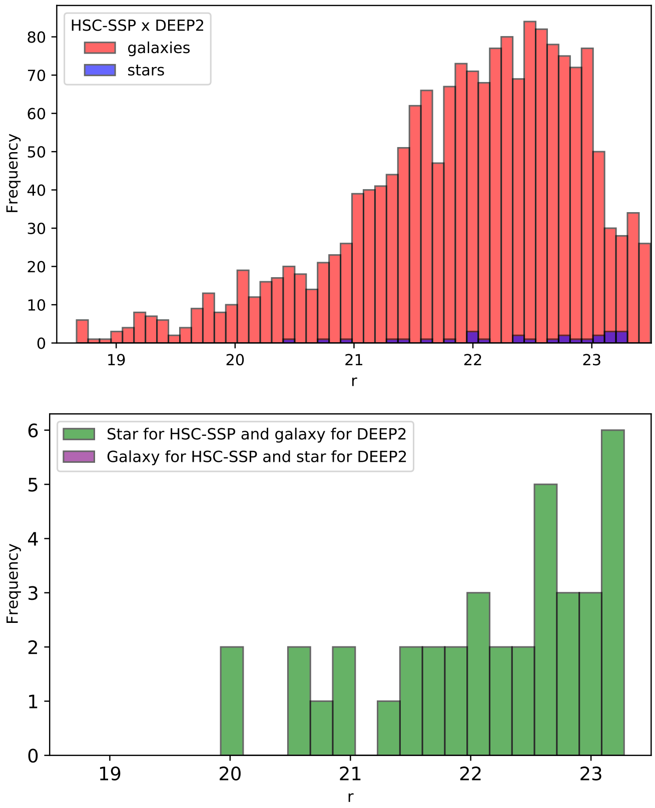

Fig. 4. *Top*: histograms of the r-band magnitudes of the objects resulting from the cross match between the HSC-SSP catalog used in this paper and DEEP2. *Bottom*: disagreement between HSC-SSP and DEEP2 as a function of r magnitude. No object was classified as a galaxy by HSC-SSP and as a star by DEEP2.

2.2. Input parameters for the ML algorithms

The features that are used as input for our algorithms can be grouped into photometric and morphological classes. In addition to these two sets of features, we also considered the average PSF in the r detection band of the four fields of miniJPAS, which is 0.70″ for AEGIS1, 0.81″ for AEGIS2, 0.68″ for AEGIS3, and 0.82″ for AEGIS4. The different PSF values signal different observing conditions: by including the PSF value, we let the ML algorithms know that the data are not homogeneous.

2.2.1. Photometric information

As photometric information we considered the MAG_AUTO magnitudes associated with the 60 filters together with their errors. The rationale behind including the errors is that in this way, we can characterize the statistical distribution associated with a magnitude measurement. Observations may be inhomogeneous because observing conditions vary, and the measurement errors should be able to account at least in part for this potential bias. As we show below, how well the magnitude associated with a filter can be measured may be more important than the actual measurement.

We reiterate that sources are detected in the r band, therefore a non-detection in the other filters may occur. Null or negative fluxes (after background subtraction) are assigned a magnitude value of 99. The ML algorithms are expected to learn that 99 marks missing values.

2.2.2. Morphological information

We considered the following four morphological parameters:

- Concentration $c_r = r_{1.5''} - r_{3.0''}$, where $r_{1.5''}$ and $r_{3.0''}$ are the r-band magnitudes within fixed circular apertures of 1.5″ and 3.0″, respectively,
- Ellipticity A/B, where A and B are the RMS of the light distribution along the maximum and minimum dispersion directions, respectively.
- The full width at half maximum, $FWHM$, assuming a Gaussian core,
- MU_MAX/MAG_APER_3_0 (r band), where MU_MAX and MAG_APER_3_0 are the peak surface brightness above background and the magnitude within 3.0″, respectively. We took the ratio in order to have a parameter that is complementary to c_r.

Figures 5 and 6 show their distributions for stars and galaxies and the two catalogs. The stellar bimodality in c_r and MU_MAX/MAG_APER_3_0 arises because the four fields feature a different average PSF. We discuss these figures when we examine the feature importance in Sect. 5.4.

2.3. J-PAS star and galaxy classifiers

We briefly discuss the star and galaxy classifiers available for miniJPAS. However, first we show how HSC-SSP classifies objects into stars and galaxies. This is performed by drawing a hard cut in the source parameter space. In Fig. 7 we plot the difference between mag_{PSF} and $\text{mag}_{\text{cmodel}}$ as a function of $\text{mag}_{\text{cmodel}}$ for the HSC-SSP data using their r band (for the definitions, see Aihara et al. 2019). Stars are expected to have $\text{mag}_{\text{PSF}} \simeq \text{mag}_{\text{cmodel}}$, while because of their extended structure, galaxies should feature $\text{mag}_{\text{PSF}} > \text{mag}_{\text{cmodel}}$. Stars can therefore be separated from galaxies through a cut in the extendedness parameter $\text{mag}_{\text{PSF}} - \text{mag}_{\text{cmodel}}$, which we show with a yellow line in Fig. 7. The disadvantage of this model is that it provides an absolute classification for a scenario in which the uncertainties increase at weaker magnitudes. For $r_{\text{cmodel}} \gtrsim 24$, the separation is not reliable because stars no longer cluster around a null extendedness.

2.3.1. CLASS_STAR

Source Extractor (SExtractor, Bertin & Arnouts 1996) is a software developed for processing large images (60k × 60k pixels). It has been widely applied to photometric surveys including miniJPAS. In addition to detecting sources, SExtractor also classifies objects into stars and galaxies. The software has two internal classifiers, CLASS_STAR and SPREAD_MODEL. miniJPAS includes the classification through CLASS_STAR, which is based on neural networks (see Sect. 3.5)[14]. The network has ten inputs: eight isophotal areas, the peak intensity, and the seeing control parameter. The output is probabilistic, and quasars are classified as stars (in agreement with our convention). CLASS_STAR is reliable up to $r \sim 21$ (see also Bertin & Arnouts 1996).

2.3.2. Stellar and galactic loci classifier

miniJPAS includes the Bayesian classifier (SGLC) developed by López-Sanjuan et al. (2019) for J-PLUS data[15]. The concentration versus magnitude diagram presents a bimodal distribution,

[14] sextractor.readthedocs.io/en/latest/ClassStar.html
[15] j-plus.es/datareleases

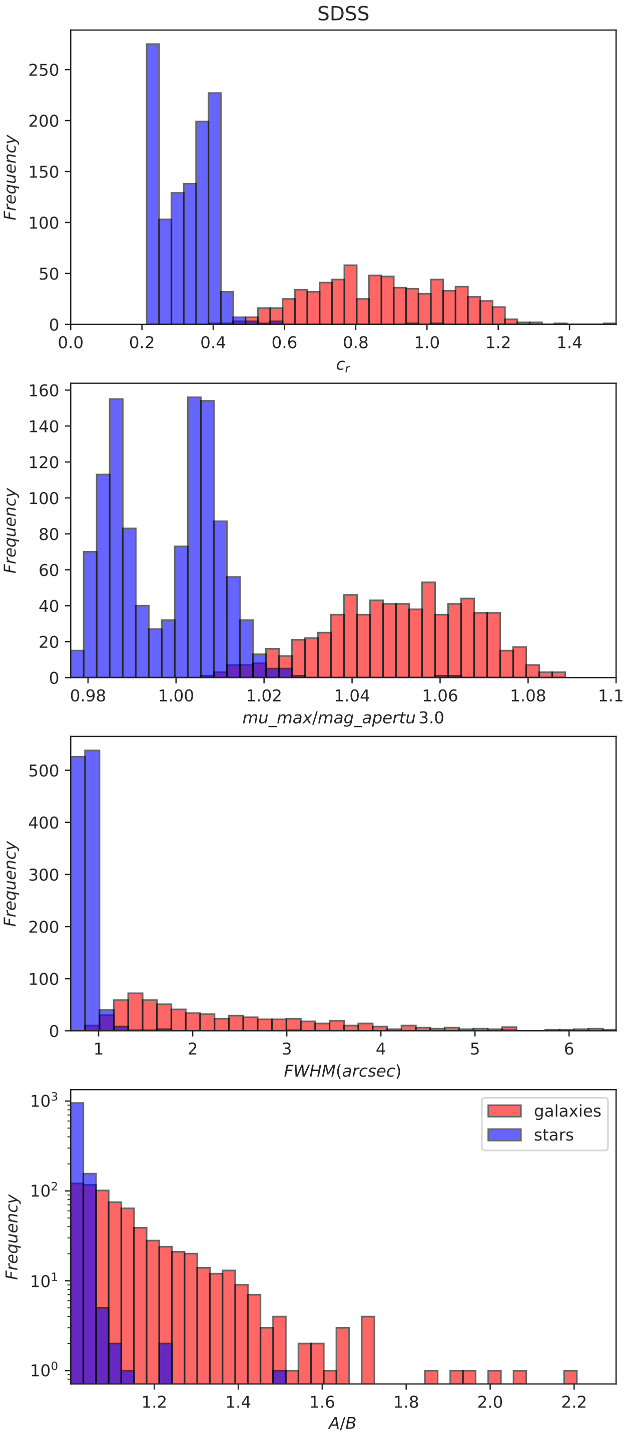

Fig. 5. Distributions of the morphological parameters of stars and galaxies for the miniJPAS catalog cross-matched with the SDSS. Galaxies are shown in red, and stars in semi-transparent blue.

corresponding to compact point-like objects and extended sources. López-Sanjuan et al. (2019) modeled both distributions to obtain the probability of each source to be compact or extended. The model with suitable priors was then used to estimate the Bayesian probability that a source is a star or a galaxy. In this case as well, quasars are expected to be classified as stars. This method was updated to miniJPAS data, in particular, we adopted a different galaxy population model. See Bonoli et al. (2020) for more details.

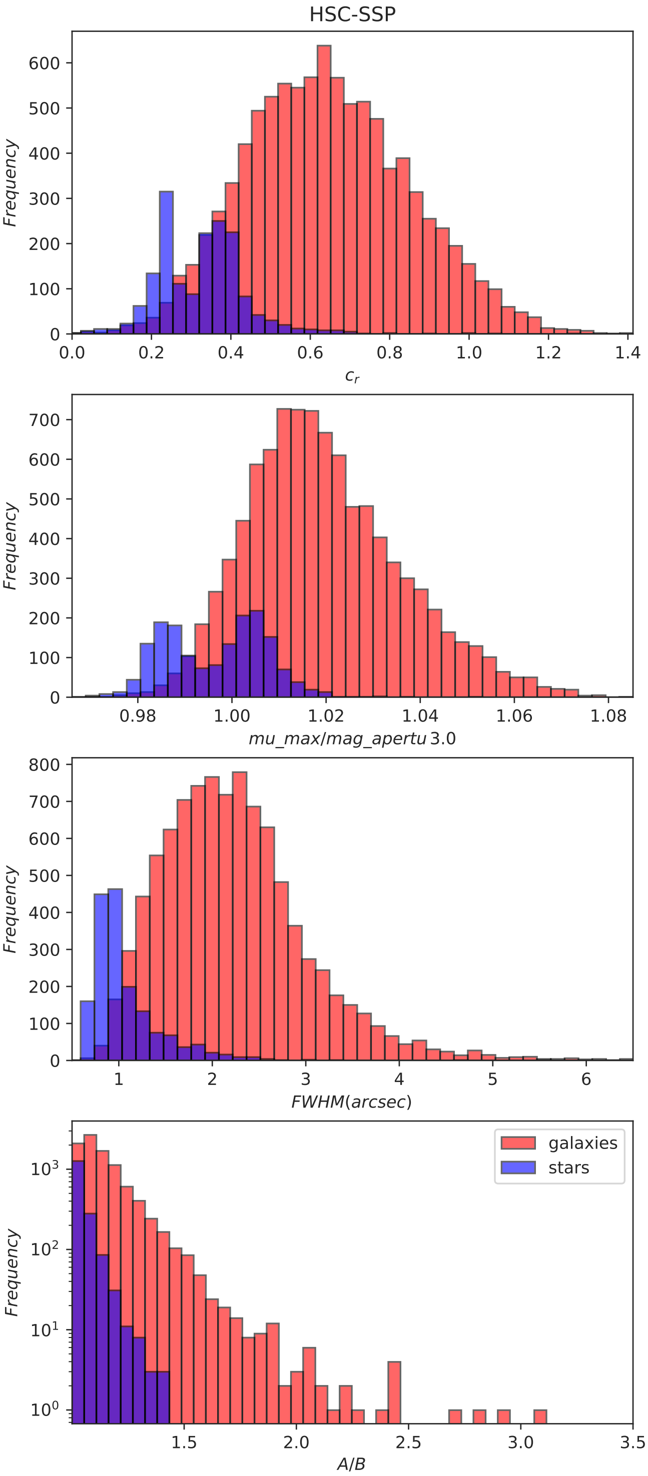

Fig. 6. Distributions of the morphological parameters of stars and galaxies for the miniJPAS catalog cross-matched with HSC-SSP. Galaxies are shown in red, and stars in semi-transparent blue.

3. Machine learning

Machine learning is a branch of artificial intelligence that includes statistical and computational methods dedicated to providing predictions or taking decisions without being explicitly programmed to perform the task. Machine learning is employed in a variety of computing tasks, for which the explicit programming of well-performing algorithms is difficult or unfeasible.

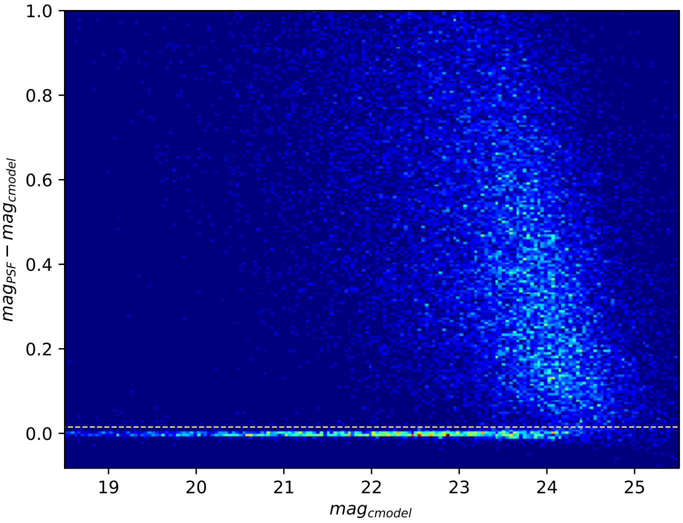

Fig. 7. Density of objects as a function of extendedness (the difference between mag_{PSF} and $\text{mag}_{\text{cmodel}}$) and $\text{mag}_{\text{cmodel}}$ for HSC-SSP data. The yellow line marks an extendedness of 0.015. According to this morphological classification, the sources below the cut are stars and those above the cut are galaxies.

ML methods can either be supervised or unsupervised. The former learn from preclassified data that have known inputs and outputs. When classification is not available, unsupervised methods are used instead, which can group items that are related in the parameter space, that is, learn without the need of external information.

We focus on binary supervised classification methods. In this case, the model (the internal parameters of the algorithm) is implicitly adjusted through the training set. Its performance is then tested with the remaining part of the dataset, the test set. Specifically, the internal parameters of the prediction function $f : \mathbb{R}^n \to Y$ are trained through the training dataset $\mathbf{x}_i \in \mathbb{R}^n$ (n is the dimensionality of the feature space, and i labels the elements of the training set) with classifications $y_i \in \{0, 1\}$, where 1 stands for galaxy and 0 for star. Classifiers are divided into nonprobabilistic and probabilistic classifiers. The former type of classifier outputs the best class, and the latter the probability of the classes (the best class is taken as the class with the highest probability). We considered only binary probabilistic classifiers so that it is $f : \mathbb{R}^n \to [0, 1]$, that is, f gives the probability that an object is a galaxy. The probability of being a star is simply $1 - f$. A value of f close to 1 means that the object is likely a galaxy.

According to the no-free lunch theorem, there is not an ideal algorithm that performs better than the others in any situation (Wolpert 1996). As it is impossible to test all the methods available with all the possible choices of hyperparameters, we followed the strategy of first exploring some of the main ML methods: K-nearest neighbors (KNN), decision trees (DT), random forest (RF), and artificial neural networks (ANN). We also tested support-vector machine (SVM) with the linear, polynomial, and radial basis function (RBF) kernels. We found results similar to DT and KNN. Subsequently, because the RF technique responded best, we decided to focus on DT algorithms and ensemble models, therefore we added extremely randomized trees (ERT) and an ensemble classifier (EC) to our analysis. These algorithms can be used for both regression and classification. While classification is used to predict whether an object belongs to a class, regression is used to predict real-value outputs that do not belong to a fixed set. For example, regression is used when photometric information is used in order to predict the source redshift. We only considered classification. We implemented these algorithms using the `scikit-learn`[16] package, which is written in python (Pedregosa 2011). For more information about supervised learning, see Mitchell (1997) and Hastie et al. (2009). For the training and test sets, we used 80% and 20% of the cross-matched catalogs, respectively. The division was performed randomly. This guarantees a good training and an accurate testing. A 70%–30% split is also a viable alternative. As mentioned in Sect. 2.1, the training sets are unbalanced as they feature a different number of galaxies and stars. We show the purity curves for stars and galaxies in order to estimate the performance for each class. We now briefly review the six ML algorithms adopted in this paper. The hyperparameters used in the algorithms can be found in Appendix F.

3.1. K-nearest neighbors

The KNN algorithm is one of the most simple ML methods (Altman 1992; Hastie et al. 2009). It calculates the distance between the element to be classified (within the test set) and the elements belonging to the training set. The predicted class is calculated using the k-nearest neighbors. Although we used the Euclidean metric, it is possible to choose others metrics to compute the distances. This method is very fast, and its computational cost is proportional to the size of the training set.

The output of the model is discrete when the majority vote from the k-nearest neighbors is used (a vote is a classification by a neighbor). We used the probabilistic version, which assigns a probability to each class. In this case, the classification is given by the average of the nearest k neighbors,

$$f(\boldsymbol{x}_q) = \frac{\sum_{i=1}^{k} w_i f(\boldsymbol{x}_i)}{\sum_{i=1}^{k} w_i} \qquad \text{with} \qquad w_i = \frac{1}{d(\boldsymbol{x}_q, \boldsymbol{x}_i)^2}\,, \tag{1}$$

where the sum over the k-nearest neighbors is weighted by the weights w_i, which are the inverse of the square of the distance $d(\boldsymbol{x}_q, \boldsymbol{x}_i)$ from the neighbors ($\boldsymbol{x}_i$) to the element to be classified ($\boldsymbol{x}_q$, q labels the test set), and $f(\boldsymbol{x}_i) = y_i$ are the classifications of the training set. As discussed in Sect. 4.3, the number k of neighbors is optimized through k-fold cross-validation. KNN has the advantage of being simple, intuitive, and competitive in many areas. However, its computational complexity increases with the number of data points.

3.2. Decision trees

The DT methods (see Breiman et al. 1984; Hastie et al. 2009) recurrently divide the parameter space according to a tree structure, following the choice of minimum class impurity of the groups at every split. To build a decision tree, we first defined an information gain (IG) function,

$$IG(D_{\text{p}}, x_{\text{t}}) = I(D_p) - \frac{N_{\text{left}}}{N_{\text{p}}} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N_{\text{p}}} I(D_{\text{right}})\,, \tag{2}$$

where D_{p} is the parent dataset of size N_{p}, D_{left} and D_{rigth} are the child datasets of sizes N_{left} and N_{right}, respectively, and I is a function called impurity. At every step the dataset is divided according to the feature and threshold x_{t} that maximize the IG function, or equivalently, that minimize the impurity in the child datasets. Within our notation, x_{t} is the threshold for the feature

[16] `scikit-learn.org`

that maximizes IG (there are n features). We considered several impurity functions, such as entropy, classification error, and Gini. For example, the latter is

$$I_G(m) = 1 - \sum_{i=0,1} p(i|m)^2 \,, \tag{3}$$

where $p(i|m)$ is the fraction of data belonging to class i (0 or 1) for a particular node m that splits the parent dataset into the child datasets. After the growth of the tree is completed, the feature space is divided with probabilities associated with each class, and the probability for a test element is that of the region it belongs to.

During the branching process described above, some features appear more often than others. Using this frequency, we can measure how important each feature is in the prediction process. We define the importance of each feature as

$$\mathrm{Imp}(x) = \sum_t \frac{N_\mathrm{p}}{N_\mathrm{tot}} IG(D_p, x_\mathrm{t}) \,, \tag{4}$$

where N_tot is the size of the dataset. The higher the number of times a feature branches a tree, the higher its importance. The first features that divide the tree tend to be of greater importance because the factor $N_\mathrm{p}/N_\mathrm{tot}$ in Eq. (4) decreases as the tree grows (N_p decreases). DT is characterized by an easy interpretability and handling, but it is sensitive to small changes in the training set and may therefore be biased.

3.3. Random forest

Random forest (Breiman 2001; Hastie et al. 2009) is an ensemble algorithm built from a set of decision trees (the forest). Each tree generates a particular classification, and the RF prediction is the combination of the different outputs. Each tree is different because of the stochastic method that is used to find the features when the IG function is mximized. Moreover, the bootstrap statistical method is used to build different datasets from the original one in order to grow more trees. For the discrete case, the output is built from the majority vote, as we showed for the KNN algorithm. For the probabilistic case, we calculated the RF output as the average of the probabilities of each class for each tree. Finally, we computed the feature importances Imp(x) for each tree of the ensemble and then averaged them to obtain the RF feature importance. The diversity of trees decreases the bias as compared to DT, generating globally better models. On the other hand, RF requires greater memory and time than DT.

3.4. Extremely randomized trees

Extremely randomized trees (Geurts et al. 2006) is an ensemble method similar to RF. There are only two differences between RF and ERT. The first is that ERT originally does not use bootstrap, although the implementation in `scikit-learn` allows inserting it into the analysis. The second is that while RF tries to find the best threshold for a features through the IG function, in ERT the division is made randomly. Of all the randomly generated splits, the split that yields the highest score is then chosen to split the node. For large datasets, ERT algorithms are faster than RF algorithms and yield a similar performance (Geurts et al. 2006).

3.5. Artificial neural networks

Artificial neural networks mimic the functioning of the nervous system; they are able to recognize patterns from a representative dataset (for an introduction, see Mitchell 1997; Hastie et al. 2009). The model we used in our analysis consists of a simple supervised model called multilayer perceptron (MLP).

The MLP consists of a set of perceptrons arranged in different layers. A perceptron, or artificial neuron, is a binary classifier. The data features are inserted in the input layer, the learning process occurs in the hidden layers, and the object classification is performed by the output layer. The information in the hidden layers is passed several times through each perceptron until convergence. In this algorithm, we can have several layers containing hundreds of perceptrons. To train the neural network, a cost function is used that should be minimized. As learning method we used backpropagation (Rumelhart et al. 1986). We used `LBFGS`, `Stochastic Gradient Descent`, and `Adam` cost functions in addition to various activation functions. The values of the hyperparameters that give the best performance are given in Appendix F. In particular, we adopted one hidden layer with 200 neurons. ANN algorithms are very competitive and have the ability to deal with complex nonlinear problems, but possess low interpretability and require powerful processors.

Finally, we briefly discuss the differences between `CLASS_STAR` and the ANN classifier used in this work. First, our ANN classifier was trained on real miniJPAS data, while `CLASS_STAR` was trained on simulated images. Second, although they both feature one hidden layer, `CLASS_STAR` has an input layer with 10 parameters and a hidden layer with 10 neurons, while our classifier uses an input layer with 64 parameters (four morphological features plus 60 photometric bands) and has a hidden layer with 200 neurons.

3.6. Ensemble classifiers

The ensemble method aims to construct a metaclassifier from the union of different algorithms. When they are efficiently combined, these classifiers can generally perform better than the single best algorithm. In order to combine the classifiers, we adopted the weighted-sum rule with equal weights. The probability prediction function f can be written as

$$f(\boldsymbol{x}_q) = \frac{\sum_{j=1}^{m} w_j f_j(\boldsymbol{x}_q)}{\sum_{j=1}^{m} w_j} \,, \tag{5}$$

where $f_j(\boldsymbol{x}_q)$ is the probabilistic binary classification from the classifier j and m is the number of classifiers considered. We implemented this algorithm using the `VotingClassifier` function from `scikit-learn`. In the following, the ensemble classifier (EC) comprises ANN, RF, and SGLC methods with equal weight ($w_j = 1/3$). The EC is not a pure ML classifier as it uses SGLC, see Sect. 2.3.2. These algorithms generally inherit the advantages and disadvantages of the methods they are based on.

4. Performance metrics

We now introduce the metrics that we adopted in order to assess the performance of the classifiers. See Mitchell (1997) and Hastie et al. (2009) for more details.

4.1. Confusion matrix

We considered probabilistic classifiers, therefore the classification of sources into stars or galaxies depends on a probability threshold p_cut to be specified. In our case, all objects with $f > p_\mathrm{cut}$ are classified as galaxies. The choice of p_cut depends on completeness and purity requirements.

When p_cut is specified, the classification performance can be summarized using the confusion matrix that thoroughly

Table 1. Performance of the classifiers considered in this paper for the miniJPAS catalog cross-matched with the SDSS catalog ($15 \leq r \leq 20$, top) and with the HSC-SSP catalog ($18.5 \leq r \leq 23.5$, bottom).

miniJPAS-SDSS	AUC_{M+P}	AUC_{P}	$\mathrm{AP}^{\mathrm{gal}}_{M+P}$	$\mathrm{AP}^{\mathrm{gal}}_{P}$	MSE_{M+P}	MSE_{P}
SGLC	0.994	–	0.989	–	0.006	–
`CLASS_STAR`	0.997	–	0.993	–	0.032	–
KNN	0.996 ± 0.003	0.991 ± 0.007	0.990 ± 0.008	0.984 ± 0.009	0.015	0.027
DT	0.992 ± 0.006	0.984 ± 0.012	0.983 ± 0.011	0.974 ± 0.018	0.011	0.032
RF	**0.997 ± 0.006**	0.996 ± 0.004	0.992 ± 0.009	0.995 ± 0.010	0.006	0.019
EC	0.997	0.997	0.995	0.996	0.006	0.014
ANN	0.997 ± 0.004	0.988 ± 0.009	**0.994 ± 0.017**	0.983 ± 0.015	0.012	0.043
ERT	0.997 ± 0.002	**0.997 ± 0.003**	0.993 ± 0.006	**0.996 ± 0.004**	**0.005**	**0.019**
miniJPAS-HSC-SSP	AUC_{M+P}	AUC_{P}	$\mathrm{AP}^{\mathrm{gal}}_{M+P}$	$\mathrm{AP}^{\mathrm{gal}}_{P}$	MSE_{M+P}	MSE_{P}
SGLC	0.970	–	0.992	–	0.040	–
`CLASS_STAR`	0.956	–	0.991	–	0.053	–
KNN	0.950 ± 0.010	0.824 ± 0.023	0.989 ± 0.003	0.959 ± 0.006	0.053	0.098
DT	0.961 ± 0.009	0.855 ± 0.017	0.990 ± 0.003	0.959 ± 0.007	0.061	0.132
RF	0.978 ± 0.005	**0.938 ± 0.007**	0.995 ± 0.002	**0.986 ± 0.002**	0.032	0.054
EC	0.979	0.967	0.996	0.993	0.031	0.040
ANN	0.970 ± 0.007	0.885 ± 0.014	0.993 ± 0.003	0.969 ± 0.005	0.036	0.070
ERT	**0.979 ± 0.006**	0.931 ± 0.006	**0.995 ± 0.002**	0.982 ± 0.002	**0.032**	**0.053**

Notes. The best performance is marked in bold (EC is not considered). P stands for the analysis that uses only photometric bands, and $M + P$ stands for the analysis that uses photometric bands together with morphological parameters.

compares predicted and true values. For a binary classifier the confusion matrix has four entries: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP are sources that are correctly classified as galaxies by the model. TN are sources that are correctly classified as stars. FN are sources that are classified as stars by the model when they are galaxies. FP are sources that are classified as galaxies when they are stars.

4.2. Metrics

The receiver-operating characteristic (ROC) curve represents a comprehensive way to summarize the performance of a classifier. It is a parametric plot of the true positive rate (TPR) and false positive rate (FPR) as a function of p_{cut},

$$\mathrm{TPR}(p_{\mathrm{cut}}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR}(p_{\mathrm{cut}}) = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{6}$$

with $0 \leq p_{\mathrm{cut}} \leq 1$. TPR is also called "recall", and in astronomy, it is the completeness. The performance of a classifier can then be summarized with the area under the curve (AUC). The AUC can assume values between 0 and 1. A perfect classifier has a value of 1, while a random classifier on average has a value of 1/2.

The purity curve is a useful method to assess the performance of an unbalanced classifier (as the training set does not feature the same number of stars and galaxies). It is a parametric plot of the completeness (or recall) and the purity (or precision) as a function of p_{cut},

$$\mathrm{Purity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{7}$$

In order to summarize the purity curve, we considered the average precision (AP), which is the area under the purity curve and takes values between 0 and 1.

Finally, the algorithm performance can be measured with the mean squared error (MSE) defined as

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{test}}} \sum_{q=1}^{N_{\mathrm{test}}} \left(y_q - f(\boldsymbol{x}_q)\right)^2 , \tag{8}$$

where y_q are the test-set classifications, and N_{test} is the test-set size. MSE = 0 characterizes a perfect performance. In the present case of a binary classifier, it is $\mathrm{MSE} = (\mathrm{FP} + \mathrm{FN})/N_{\mathrm{test}}$.

4.3. k-fold cross validation

We used the k-fold cross-validation method in order to optimize the algorithm hyperparameters, test for overfitting and underfitting, and estimate the errors on AUC and AP. k-fold cross validation separates the training data into k equal and mutually exclusive parts (we adopted $k = 10$). The model is trained in $k - 1$ parts and validated in the remaining one, called validation. This process is repeated cyclically k times. The final result is the mean and standard deviation of the metric.

The ML methods described in Sect. 3 depend on several internal hyperparameters (e.g., the number k of neighbors in KNN). In order to optimize them, we performed k-fold cross validation for several hyperparameter configurations. The results of the next section are relative to the best configuration according to the AUC. See Appendix F for details.

We also tested the ML algorithms against overfitting and underfitting. The former happens when the training is successful (high AUC) but not the testing (low AUC). The latter occurs when neither training nor testing are successful (both AUCs are low). We confirmed that the average AUC from the k-fold cross validation agrees with the AUC from the test set; all the methods passed this test.

Finally, we can use k-fold cross validation in order to estimate the error in the determination of the AUC and AP. This will help us understand if the differences between two estimators are significative and also how sensitive a classifier is with respect to the division of the dataset into training and test sets.

5. Results

We now present our results for the algorithms introduced in Sect. 3 applied to the cross-matched catalogs described in Sect. 2.1. For stars and galaxy number counts, we refer to the miniJPAS presentation paper (Bonoli et al. 2020).

5.1. miniJPAS-SDSS catalog

The performance of the star and galaxy classifiers considered in this paper for the miniJPAS catalog cross-matched with the SDSS catalog in the magnitude interval $15 \leq r \leq 20$ is excellent. The results are summarized in Table 1, where the best result are marked in bold (EC is not considered as it is not a pure ML classifier). We omit the corresponding figures as they are not informative given the excellent performance. The errors on the pure-ML classifiers are estimated through k-fold cross validation. In order to assess the importance of photometric bands and morphological parameters, the analysis considers two cases: only photometric bands (P subscript in the table), and photometric bands together with morphological parameters ($M + P$ subscript in the table). This distinction does not apply to SGLC and CLASS_STAR as they always include the use of morphological parameters.

In the analysis with photometric bands only, the best ML methods are RF and ERT, showing the power of combining several trees when a prediction is made. Remarkably, using only photometric information, RF and ERT outperform SGLC and CLASS_STAR. When we added morphological information, the almost perfect performance of RF and ERT did not improve, showing again that in this magnitude range, photometric information is sufficient. In Table 1 we also show the MSE, whose results agree with those from the ROC and purity curves.

Another way to qualitatively analyze the performance of a classifier is through a color-color diagram for objects classified as stars ($p \leq p_{\rm cut} = 0.5$). Figure 8 shows the stellar locus in the $g - r$ versus $r - i$ color space. The blue line is a fifth-degree polynomial fit, based on miniJPAS data that were classified as stars by SDSS. The various markers represent the averages of each classifier for different bins. We observe a small dispersion around the curve, which decreases when morphological parameters are included. This indicates that the classifiers and the classification from SDSS agree well.

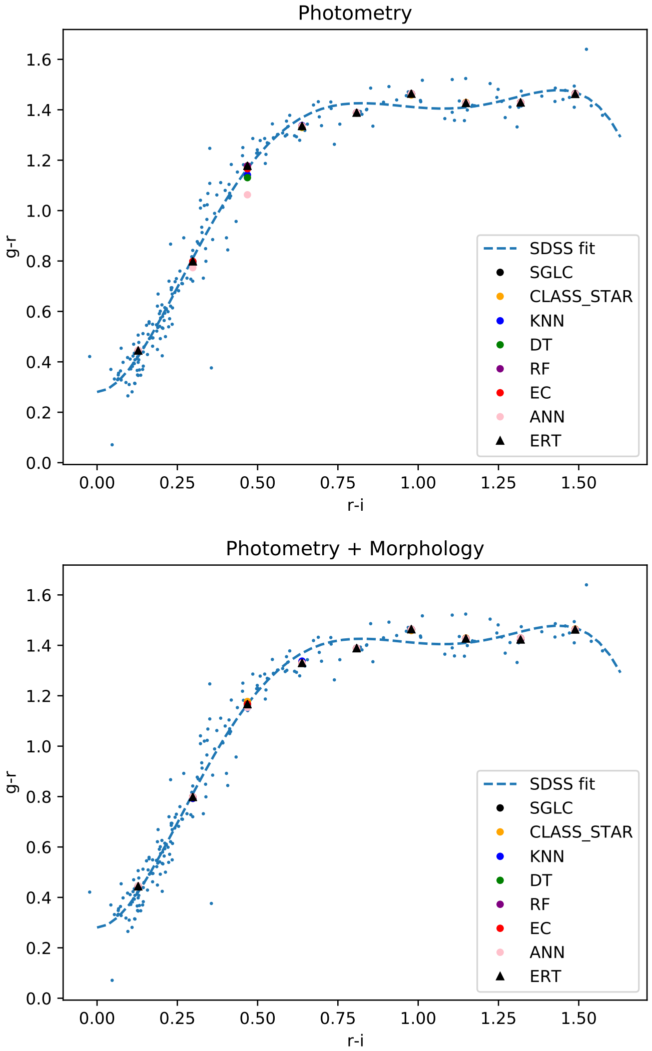

Fig. 8. Stellar locus characterization. The small blue dots represent the stellar locus for the objects classified as stars ($p \leq p_{\rm cut} = 0.5$) of the miniJPAS catalog cross-matched with the SDSS catalog in the magnitude interval $15 \leq r \leq 20$. The dashed line represents a polynomial fit to the stellar locus. *Top panel*: analysis that uses only photometric bands, and *bottom panel*: analysis that also uses morphological information. The larger colored symbols represent the mean stellar locus provided by the different ML models. For comparison we also show the classification by CLASS_STAR and SGLC, which always use morphological parameters.

5.2. miniJPAS-HSC-SSP catalog

As shown in the previous section, the star and galaxy classification in the range $15 \leq r \leq 20$ is not problematic. However, the scenario changes at fainter magnitudes. As the amount of light decreases and less information reaches the telescope, the performance of the algorithms decreases to the point that it is important to look for alternative solutions such as ML. We present the analysis of the previous section applied to the miniJPAS catalog cross-matched with the HSC-SSP catalog in the magnitude interval $18.5 \leq r \leq 23.5$.

Figure 9 and Table 1 show the results. Using only photometric information, the RF algorithm achieves the remarkable score of AUC = 0.938. Although it performance not as well as SGLC and CLASS_STAR (which use morphology), this result shows that ML has the potential of identifying compact galaxies, which have the same morphology as stars. It has also been argued that models that use just photometry can classify quasars as extragalactic objects better than models that use morphological parameters (Costa-Duarte et al. 2019). The use of the morphological parameters improves the performance of the ML methods to the point that ERT and RF perform better than CLASS_STAR and SGLC. In Appendix C we repeat the analysis of Fig. 9 for the mJP-AEGIS1 field, which is the miniJPAS pointing with the best PSF.

It is interesting to note that although the classifiers feature lower AUCs and higher MSEs than the analyses of the previous section, the APs reach similar values even when we use only photometric bands. The reason is that this dataset has many more galaxies and only 15.3% stars. This means that even if there are contaminations by stars, the effect is weaker.

Finally, we show the stellar locus in Fig. 10. The dispersion is greater than in Fig. 8, especially when only photometric bands

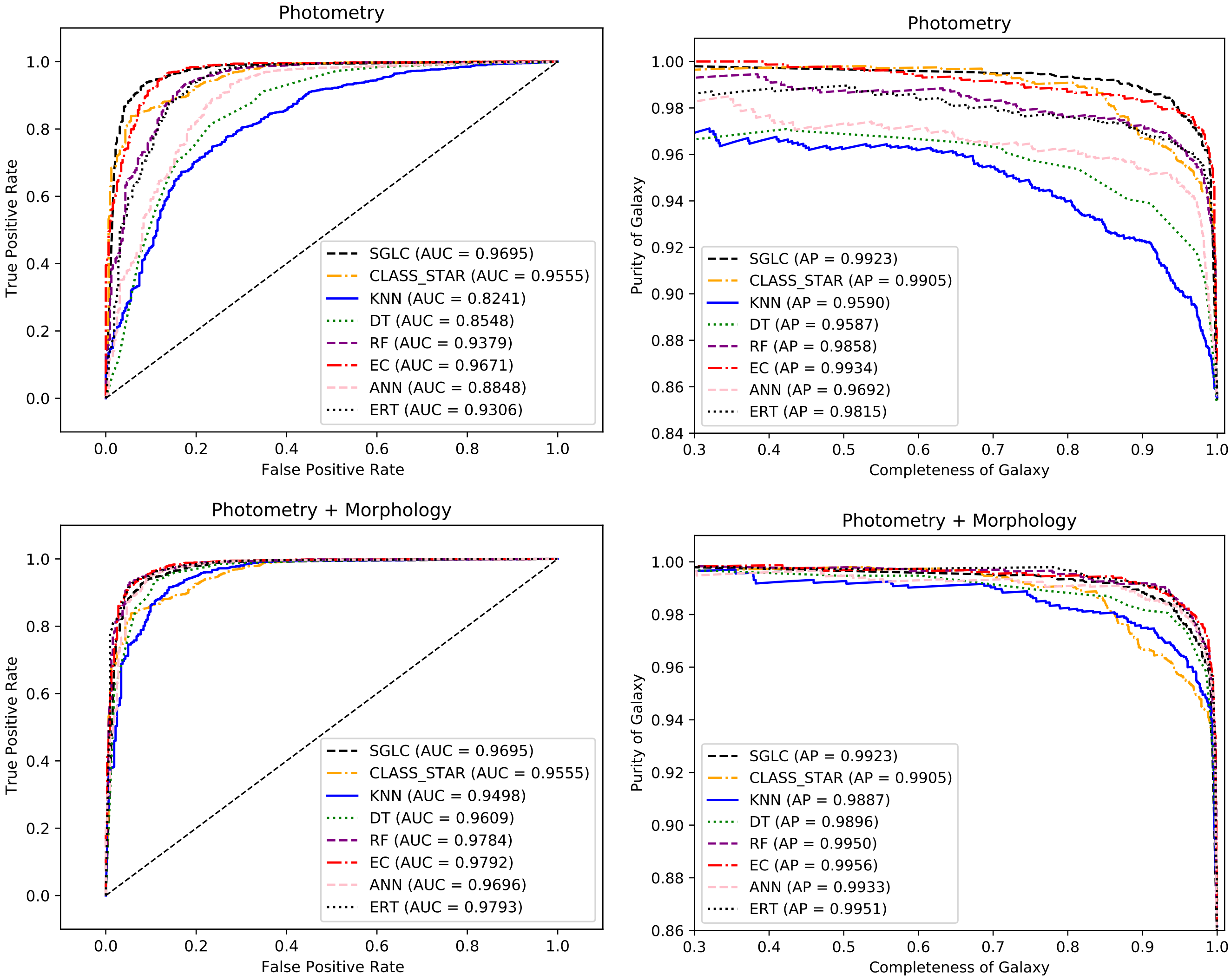

Fig. 9. ROC curves (*left panels*) and purity curves for galaxies (*right panels*) for the classifiers considered in this paper for the miniJPAS catalog cross-matched with the HSC-SSP catalog in the magnitude interval $18.5 \leq r \leq 23.5$. *Top panels*: analysis that uses only photometric bands, and *bottom panels*: analysis that uses photometric bands and morphological parameters. For comparison we also show the classification by `CLASS_STAR` and SGLC, which always use morphological parameters. The axes ranges are varied in order to better show the curves. The results are summarized in Table 1 (bottom).

are used in the analysis. Nevertheless, the ML methods return the correct shape of the stellar locus, and their performance is similar to the one by SGLC.

5.3. Value-added catalog

The ultimate goal of this work is to release a value-added catalog with our best alternative classification. In the previous section we studied star and galaxy classification in the (partially overlapping) magnitude ranges $15 \leq r \leq 20$ and $18.5 \leq r \leq 23.5$. In order to have a uniform dependence on $p_{\rm cut}$, we wish to produce a catalog that is obtained using a single classifier. Section 2.1 showed that in the magnitude range $18.5 \leq r \leq 20$, the classification by HSC-SSP is more reliable than the one by SDSS. We therefore considered the classification by SDSS in the range $15 \leq r < 18.5$ and the one by HSC-SSP in the range $18.5 \leq r \leq 23.5$. This catalog spans the magnitude range $15 \leq r \leq 23.5$ and features a total of 11 763 sources, 9517 galaxies and 2246 stars. We call it XMATCH catalog.

Next, we trained and tested all the models on this catalog. When only photometric information is used, the best classifier is RF, which reaches AUC = 0.957 ± 0.008; this is close to the performance of SGLC, which uses morphological information. When photometric and morphological information is used, the best classifier is ERT, which with AUC = 0.986 ± 0.005 outperforms SGLC. In Appendix E we show the analysis that considers only morphological information. We find that RF and ANN yield AUC = 0.970 ± 0.006. Figure 11 shows the ROC curve and the purity curve for galaxies and stars for the three classifiers above, with the addition of the probability threshold $p_{\rm cut}$ through color-coding. These plots are meant to help choose the probability threshold that best satisfies the completeness and purity requirements (see also Appendix B). These plots were made with the code available online[17]. As shown in the bottom panel of Fig. 11, the AP of stars is quite good (and significantly better than SGLC), showing that the fact that we used an unbalanced set did not affect the results for the least represented class.

[17] github.com/PedroBaqui/minijpas-astroclass

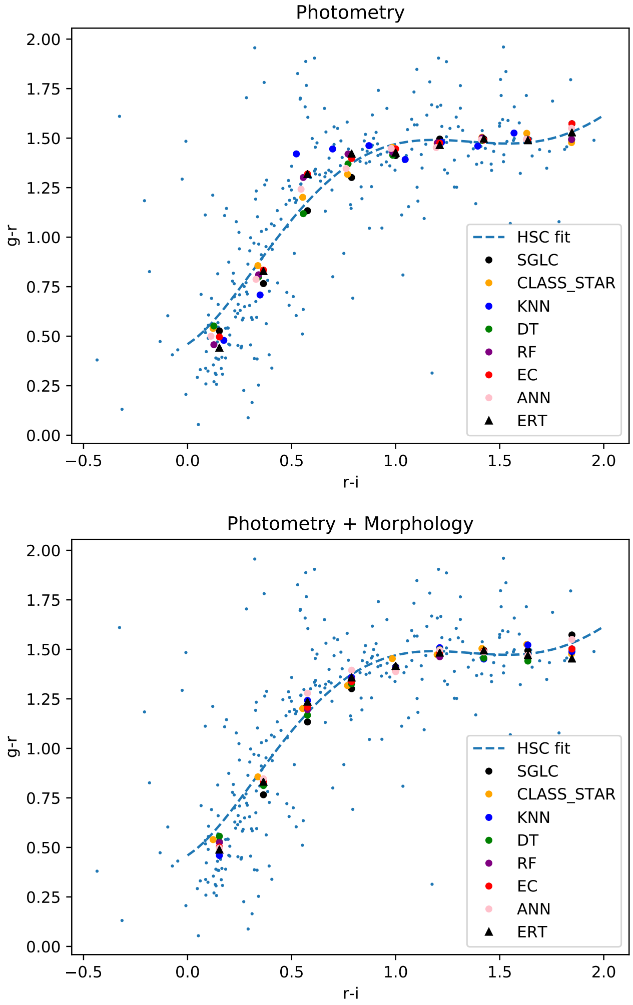

Fig. 10. Stellar locus characterization. The small blue dots represent the stellar locus for the objects classified as stars ($p \leq p_{\mathrm{cut}} = 0.5$) of the miniJPAS catalog cross-matched with the HSC-SSP catalog in the magnitude interval $18.5 \leq r \leq 23.5$. The dashed line represents a polynomial fit to the stellar locus. *Top panel*: analysis that uses only photometric bands, and *bottom panel*: analysis that also uses morphological information. The larger colored symbols represent the mean stellar locus provided by the different ML models. For comparison we also show the classification by CLASS_STAR and SGLC, which always use morphological parameters.

Finally, we show in Fig. 12 the cumulative purity of the galaxy and star samples as a function of r magnitude for a fixed completeness of 95% and 99%, which were achieved by choosing a suitable p_{cut}. For a completeness of 95% and the ERT classifier, the purity of the galaxy sample remains higher than 99% throughout the magnitude range, which is better than SGLC. Regarding stars, for a completeness of 95% and ERT, purity remains higher that 90% for $r < 22.5$. For fainter stars, ERT outperforms SGLC.

In order to build our catalog, we applied our two best classifiers (RF without morphology and ERT with morphology) to the 29 551 miniJPAS sources in the magnitude range $15 \leq r \leq 23.5$. It is important to note that given the completeness of miniJPAS (see Bonoli et al. 2020), sources outside this magnitude interval are less likely to enter scientific studies. The catalog is publicly available[18] via the ADQL table `minijpas.StarGalClass`. See Appendix D for more information and an ADQL query example.

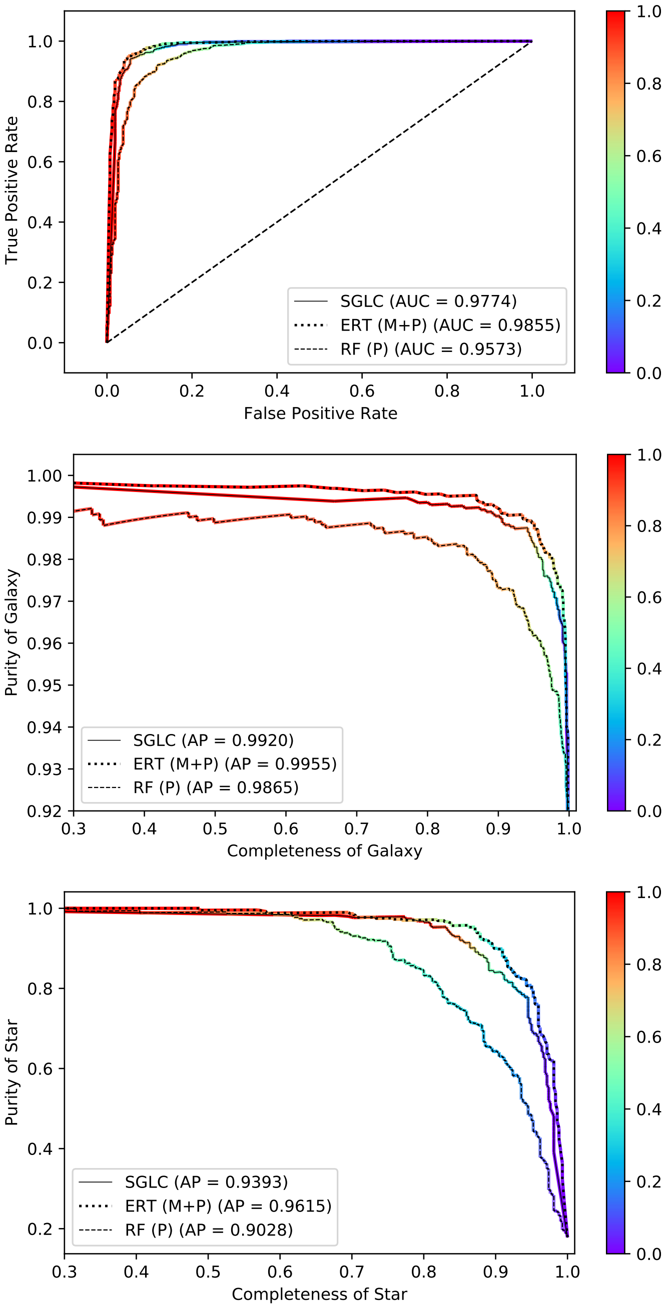

Fig. 11. ROC curve (*top panel*) and purity curve for galaxies (*middle panel*) and stars (*bottom panel*) for RF (no morphology), ERT (with morphology), and SGLC for sources in the magnitude range $15 \leq r \leq 23.5$. The color-coding indicates the probability threshold p_{cut}. The axes ranges are varied in order to better show the curves.

5.4. Feature importance

We used the RF algorithm (see Eq. (4)) to assess feature importance, which can give us insights on the way objects are classified. The 15 most important features are listed in Table 2. The full tables are provided as machine-readable supplementary material.

When morphological parameters are included, the FWHM is the most important feature. This agrees with the distributions

[18] http://j-pas.org/datareleases

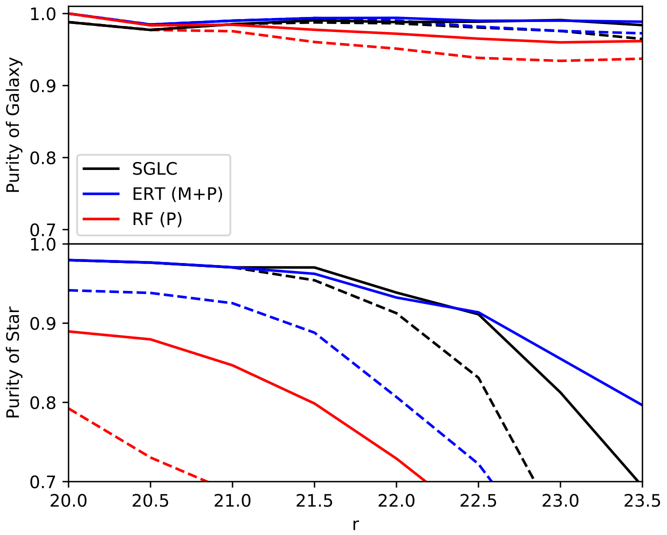

Fig. 12. Cumulative purity of the galaxy (*top*) and star (*bottom*) samples as a function of magnitude for the ML classifiers of Fig. 11 for a fixed completeness of 95% (solid line) and 99% (dashed line).

Table 2. Feature importance with (M + P) and without (P) morphological parameters for the analysis of the full cross-matched catalog XMATCH ($15 \leq r \leq 23.5$, see Sect. 5.3).

XMATCH (P)		XMATCH ($P + M$)	
Feature	Importance	Feature	Importance
iSDSSerr	1.00	FWHM	1.00
J0810err	0.31	c_r	0.30
J0390	0.22	max/ap3	0.18
J0460err	0.18	PSF	0.10
J0680	0.18	iSDSSerr	0.08
rSDSSerr	0.14	J0820err	0.02
J1007err	0.12	J0390err	0.02
J0820err	0.09	A/B	0.01
gSDSSerr	0.09	J1007err	0.01
iSDSS	0.08	J0810err	0.01
J0720	0.08	J0390	0.01
J0660err	0.07	gSDSS	0.009
uJAVA	0.05	uJAVAerr	0.008
J1007	0.05	J0790err	0.008
uJPAS	0.05	J0680	0.007
...	...	...	...

Notes. The importance is normalized relative to the best feature. The quantity max/ap3 is `MU_MAX/MAG_APER_3_0`. The full table is available at the CDS as machine-readable supplementary material. See also Fig. 13.

of the FWHM in Figs. 5 and 6, which show a good separation between stars and galaxies. Although this separation is less evident for the other parameters, they also contribute to the classification. In particular, the mean PSF is the fourth most important feature, while the least important morphological feature is the ellipticity parameter A/B. To some extent, these results might depend on the choice of the impurity function (see Eq. (3)). We tested different impurity functions and confirmed that morphological parameters are generally more important than photometric bands.

When only photometric information is used, the importance of the features is more evenly distributed as more features work together toward object classification. In particular, broad bands are not necessarily more important than narrow bands, and errors (the width of the distribution) are as important as the measurements (central value of the distribution). In other words, the full characterization of the measurement seems to be important.

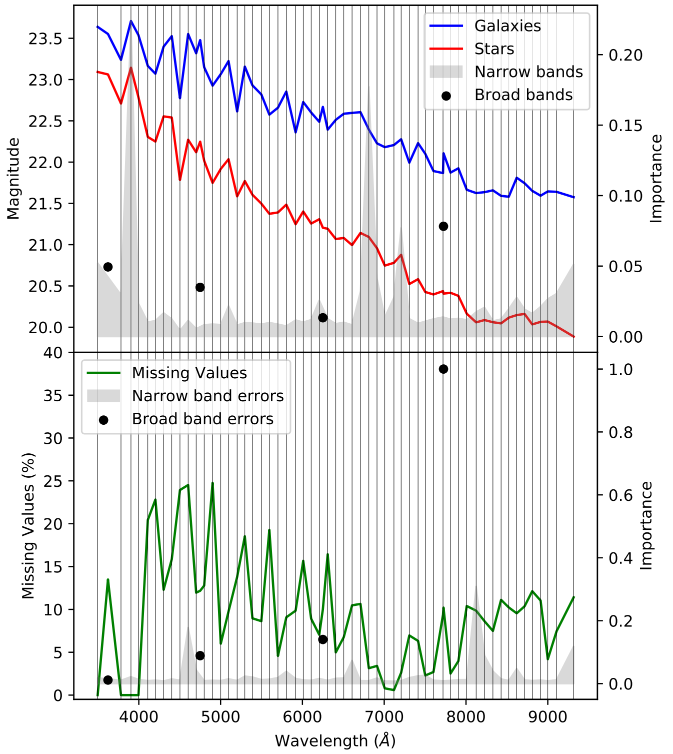

Fig. 13. Feature importance and average J-spectra. *Top*: the shaded area represents the relative importance (see Eq. (4)) of the narrow-band filters as function of the filter wavelength for the analysis relative to the full magnitude range $15 \leq r \leq 23.5$ (see Sect. 5.3). The importance of the four broad-band filters is shown using black circles. The red and blue lines show the average J-spectrum of stars and galaxies, respectively. *Bottom*: same as the top panels, but for the relative importance of the magnitude errors. The green line shows the percentage of missing values (magnitude of 99) for the narrow-band filters.

In order to obtain physical insight into the regions of the spectrum that matter most for classification, we show in Fig. 13 (top) the relative importance of the filter magnitudes as a function of the filter wavelength together with the median star and galaxy J-spectrum. It is clear that there are regions that are systematically more important than others (neighboring filters with higher importance) and that the most important regions and the average features in the spectra are correlated. In the bottom panel of Fig. 13 we show the importance of the magnitude errors, which also show regions that are systematically more important than others. The error on the i band is particularly important. In the same panel we also show the fraction of missing values (magnitude of 99) for each narrow-band filter. This fraction is anticorrelated with the filter importance (top panel).

5.5. Transmission curve variability

The transmission curves of the narrow-band filters vary according to the relative position in the filters. In particular, the transmission curve variability depends on the spectral energy distribution (SED) of each object so that the map of relative variation in flux for a given filter is different for objects with different

Table 3. Area under the curve for the four filter quadrants relative to the best classifiers shown in Fig. 11.

RF (P)	$\mathtt{X} < 4500$	$4500 \leq \mathtt{X} \leq 9000$
$\mathtt{Y} < 4500$	0.9633	0.9592
$4500 \leq \mathtt{Y} \leq 9000$	0.9449	0.9588
ERT ($P + M$)	$\mathtt{X} < 4500$	$4500 \leq \mathtt{X} \leq 9000$
$\mathtt{Y} < 4500$	0.9917	0.9775
$4500 \leq \mathtt{Y} \leq 9000$	0.9822	0.9938

SEDs. This should affect classifications that strongly depend on particular narrow spectral features (even more so when they fall at one of the edges of the narrow-band transmission curve) and would have almost no effect when mainly the continuum is considered. As we used photometric data, our results might show this effect.

miniJPAS data, in particular, the size of the XMATCH catalog, do not allow us to thoroughly investigate this effect. We therefore explored this problem by dividing the test set into the four quadrants of the filter area and computed the AUC for each quadrant. The filter coordinates are given in pixels via the X_IMAGE and Y_IMAGE variables (9000 × 9000 pixels). Table 3 shows that the AUC variation is compatible with the overall performance of AUC = 0.957 ± 0.008 (RF) and AUC = 0.986 ± 0.005 (ERT), showing that the effect probably does not bias our results strongly.

6. Conclusions

We applied different machine-learning methods for the classification of sources of miniJPAS. The goal was to build models that are competitive with and complementary to those adopted in the miniJPAS 2019 public data release and to offer to the astronomical community a value-added catalog with an alternative classification. As we considered supervised ML algorithms, we classified the miniJPAS objects that are in common with SDSS and HSC-SSP, whose classifications are trustworthy within the magnitude intervals $15 \leq r \leq 20$ and $18.5 \leq r \leq 23.5$, respectively. We used the magnitudes associated with the 60 filters along with their errors, four morphological parameters, and the mean PSF of the pointings as input. The output of the algorithms is probabilistic. We tested K-nearest neighbors, decision trees, random forest, artificial neural networks, extremely randomized trees, and ensemble classifier.

Our results show that ML is able to classify objects into stars and galaxies without the use of morphological parameters. This makes ML classifiers quite valuable as they can distinguish compact galaxies from stars, differently from methods that require using morphological parameters in the classification process. Including morphological parameters improves the results to the point that ERT can outperform CLASS_STAR and SGLC (the default classifier in J-PAS).

We used the RF algorithm to assess feature importance. When morphological parameters are used, the FWHM is the most important feature. When only photometric information is used, broad bands are not necessarily more important than narrow bands, and errors (the width of the distribution) are as important as the measurements (central value of the distribution). In other words, the full characterization of the measurement seems to be important. We also showed that ML can give meaningful insights into the regions of the spectrum that matter most for classification.

After validating our methods, we applied our best classifiers, with and without morphology, to the full dataset. This classification is available as a value-added catalog[19] via the ADQL table minijpas.StarGalClass. The ML models are available online[20]. Our catalog both validates the quality of SGLC and produces an independent classification that can be useful to test the robustness of subsequent scientific analyses. In particular, our classification uses the full photometric information, with and without morphology, which is important for faint galaxies whose morphology is similar to that of stars.

We conclude by stressing that our method can be further improved both at the algorithmic and at the data input level. A promising avenue is the direct use of the object images with convolutional neural networks. This approach has the potential of outperforming currently available classifiers.

Acknowledgements. POB thanks, for financial support, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. VM thanks CNPq (Brazil) and FAPES (Brazil) for partial financial support. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 888258. LADG is supported by the Ministry of Science and Technology of Taiwan (grant MOST 106-2628-M-001-003-MY3), and by the Academia Sinica (grant AS-IA-107-M01). ES has been partly supported by the Spanish State Research Agency (AEI) Projects AYA2017-84089 and MDM-2017-0737 at Centro de Astrobiología (CSIC-INTA), Unidad de Excelencia María de Maeztu. MQ is supported by the Brazilian research agencies CNPq and FAPERJ. RGD acknowledges financial support from the State Agency for Research of the Spanish MCIU through the "Center of Excellence Severo Ochoa" award to the Instituto de Astrofísica de Andalucía (SEV-2017-0709) and through the projects AyA2016-77846-P and PID2019-109067GB-100. LS acknowledges support from Brazilian agencies CNPq (grant 304819/2017-4) and FAPESP (grant 2012/00800-4). This work made use of the Virgo Cluster at Cosmo-ufes/UFES, which is funded by FAPES and administrated by Renan Alves de Oliveira. Based on observations made with the JST/T250 telescope and PathFinder camera for the miniJPAS project at the Observatorio Astrofísico de Javalambre (OAJ), in Teruel, owned, managed, and operated by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA). We acknowledge the OAJ Data Processing and Archiving Unit (UPAD) for reducing and calibrating the OAJ data used in this work. Funding for OAJ, UPAD, and CEFCA has been provided by the Governments of Spain and Aragón through the Fondo de Inversiones de Teruel; the Aragón Government through the Research Groups E96, E103, and E16_17R; the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, UE) with grant PGC2018-097585-B-C21; the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) under AYA2015-66211-C2-1-P, AYA2015-66211-C2-2, AYA2012-30789, and ICTS-2009-14; and European FEDER funding (FCDD10-4E-867, FCDD13-4E-2685). Based on data from ALHAMBRA Data Access Service the at CAB (CSIC-INTA). Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G. The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from the Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University. This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at dm.lsst.org. This paper is based [in part] on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center (ADC) at National Astronomical Observatory of Japan. Data analysis was in

[19] http://j-pas.org/datareleases

[20] github.com/J-PAS-collaboration/StarGalClass-MachineLearning

part carried out with the cooperation of Center for Computational Astrophysics (CfCA), National Astronomical Observatory of Japan. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III website is 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 Astrofisica 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.

References

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, PASJ, 71, 114
Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Altman, N. 1992, Am. Stat., 46, 175
Amorín, R. O., Pérez-Montero, E., & Vílchez, J. M. 2010, ApJ, 715, L128
Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342
Benitez, N., Dupke, R., Moles, M., et al. 2014, ArXiv e-prints [arXiv:1403.5237]
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bilicki, M., Hoekstra, H., Brown, M. J. I., et al. 2018, A&A, 616, A69
Biswas, R., Blackburn, L., Cao, J., et al. 2013, Phys. Rev. D, 88, 062003
Bonoli, S., Marín-Franch, A., Varela, J., et al. 2020, ArXiv e-prints [arXiv:2007.01910]
Breiman, L. 2001, Mach. Learn., 45, 5
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1984, Int. Group, 432, 151
Cabayol, L., Sevilla-Noarbe, I., Fernández, E., et al. 2019, MNRAS, 483, 529
Cardamone, C., Schawinski, K., Sarzi, M., et al. 2009, MNRAS, 399, 1191
Carrillo, M., González, J. A., Gracia-Linares, M., & Guzmán, F. S. 2015, J. Phys. Conf. Ser., 654, 012001
Cavuoti, S., Brescia, M., Tortora, C., et al. 2015, MNRAS, 452, 3100
Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2010, in The Javalambre Astrophysical Observatory project, SPIE Conf. Ser., 7738, 77380V
Charnock, T., & Moss, A. 2017, Supernovae: Photometric Classification of Supernovae
Costa-Duarte, M. V., Sampedro, L., Molino, A., et al. 2019, MNRAS, submitted [arXiv:1909.08626]
Davis, M., Guhathakurta, P., Konidaris, N., et al. 2007, ApJ, 660, L1
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
Díaz-García, L. A., Cenarro, A. J., López-Sanjuan, C., et al. 2019, A&A, 631, A156
Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15
Garofalo, M., Botta, A., & Ventre, G. 2016, Proc. Int. Astron. Union, 12, 345
Gauci, A., Adami, K. Z., Abela, J., & Magro, A. 2010, ArXiv e-prints [arXiv:1005.0390]
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Mach. Learn., 63, 3
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data mining, Inference, and Prediction (Springer Science & Business Media)
Henrion, M., Mortlock, D. J., & Hand, D. J. 2011, MNRAS, 412, 2286
Ishak, B. 2017, Contemp. Phys., 58, 99
Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463
Kim, E. J., Brunner, R. J., & Carrasco Kind, M. 2015, MNRAS, 453, 507
Le Fevre, O., Crampton, D., Lilly, S. J., Hammer, F., & Tresse, L. 1995, ApJ, 455, 60
Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31
López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, A&A, 622, A177
Marshall, P., Anguita, T., Bianco, F. B., et al. 2017, https://doi.org/10.5281/zenodo.842713
Matthews, D. J., Newman, J. A., Coil, A. L., Cooper, M. C., & Gwyn, S. D. J. 2013, ApJS, 204, 21
Mitchell, T. M. 1997, Machine Learning (McGraw-Hill)
Moles, M., Benítez, N., Aguerri, J. A. L., et al. 2008, AJ, 136, 1325
Molino, A., Benítez, N., Moles, M., et al. 2014, MNRAS, 441, 2891
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5
Odewahn, S. C., de Carvalho, R. R., Gal, R. R., et al. 2004, AJ, 128, 3092
Pedregosa, F., et al. 2011, J. Mach. Learn. Res., 12, 2825
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533
Sevilla-Noarbe, I., Hoyle, B., Marchã, M. J., et al. 2018, MNRAS, 481, 5451
Vargas dos Santos, M., Quartin, M., & Reis, R. R. 2020, MNRAS, 497, 2974
Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, AJ, 141, 189
Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2019, A&A, 622, A182
Wolpert, D. H. 1996, Neural Comput., 8, 1341

[1] PPGFis & Núcleo de Astrofísica e Cosmologia (Cosmo-ufes), Universidade Federal do Espírito Santo, 29075-910 Vitória, ES, Brazil
e-mail: marra@cosmo-ufes.org
[2] PPGCosmo & Departamento de Física, Universidade Federal do Espírito Santo, 29075-910 Vitória, ES, Brazil
[3] INAF – Osservatorio Astronomico di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[4] IFPU – Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy
[5] Departamento de Física, Universidade Federal de Sergipe, 49100-000 Aracaju, SE, Brazil
[6] Donostia International Physics Center (DIPC), Manuel Lardizabal Ibilbidea, 4, San Sebastián, Spain
[7] Ikerbasque, Basque Foundation for Science, 48013 Bilbao, Spain
[8] Academia Sinica Institute of Astronomy & Astrophysics (ASIAA), 11F of Astronomy-Mathematics Building, AS/NTU, No. 1, Section 4, Roosevelt Road, Taipei 10617, Taiwan
[9] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Unidad Asociada al CSIC, Plaza San Juan, 1, 44001 Teruel, Spain
[10] Instituto de Astrofísica de Canarias, C/Vía Láctea, s/n, 38205 La Laguna, Tenerife, Spain
[11] Departamento de Astrofísica, Universidad de La Laguna, 38206 La Laguna, Tenerife, Spain
[12] Observatório do Valongo, Universidade Federal do Rio de Janeiro, 20080-090 Rio de Janeiro, RJ, Brazil
[13] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan 1, 44001 Teruel, Spain
[14] NSF's Optical-Infrared Astronomy Research Laboratory, Tucson, AZ 85719, USA
[15] Instituto de Física, Universidade Federal do Rio de Janeiro, 21941-972 Rio de Janeiro, RJ, Brazil
[16] Instituto de Física, Universidade de São Paulo, 05508-090 São Paulo, SP, Brazil
[17] Physics Department, Lancaster University, Lancashire, UK
[18] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESAC Campus, Camino Bajo del Castillo s/n, 28692 Villanueva de la Cañada, Madrid, Spain
[19] Tartu Observatory, University of Tartu, Observatooriumi 1, 61602 Tõravere, Estonia
[20] Instituto de Astrofísica de Andalucía – CSIC, Apdo 3004, 18080 Granada, Spain
[21] Observatório Nacional, Ministério da Ciencia, Tecnologia, Inovação e Comunicações, 20921-400 Rio de Janeiro, RJ, Brazil
[22] Instituto de Física, Universidade Federal da Bahia, 40210-340 Salvador, BA, Brazil
[23] Departamento de Física-CFM, Universidade Federal de Santa Catarina, 88040-900 Florianópolis, SC, Brazil
[24] Departamento de Astronomia, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, 05508-090 São Paulo, SP, Brazil
[25] Department of Astronomy, University of Michigan, 311West Hall, 1085 South University Ave., Ann Arbor, USA
[26] Department of Physics and Astronomy, University of Alabama, Box 870324, Tuscaloosa, AL, USA
[27] Instruments4, 4121 Pembury Place, La Cañada Flintridge, CA 91011, USA

Appendix A: Purity curves for stars

For completeness we report in Fig. A.1 the purity curves for the stars. For a comparison see, for example, Sevilla-Noarbe et al. (2018), Fadely et al. (2012) and Cabayol et al. (2019).

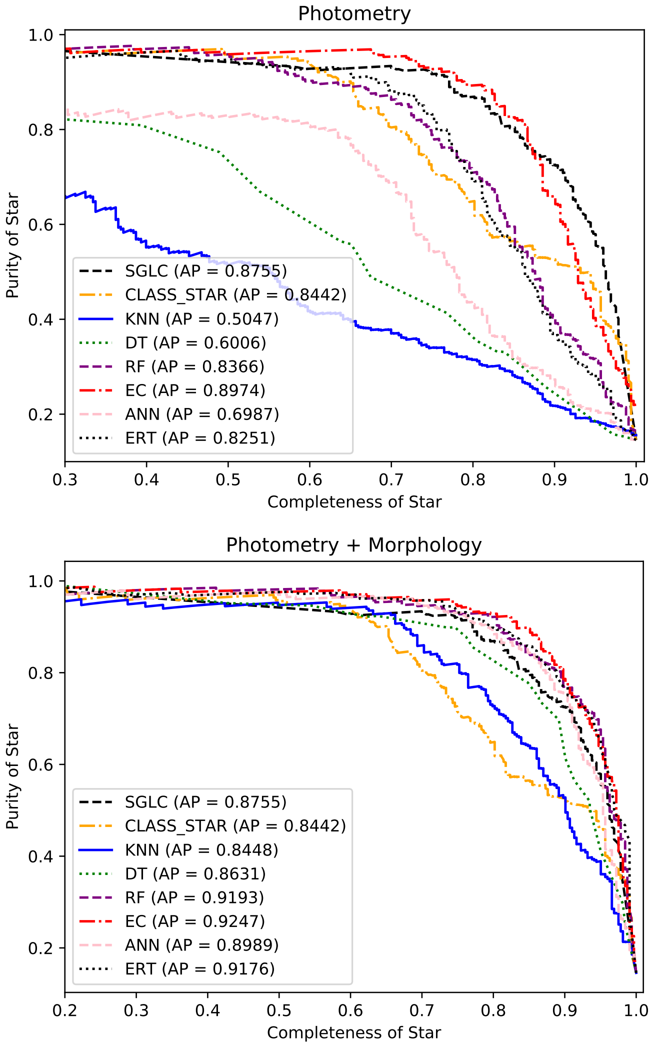

Fig. A.1. Purity curves for stars using J-PAS data with HSC-SSP classification. *Top panel*: uses only photometric information, and *bottom panel*: also uses morphology. For comparison we show the classification by CLASS_STAR and SGLC, which always use morphological parameters. The axes ranges are varied in order to show the curves better.

Appendix B: Classification vs. probability threshold

We show in Fig. B.1 the histograms of the probabilities that the objects received from the classifiers. In red we plot objects that are classified as galaxies, and in blue we show stars. These plots allow us to assess the performance of the algorithms from a different point of view. When a value for p_{cut} is chosen, all the objects to the right of this value are classified as galaxies, and those with lower probability are classified as stars. It is then clear that an ideal algorithm should have well-separated and nonintersecting probability distributions for stars and galaxies.

The addition of morphology makes the distributions tighter and causes fewer intersections. Similar results were obtained with CFHTLenS data (Kim et al. 2015). We observe instead that the probability distribution of galaxies for CLASS_STAR is more concentrated than the probability distribution for stars. This leads us to the conclusion that CLASS_STAR has a tendency to classify galaxies better than stars. It is also clear that by varying p_{cut}, we can sacrifice the completeness of the dataset in favor of a higher purity of galaxies.

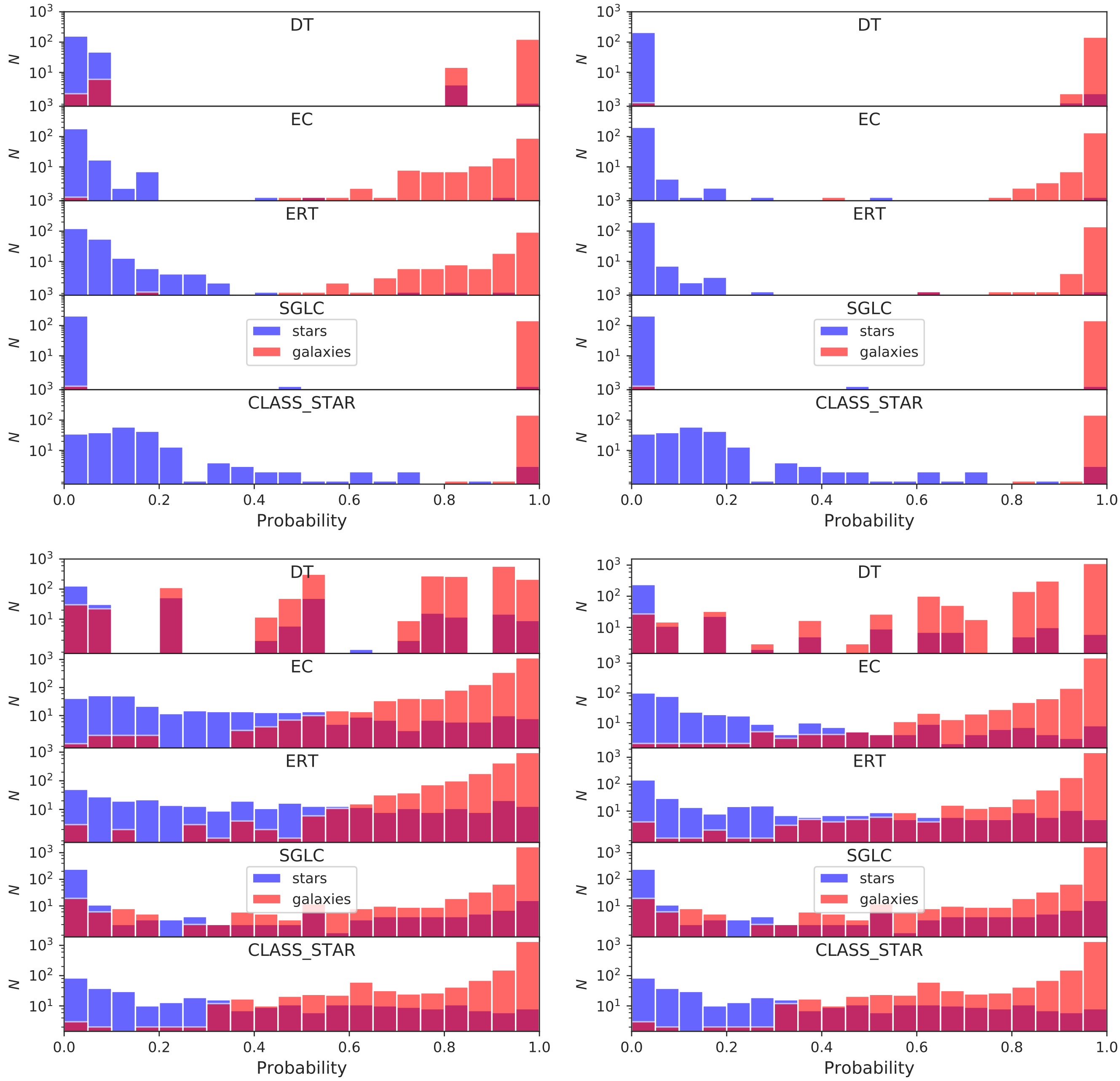

Fig. B.1. Histograms of the probability that a source belongs to the galaxy class. The histograms for actual stars and galaxies as classified by SDSS (*top*) and HSC-SSP (*bottom*) are plotted in blue and semi-transparent red, respectively. The panels on the *left* use only photometric information, and those on the *right* also use morphology. For comparison we also show the classification by CLASS_STAR and SGLC, which always use morphological parameters.

Appendix C: mJP-AEGIS1 field

We explained above that miniJPAS consists of four fields, each with an approximately 0.25 deg^2 field of view (for details, see Bonoli et al. 2020). The mJP-AEGIS1 has 20016 objects and features an r-band PSF that is similar to mJP-AEGIS3 (∼0.7″) and better than mJP-AEGIS2 and mJP-AEGIS4 (∼0.8″). It is then interesting to repeat for mJP-AEGIS1 the analysis relative to HSC-SSP (see Sect. 5.2). We did not consider the analysis for the SDSS as the crossmatched catalog would be too small.

The cross match of mJP-AEGIS1 with HSC-SSP in the range $18.5 \leq r \leq 23.5$ has 4486 objects, 3809 galaxies and 677 stars. We show the results in Fig. C.1, which should be compared with the analysis that considers the full miniJPAS catalog in Fig. 9. The results for the various classifiers clearly improve, as expected. In particular, when both morphological and photometric features are considered, ERT changes from AUC = 0.979 (Fig. 9) to AUC = 0.987 (Fig. C.1).

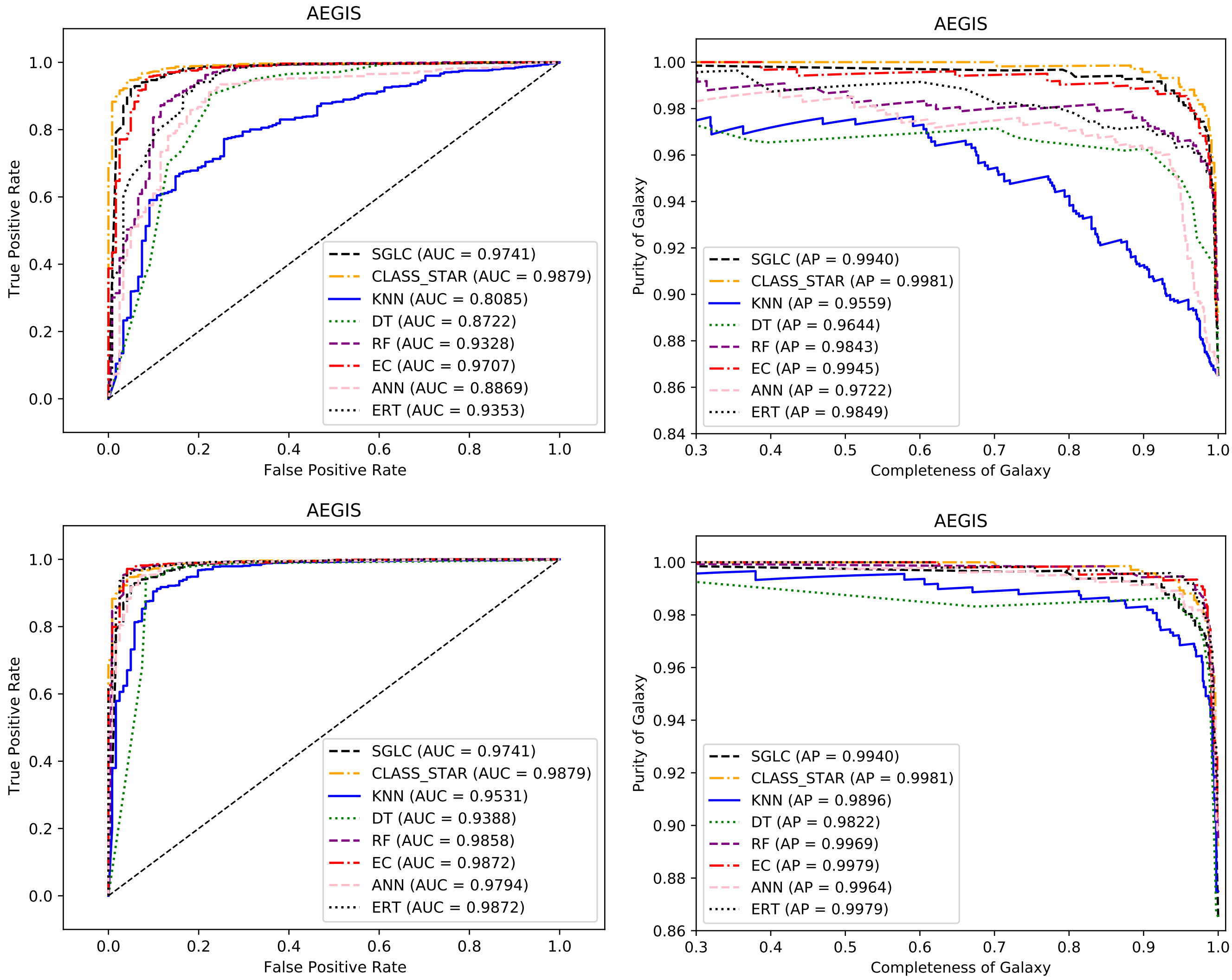

Fig. C.1. ROC curves (*left panels*) and purity curves for galaxies (*right panels*) for the classifiers considered in this paper for the AEGIS1 field cross-matched with the HSC-SSP catalog in the magnitude interval $18.5 \leq r \leq 23.5$. *Top panels*: analysis that uses only photometric bands, and *bottom panels*: analysis that uses photometric bands and morphological parameters. For comparison we also show the classification by `CLASS_STAR` and SGLC, which always use morphological parameters. The axes ranges are varied in order to show the curves better.

Appendix D: ADQL query

The value-added catalog with the ERT and RF classifications is publicly available[21] via the ADQL table `minijpas.StarGalClass`. The column `ert_prob_star` gives the probability $1 - f$ of being a star provided by the ERT classifier, using both morphological and photometric information. The column `rf_prob_star` gives the probability $1 - f$ of being a star provided by the RF classifier, using only photometric information. In order to follow the convention of the `minijpas.StarGalClass` table, we used the probability $1 - f$ of being a star. In the remaining work, we used the probability f of being a galaxy.

In order to facilitate access to our results, we now report a simple query example that allows access to the classifications generated by ML along with the miniJPAS photometric bands with flag and mask quality cuts.

```
SELECT

t1.MAG_AUTO[minijpas::uJAVA] as uJAVA,
t1.MAG_AUTO[minijpas::J0378] as J0378,
t1.MAG_AUTO[minijpas::J0390] as J0390,
t1.MAG_AUTO[minijpas::J0400] as J0400,
t1.MAG_AUTO[minijpas::J0410] as J0410,
t2.ert_prob_star,
t2.rf_prob_star

FROM

minijpas.MagABDualObj t1

JOIN

minijpas.StarGalClass t2

ON

t1.tile_id = t2.tile_id AND
t1.number=t2.number

WHERE

t1.flags[minijpas::rSDSS]=0 AND
t1.mask_flags[minijpas::rSDSS]=0
```

Appendix E: Analysis using only morphology

W applied the ML methods discussed in the main text to the XMATCH catalog using only morphological parameters as input. Figure E.1 shows the performances of the different algorithms. Comparing these results with those of Fig. 11 (morphology + photometry), we observe that the inclusion of the photometric bands in the analysis increases the performance of the models. When only morphological parameters are used, the AUC, galaxy AP, and star AP of the best pure ML classifier are 0.9696, 0.9918, and 0.9152, respectively. When morphological parameters and photometric bands are used, the AUC, galaxy AP, and star AP of the best pure ML classifier are 0.9855, 0.9955, and 0.9615, respectively.

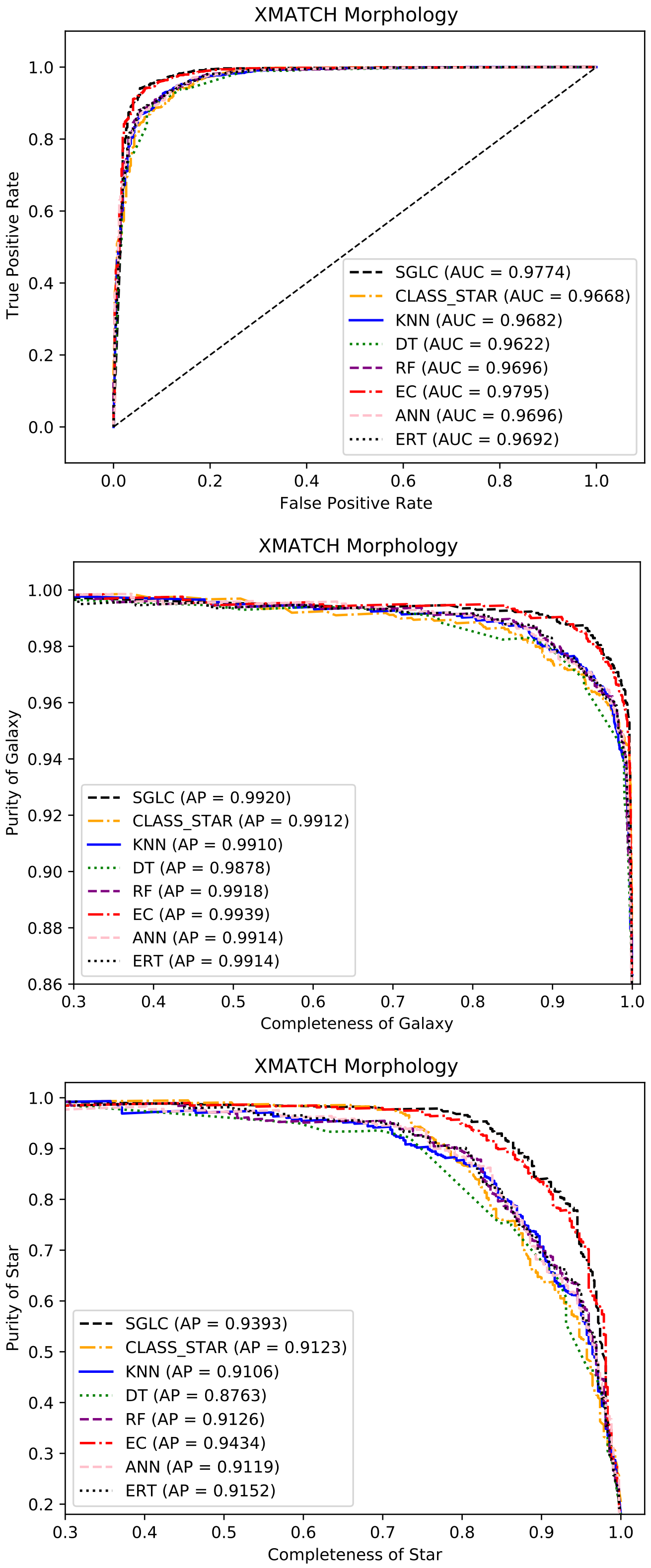

Fig. E.1. ROC curve (*top panel*) and purity curve for galaxies (*middle panel*) and stars (*bottom panel*) when only morphological parameters in the magnitude range $15 \leq r \leq 23.5$ of the XMATCH catalog are used.

[21] http://j-pas.org/datareleases

Appendix F: Hyperparameter optimization

The best values of the hyperparameters were found with the `GridSearchCV` method. They were selected according to the best mean from $k = 10$ folds using the AUC-ROC metric. Below are the hyperparameters from the XMATCH catalog; the unspecified hyperparameters take the default values by `scikit-learn` version 0.23.1. We fixed `random seed = 5`.

F.1. Using photometric bands only

KNN

`n_neighbors: 50, weights: distance`

DT

`class_weight: balanced, criterion: gini, max_depth: 5, max_features: None, random_state: 5`

RF

`n_estimators: 100, bootstrap: True, class_weight: balanced_subsample, criterion: entropy, max_depth: 10, max_features: None`

ANN

`activation: logistic, hidden_layer_sizes: 200, learning_rate: constant, solver: adam, max_iter: 200, tol=0.0001`

ERT

`n_estimators: 200, bootstrap: False, class_weight: balanced_subsample, criterion: entropy, max_depth: 20, max_features: None`

F.2. Using morphological parameters only

KNN

`n_neighbors: 100, weights: distance`

DT

`class_weight: balanced, criterion: entropy, max_depth: 5, max_features: None`

RF

`bootstrap: True, class_weight: balanced_subsample, criterion: entropy, max_depth: 5, max_features: None, 'n_estimators': 100`

ANN

`activation: relu, hidden_layer_sizes: 200, learning_rate: constant, solver: adam, max_iter: 200, tol=0.0001`

ERT

`bootstrap: False, class_weight: balanced_subsample, criterion: entropy, max_depth: 10, max_features: None, n_estimators: 200`

F.3. Using photometric bands together with morphology parameters

KNN

`n_neighbors: 100, weights: distance`

DT

`class_weight: balanced, criterion: entropy, max_depth: 5, max_features: None`

RF

`n_estimators: 100, bootstrap: True, class_weight: balanced_subsample, criterion: entropy, max_depth: 20, max_features: None`

ANN

`activation: logistic, hidden_layer_sizes: 200, learning_rate: constant, solver: sgd, max_iter: 200, tol=0.0001`

ERT

`n_estimators: 200, bootstrap: False, class_weight: balanced_subsample, criterion: entropy, max_depth: 20, max_features: None`