The miniJPAS survey quasar selection II: Machine learning classification with photometric measurements and uncertainties

Natália V.N. Rodrigues, L. Raul Abramo, Carolina Queiroz, Ginés Martínez-Solaeche, Ignasi Pérez Ràfols, Silvia Bonoli, Jonás Chaves-Montero, Matthew M. Pieri, Rosa M. González Delgado, Sean S. Morrison, et al.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **520,** 3494–3509 (2023) https://doi.org/10.1093/mnras/stac2836

The miniJPAS survey quasar selection – II. Machine learning classification with photometric measurements and uncertainties

Natália V. N. Rodrigues,[1★] L. Raul Abramo,[1] Carolina Queiroz,[1,2] Ginés Martínez-Solaeche,[3] Ignasi Pérez-Ràfols,[4,5] Silvia Bonoli,[6,7] Jonás Chaves-Montero,[6] Matthew M. Pieri,[8] Rosa M. González Delgado,[3] Sean S. Morrison,[8,9] Valerio Marra,[10,11] Isabel Márquez,[3] A. Hernán-Caballero,[12] L. A. Díaz-García,[3] Narciso Benítez,[3] A. Javier Cenarro,[13] Renato A. Dupke,[14,15,16] Alessandro Ederoclite,[12] Carlos López-Sanjuan,[13] Antonio Marín-Franch,[13] Claudia Mendes de Oliveira,[17] Mariano Moles,[3,12] Laerte Sodré, Jr,[17] Jesús Varela,[13] Héctor Vázquez Ramió[13] and Keith Taylor[18]

[1]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, Rua do Matão 1371, CEP 05508-090, São Paulo, Brazil*
[2]*Departamento de Astronomia, Instituto de Física, Universidade Federal do Rio Grande do Sul (UFRGS), Avenida Bento Gonçalves 9500, Porto Alegre, RS, Brazil*
[3]*Instituto de Astrofísica de Andalucía (CSIC), PO Box 3004, E-18080 Granada, Spain*
[4]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, E-08193 Bellaterra (Barcelona), Spain*
[5]*Laboratoire de Physique Nucléaire et de Hautes Energies, Sorbonne Université, Université Paris Diderot, CNRS/IN2P3, LPNHE, 4 Place Jussieu, F-75252 Paris, France*
[6]*Donostia International Physics Center, Paseo Manuel de Lardizabal 4, E-20018 Donostia-San Sebastian, Spain*
[7]*Ikerbasque, Basque Foundation for Science, E-48013 Bilbao, Spain*
[8]*Aix-Marseille University, CNRS, CNES, LAM, Marseille, France*
[9]*Department of Astronomy, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA*
[10]*INAF, Osservatorio Astronomico di Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[11]*IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy*
[12]*Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan, 1, E-44001 Teruel, Spain*
[13]*Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Unidad Asociada al CSIC, Plaza San Juan 1, E-44001 Teruel, Spain*
[14]*Observatório Nacional/MCTI, Rua General José Cristino, 77, São Cristóvão, CEP 20921-400, Rio de Janeiro, Brazil*
[15]*Department of Astronomy, University of Michigan, 311 West Hall, 1085 South University Avenue, Ann Arbor, MI, USA*
[16]*Department of Physics and Astronomy, University of Alabama, Gallalee Hall, Tuscaloosa, AL 35401, USA*
[17]*Depto. de Astronomia, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão, 1226, CEP 05508-090, São Paulo, Brazil*
[18]*Instruments4, 4121 Pembury Place, La Canada Flintridge, CA 91011, USA*

Accepted 2022 September 27. Received 2022 September 26; in original form 2022 August 22

ABSTRACT

Astrophysical surveys rely heavily on the classification of sources as stars, galaxies, or quasars from multiband photometry. Surveys in narrow-band filters allow for greater discriminatory power, but the variety of different types and redshifts of the objects present a challenge to standard template-based methods. In this work, which is part of a larger effort that aims at building a catalogue of quasars from the miniJPAS survey, we present a machine learning-based method that employs convolutional neural networks (CNNs) to classify point-like sources including the information in the measurement errors. We validate our methods using data from the miniJPAS survey, a proof-of-concept project of the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) collaboration covering $\sim$1 deg^2 of the northern sky using the 56 narrow-band filters of the J-PAS survey. Due to the scarcity of real data, we trained our algorithms using mocks that were purpose-built to reproduce the distributions of different types of objects that we expect to find in the miniJPAS survey, as well as the properties of the real observations in terms of signal and noise. We compare the performance of the CNNs with other well-established machine learning classification methods based on decision trees, finding that the CNNs improve the classification when the measurement errors are provided as inputs. The predicted distribution of objects in miniJPAS is consistent with the putative luminosity functions of stars, quasars, and unresolved galaxies. Our results are a proof of concept for the idea that the J-PAS survey will be able to detect unprecedented numbers of quasars with high confidence.

Key words: quasars: general – methods: data analysis – techniques: photometric – cosmology: observations.

★ E-mail: natalia.villa.rodrigues@usp.br

1 INTRODUCTION

Galaxy surveys have evolved to tackle a broad range of fundamental questions, from dark energy and neutrino masses to galaxy evolution and the halo–galaxy connection (Cole et al. 2005; Blake et al. 2012; Hikage et al. 2019; Alam et al. 2021; DES Collaboration 2021). Technological advances and investment in new instruments have amplified the scope of these surveys, which demand increasingly sophisticated toolboxes for data reduction, statistical analysis, and phenomenology.

The first step in any survey is finding luminous sources behind the foregrounds of the sky and the Milky Way – a task that is often performed using optical data. Typically, a large number of sources are detected using photometry in broad optical filters, but only a small fraction of those sources are then selected for spectroscopic follow-up observations. This target selection can be made on the basis of the multiband photometry, by inspecting variability in the time domain (Morganson et al. 2015; Ivezić et al. 2019), by cross-matching the sources with other wavelengths (Jansen et al. 2001; Wright et al. 2010), or by some combination thereof. In fact, the decision process about which of those luminous sources are likely to be the kinds of objects of interest to a given survey is the crucial first step that determines how we employ valuable resources, such as a multi-object spectrograph on a large telescope.

The Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al. 2014) was designed to take multiband photometry in narrow filters (of width ∼100 Å) of all sources in its field of view, providing low-resolution spectra ($R \sim 60$) in the interval $3500\,\text{Å} \lesssim \lambda \lesssim 9000\,\text{Å}$ – in that context, see also Wolf et al. (2003) and Martí et al. (2014) for other narrow-band surveys. The *science verification* phase of the survey, *miniJPAS* (Bonoli et al. 2021), achieved 5σ limiting magnitudes (for an aperture of 3 arcsec) of approximately ∼23–24 for the broad-bands (u, g, r, and i), and between ∼22 and 23 for the narrow bands. MiniJPAS has demonstrated that optical 'pseudo-spectra' are often sufficient to determine with high confidence whether an object is a star, a galaxy, a quasar, or some other type of source – and, in the case of extragalactic sources, to determine the redshifts of those objects with sub-per cent precision.

However, even with exquisite photometry a precise determination of the classes of very large numbers (millions or even billions) of objects is a challenge to established methods such as magnitude and/or colour cuts, as well as techniques that rely on template fitting (Takada et al. 2014; Dawson et al. 2016). This is particularly problematic in the case of rare objects such as quasars, which can be drowned by the heaps of stars and galaxies that constitute the bulk of sources in photometric surveys (Myers et al. 2015; Dwelly et al. 2017).

Given the advantages of narrow-band photometry to classify astrophysical sources, and in particular objects with strong emission lines such as quasars (Chaves-Montero et al. 2017), the J-PAS and WEAVE-QSO (Pieri et al. 2016) surveys have partnered to produce the largest, most complete high-redshift quasar survey to date. The goal is to build a near-complete sample of quasars identified with the help of the J-PAS multiband photometry (hereafter, J-spectra), targeting in particular the $z \geq 2.1$ quasars for follow-up using the WEAVE multi-object spectrograph (Dalton 2016). The WEAVE instrument will confirm whether those objects are really quasars, helping refine the J-PAS classification and redshift estimates. WEAVE will also be able to measure the Ly α absorption systems along the lines of sight to those high-redshift quasars, providing crucial information about the large-scale structures along those lines of sight. This data set, which will eventually cover approximately 6000 deg^2, will allow us to compute the clustering of matter using both the Ly α systems and the quasars themselves, measuring distances using the baryon acoustic oscillation scale and imposing constraints on cosmological parameters at high redshifts.

In this paper, we show how machine learning (ML) techniques can be used to classify astrophysical objects using as input data the J-spectra yielded by multiband photometric data, including the measurement errors. Here, we employ only photometric features such as the fluxes and their associated errors. Additional features, such as morphology, time domain, or other ancillary data, were not included in our analysis at this moment.

The main innovation in this paper is a systematic inclusion of information about the uncertainties in the fluxes, which are key ingredients of any measurements, but are often ignored in ML applications that take scientific data as input (Reis, Baron & Shahaf 2018; Baqui et al. 2021; Villacampa-Calvo et al. 2021; Shy et al. 2022). Here, we focus on convolutional neural networks (CNNs; LeCun et al. 1989), which have been developed primarily as tools to extract features from two-dimensional (2D) images (Simonyan & Zisserman 2014). CNNs have also been employed for classification purposes in astrophysics due to its general ability to detect features in images (Burke et al. 2019; Pasquet et al. 2019), on multiband photometric data (Sharma et al. 2020), and even in the time-spectral domain (Qu et al. 2021). It is straightforward to apply CNNs to sequential data, and to incorporate the information about measurement errors – for a general description of the technique, see also Rodrigues, Abramo & Hirata (2021). In order to compare our CNN-based techniques with other well-established ML classification methods, we have also tested the performance of random forests (RFs; Breiman 2001) and the light gradient boosting machine (LightGBM; Ke et al. 2017), two powerful decision tree (DT; Breiman et al. 1984)-based algorithms.

This work is part of a larger effort to classify miniJPAS point-like sources. The first paper (Queiroz et al. 2022) describes the construction of simulated data sets (mocks) that we used to train our algorithms, and in this paper we apply CNN and DT-based ML models to those mocks. In particular, we present a technique that enables us to take into account the measurement errors in the J-spectra. We evaluate the performances of the classifiers not only with respect to validation data sets, but also for the real miniJPAS point sources, by comparing the numbers of objects with those expected from the luminosity functions (LFs) in different magnitude ranges, redshift ranges, and for the different stellar types. Finally, we test the robustness of the classification against changes in the training sets, and we perform a feature importance analysis to evaluate which miniJPAS filters are more relevant to distinguish between the different classes. In a closely related work, Martínez-Solaeche et al. (in preparation) focus on a class of well-established ML models, the artificial neural networks (NNs), to explore different input features as well as to implement data augmentation techniques that introduce hybrid objects (ad-mixtures of single, pure populations) and study how this affects the confidence of the classification. In another forthcoming paper (Pérez-Ràfols et al., in preparation), a spectral fitting method (Pérez-Ràfols et al. 2020) is used to estimate the probability that an object is a quasar at a given redshift. Finally, in Pérez-Ràfols et al. (in preparation), we will show how to combine all the previous classifiers, as well as any additional external information, into a 'consensus' catalogue of stars, galaxies, and low-redshift ($z < 2.1$) and high-redshift ($z \geq 2.1$) quasars. That combined classification will constitute the final output of our mocks and of our suite of ML techniques, and will be validated with the help of the spectroscopically confirmed miniJPAS sources (the 'truth table').

The paper is organized as follows: In Section 2, we describe the real and mock data sets. In Section 3, we introduce the methodology and the ML algorithms. In Section 4, we evaluate the performance of the models and present the results when the methods are applied to the mock test sets. In Section 5, we show the results for the point sources in the miniJPAS data. Finally, in Section 6 we draw our main conclusions, and give perspectives for future improvements and applications.

2 DATA

In this section, we describe the miniJPAS data sample, and briefly introduce the mocks used to train and validate the ML models – a full description of the method used in the construction of the mocks, as well as tests used to compare them to the miniJPAS data, can be found in Queiroz et al. (2022).

2.1 The J-PAS and miniJPAS surveys

J-PAS is soon starting full survey operations, using a 1.2-Gpixel camera mounted on a telescope with a 2.55-m mirror and a field of view of 4.2 deg^2 (Benitez et al. 2014). The J-PAS photometric system (Marín-Franch et al. 2012) consists of 54 narrow-band filters and 2 medium-band filters (named uJAVA and J1007). In 2020, before the full instrument was completed, the J-PAS Pathfinder camera conducted an ∼1-deg^2 science verification survey (the miniJPAS survey) on the area of the All-wavelength Extended Groth Strip International Survey (AEGIS; Davis et al. (2007)). In addition to the narrow-band and medium-band filters, miniJPAS includes four Sloan Digital Sky Survey (SDSS)-like filters u, g, r, and i (total of 60 filters). The primary catalogue contains 64 293 sources, and is estimated to be complete for point sources up to a magnitude of $r \simeq 23.6$. More details about miniJPAS can be found in Bonoli et al. (2021).

Starting from the dual-mode photometry catalogue, we make a quality cut that eliminates all objects with any of the flags that could indicate a problem with the photometry in any of the filters. This first cut lowers the number of sources down to 46 440 objects. Next, since we are not interested in extended sources (these are almost unequivocally classified as galaxies), we selected only the point-like sources from the miniJPAS full sample, by imposing the cut $\texttt{ERT} \geq 0.1$, which is a stellarity index constructed from image morphological information, with the help of Extremely Randomized Trees (Baqui et al. 2021), and which is provided in the miniJPAS catalogue. If that classification failed ($\texttt{ERT} = -99.0$), we then used the stellar-galaxy locus classification, with a cut of $\texttt{SGLC} \geq 0.1$ (López-Sanjuan et al. 2019). After these refinements, we end up with 11 419 sources that we must now classify as stars, galaxies, low-redshift ($z < 2.1$) quasars, or high-redshift ($z \geq 2.1$) quasars.[1] We then extract the fluxes and flux errors for all these objects in each filter, using the photometry for a fixed aperture of 3 arcsec and correcting for the light profile outside of that area, as detailed in Queiroz et al. (2022). We refer to this sample as the miniJPAS point-like source subsample.

The area of the miniJPAS survey was chosen to overlap with the AEGIS field (Davis et al. 2007) because in that region there is a wealth of information such as optical spectra from the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), SDSS and DEEP2/DEEP3 (Cooper et al. 2011; Newman et al. 2013), as well as X-ray data from *XMM*. However, in our applications we consider only the cross-match of miniJPAS with the SDSS Data Release (DR) 12 Superset (Pâris et al. 2017), which contains visually inspected spectra and redshifts of all BOSS quasar candidates. As a result of that cross-match, we end up with 117 quasars, 40 galaxies, and 115 stars. Fig. 1 shows the histograms of the r magnitudes of the objects in the miniJPAS point-like source subsample, as well as the objects from the SDSS cross-match sample, which is also split into the different classes. The cross-match sample constitutes a 'truth table' that we can use to check the classification derived on the basis of the miniJPAS J-spectra. Although the SDSS cross-match sample constitutes an important test set, one should bear in mind that it is not only extremely small, but it is also biased in terms of brighter sources, stellar types, redshifts, etc. The scarcity of spectroscopically confirmed objects is a problem not only for testing the methods, but mainly for training the ML methods, which require very large data sets in order to tune the weights of the network. Therefore, in order to train and to validate our classifiers with reliable statistics, we employ simulated data, the mock J-spectra, which are described in the following section.

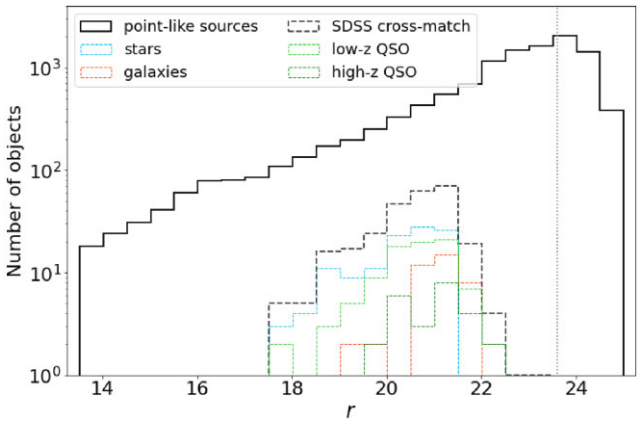

Figure 1. Histogram of the r magnitudes of the miniJPAS point-like source subsample (solid line), compared with those for the SDSS cross-match sample (dashed lines). The distribution of objects classified by SDSS as stars, galaxies, and low-z and high-z QSOs is shown in coloured dashed lines. The cross-match sample is effectively limited at $r \lesssim 22$, while the miniJPAS sample reaches up to $r \lesssim 24.0$. The vertical dotted line shows $r = 23.6$.

2.2 Mock J-spectra

ML algorithms are usually trained and validated using real-world data sets, and are subsequently applied to data that are as similar as possible to the training sets. However, when real data are not available or are too scarce, simulations can be employed to either complement existing real-world training sets or to build entire training sets (see e.g. Hoyle et al. 2015; Ramachandra et al. 2021).

Supervised learning algorithms depend on large and complete training sets with verified labels in order for the models to be properly fitted (Deng et al. 2009). In the case of J-PAS/miniJPAS data, the numbers of objects with confirmed labels are barely large enough for us to test the algorithms – never mind training them. Moreover, catalogues of astrophysical objects with confirmed labels are typically biased due to the target selection processes prior to the spectroscopic observations. They are also typically brighter, allowing for better signal-to-noise ratio (SNR) observations, and as a result may not contain a faithful representation of the variety of objects

[1]The $z = 2.1$ pivot was chosen because of the Lyman α feature. Hence, our classification provides a preliminary sample of high-redshift quasars that will be improved with appropriate redshift estimators.

expected to be found in a deeper, complete sample. Therefore, mocks are important in astronomy not only to augment the volume of the training sets, but also to fill in the sample where it lacks in diversity, in terms of magnitude ranges, types, and redshifts.

However, the construction of realistic simulated data sets is beset with substantial challenges. First, the frequencies of the objects in the training sets need to be kept under control; otherwise, we may bias the classes in the validation and test sets. Secondly, the properties of the simulated data itself must mimic, as much as possible, those of the real data sets. This means that not only the measurements, but also their uncertainties, must observe the same distributions in terms of luminosity, object class, and SNR.

In Queiroz et al. (2022), we have described in detail how we have constructed a mock catalogue of quasars, stars, and galaxies that reproduce the frequencies of those classes of objects that we expect to find in the real data sets. The first step in those simulations is a random sampling of objects drawn from given distribution functions: the quasars obey a standard LF (Palanque-Delabrouille et al. 2016), the galaxies follow a distribution based on the miniJPAS sample cross-matched with DEEP3 and SDSS DR12Q used in the quasar selection, and the stellar types and magnitudes follow the distribution expected for the specific region of the Milky Way that overlaps with the AEGIS field (Robin et al. 2003).

After specifying the types, luminosities, and redshifts of the objects in the mocks, we search for SDSS optical spectra and compute the fluxes and magnitudes in the J-PAS filters by convolving those spectra with the filters. The synthetic fluxes are similar to those measured by miniJPAS, except for the fact that their SNRs are typically much higher due to the nature of the SDSS spectroscopic observations. The next step is therefore to add noise to the synthetic fluxes in such a way that the final simulated data set has an SNR distribution that is consistent with the miniJPAS observations.

At this point, care must be taken to reproduce the actual noise properties of the underlying real data set. As shown in Queiroz et al. (2022), for some filters the noise models turned out not to be well fitted by a Gaussian, but some are better fitted by slightly different distributions. In this paper, unless noted otherwise, we train and test our ML methods with the mocks produced using the best-fitting noise models, labelled as noise 'model 11'.

Finally, the mocks also model the pattern of non-detections (NDs) from the miniJPAS point-like source subsample. In order to train the ML models, we leave the fluxes exactly as they are in the catalogues, without any special treatment of those low SNR measurements.

We used four data sets to train and validate our models. The training set is a balanced data set (Johnson & Khoshgoftaar 2019) containing equal numbers of stars, galaxies, and quasars (10^5 of each); the validation set, which we used for the ML model selection, contains 10^4 objects of each class; and the 'balanced test set' contains another 10^4 stars, galaxies, and quasars. In addition, we used an alternative test set, the '1-deg^2 test set', that contains the expected numbers of objects within 1 deg^2, down to the photometric depths of miniJPAS. Thus, this test set is not balanced. As usual in ML, both test sets remained completely blind to the training procedure.

3 MACHINE LEARNING MODELS

Dividing complex objects into classes is one of the tasks where ML has become widely used: identifying letters in written manuscripts, detecting different types of animals in images, or addressing financial risks from socio-economic data are some of the simplest examples where the applications of ML methods have shown remarkable success.

Here, we consider classification using photometric catalogues as the basic data set for classifying the objects, and we focus on the fluxes and their associated errors – i.e. we will rely on the averaged spectral features of those astrophysical sources. The set of fluxes (or, equivalently, magnitudes) in broad-band photometric surveys is typically treated as 'tabular data', since there are only a few measurements that follow a certain order, which can be thought as the central wavelengths of the filters (the photometric bands).

There are in fact some particular ML models that are considered as standard benchmarks for tabular data classification – e.g. RFs, NNs, gradient boosting, etc. – and these methods are also commonly used to separate astrophysical sources (Nakoneczny et al. 2019, 2021; Baqui et al. 2021; Nakazono et al. 2021). In the case of narrow-band surveys, however, we have a significantly higher spectral resolution compared with broad-band surveys. This means not only that there is much more data, but also that the relevant local features (e.g. breaks, emission and absorption lines) can involve complex combinations of several different points in the input data sequence.

The classification of astrophysical sources involves several additional challenges related to ML such as biased training sets, handling missing data (e.g. non-observations and NDs), noisy labels,[2] and noisy attributes. Moreover, one could also raise the issues of model interpretation and uncertainty quantification.

The problem of biased training sets arises because in astronomy the training sets are usually built based on cross-matches with spectroscopic surveys – from which we get reliable labels. Apart from the fact that spectroscopic surveys require significantly more resources compared with imaging, spectroscopic training sets may be biased over bright sources, redshift ranges, etc. This is an issue for ML since these models are unreliable on 'out of domain' samples, i.e. data that extrapolate the training set. In this work, this problem is partially alleviated with the help of the mocks, which were built not only to increase the size of the training sample, but also to be more representative of what we expect to find in the real data, in terms of brightness, redshift, and stellar types.

Queiroz et al. (2022) also avoided the problem of noisy labels as much as possible, by building the mocks only with the sources from the SDSS Superset catalogue, which should return the most reliable classification based on high-resolution spectra complemented by visual inspection.

In this work, we also draw special attention to noisy attributes (the errors in input data). Our catalogues contain, for each object, the 60 fluxes and associated uncertainties provided by the J-PAS filter set. We test several ML models to classify miniJPAS quasars, stars, and galaxies and focus on CNNs because of their flexibility as well as the ease with which we can include the information conveyed by the measurement errors while keeping the context of those errors – i.e. the fact that a given uncertainty is related to its corresponding measurement (Rodrigues et al. 2021). These uncertainties inform the significance of individual measurements – and this is equally true both for template fitting using a χ^2 as for ML methods. If the data set is very homogeneous, with nearly identical uncertainties for all data points, then of course there is no information in the errors. However, for extremely diverse data sets such as astronomical catalogues, with both bright and faint objects, and a complex distribution of SNRs as a function of magnitude, this information is critical.

We compare the CNNs with two additional ML baseline models: RFs (Breiman 2001) and LightGBM (LGBM; Ke et al. 2017), for

[2]By noisy labels we mean objects that have been assigned the wrong class, e.g. galaxies labelled as stars.

which we discard the uncertainties. Feedback from intrinsically different ML methods gives important hints on how to improve the models, on pre-processing of the input data, and on the validation of the mock data sets. In this paper, we also performed a feature importance analysis (Appendix B), which can be used to address the problem of model interpretation.

Regarding pre-processing, in order to pass the inputs to our ML models, we normalize the fluxes and flux uncertainties of any given object according to the root mean square flux for that object:

$$f_\lambda \rightarrow \frac{f_\lambda}{\sqrt{\sum_\lambda f_\lambda^2}}, \quad \sigma_\lambda \rightarrow \frac{\sigma_\lambda}{\sqrt{\sum_\lambda f_\lambda^2}}, \tag{1}$$

where the wavelength here is just a label corresponding to the central wavelength of each filter, $\lambda \in$ (uJAVA, uJPAS, . . . , J1007).

In the next subsections, we introduce the ML algorithms used in this work.

3.1 Convolutional neural networks

A NN is a type of learning algorithm where multiple activation units (neurons) are combined through layers to extract information from the data and return a prediction. The input layer receives the set of features of some instance from the data set and to each feature is assigned a weight. The activation functions encoded in the neurons from the following layer operate in the scalar product between the features and corresponding weights. This procedure is repeated recursively until the last layer, which outputs the predictions. The layers from NN structures where all neurons are fully connected are called 'dense layers', and they are designed to learn how to recognize global patterns from the input features.

CNNs work similarly, but were developed to learn how to detect local patterns using convolution kernels. For this reason, CNNs have become the benchmark for feature extraction on data sets such as images and sequential data. The architectures of CNNs are usually composed of sets of convolution and dense layers: local features are extracted from the input data with the convolution kernels, and are then combined into the dense layers to output the prediction.

In our context, CNNs can be used to search for local features in the J-spectra. A similar idea has already been used to classify astrophysical sources from narrow-band surveys in Cabayol et al. (2018), where they show that one-dimensional (1D) convolution kernels can be used to classify galaxies and stars, leading to better results when compared to usual ML algorithms, which are due to the ability of the CNNs to extract these local features. For an application in the context of high-resolution spectroscopic data, see e.g. Busca & Balland (2018), Lovell et al. (2019), and Sharma et al. (2019).

We created our own CNN architectures with the help of the `keras` framework (Chollet et al. 2015). We used the `adam` optimizer to minimize the categorical cross-entropy loss function

$$\text{cross-entropy} = -\frac{1}{N}\sum_n^N \sum_k^K y_{nk} \log p_{nk}, \tag{2}$$

where N is the number of instances, K is the number of classes, y_{nk} is the true class, and p_{nk} is the assigned probability. The convergence of the models was monitored using learning curves of the F_1 score (see Section 4.1) and the loss function on the training and validation sets. The number of epochs is constrained to the `EarlyStopping` callback: the training is interrupted when the validation loss stops improving for a number of epochs specified by the `patience`. In order to prevent the training from stagnating, we vary the learning rate using the `ReduceLROnPlateau` callback, which reduces the learning rate when the validation loss stops decreasing for a chosen number of epochs. We also use the `ModelCheckPoint` callback to save the set of weights that leads to the best F_1 macro-averaged score in the validation set. The final model corresponds to this set of weights, ensuring that the model has varied very little in the last epochs. In all intermediate layers, both convolution and dense, we use as activation the rectified linear unit (ReLU) function $f(x) = max(0, x)$ (Nair & Hinton 2010). In the last dense layer, on the other hand, we use the *softmax* activation function in order to obtain a probabilistic interpretation of the output value; i.e. the scores assigned to the four classes add up to one.

The input feature maps and architectures for each CNN version are illustrated in Fig. 2. We call a set of convolution (`Conv1D` or `Conv2D`), `BatchNormalization`, `MaxPooling`, and `Dropout` layers a 'block'.

3.1.1 CNN1

The first CNN version receives as input the set of fluxes (J-spectra) and nominal errors organized as 1D vectors in two channels (upper panel in Fig. 2). In this way, the learned features from both channels are combined in the output feature map. We also trained and tested CNN1 without the second channel, i.e. only with the fluxes, without including the uncertainties.

After the set of convolution layers processes the J-spectrum, it returns a tensor that is converted into a 1D vector in the `Flatten` layer. In addition to the J-spectrum 'tensor', we also add as input the r magnitude in the `Flatten` layer.[3] This vector then serves as input for two intermediate dense layers with 64 and 32 neurons, which are finally connected to the output layer that returns the scores assigned to each class.

3.1.2 CNN2

The strategy used to account for the uncertainties as input features in CNN2 is to treat the measurements as probability distributions with mean value equal to the flux measurements, and standard deviation equal to the corresponding nominal errors (Rodrigues et al. 2021). These distributions are then represented as 2D matrices, as illustrated in Fig. 3. This format for the input data can be particularly useful to represent errors that do not follow a simple form such as a Gaussian distribution. Furthermore, since the matrix representation is identical to an image, it is naturally suited for CNNs with 2D convolution kernels. The idea of representing fluxes and uncertainties as heatmaps has already been used in the context of astrophysical source classification (Qu et al. 2021, 2022). The bottom panel of Fig. 2 illustrates the architecture of our CNN2 method. Once again, we add the r magnitude in the `Flatten` layer feature map, and the dense layers contain 64 and 32 neurons, as in CNN1.

3.2 Decision tree-based algorithms

In the following subsections, we introduce the DT-based models used to compare with the performance of the CNNs. The details

[3]With this strategy, it is possible to include any other 'tabular' features, such as morphological parameters or time-domain data, in addition to the J-spectra.

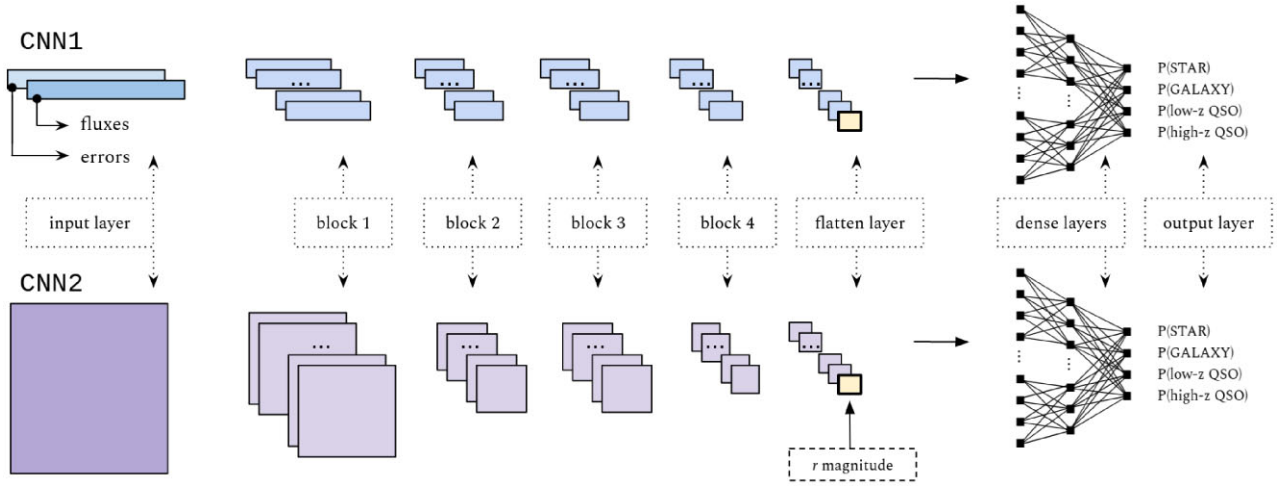

Figure 2. CNN1 (top) and CNN2 (bottom) architectures. CNN1 input data are the set of normalized fluxes and corresponding uncertainties represented as a vector with two channels. One can also train CNN1 without the errors by including only the first channel. The input data of CNN2 are the set of normalized fluxes and corresponding uncertainties represented in two dimensions (see Fig. 3). A 'block' contains a convolution, batch normalization, max pooling, and dropout layers. The yellow box in the feature maps from both flatten layers represents the *r* magnitude, which is added to the feature map after the convolution layers have processed the J-spectra.

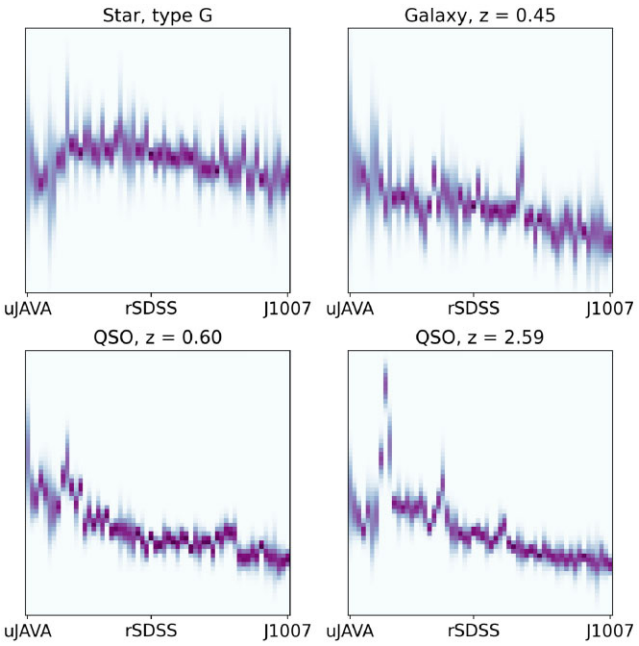

Figure 3. Diagram representing the CNN2 input data. Columns correspond to the miniJPAS filters and rows correspond to normalized fluxes. Darker pixels correspond to higher probability, i.e. denser regions of the probability distribution. The top panels show a G-type star (left) and a galaxy at $z = 0.45$ (right). The bottom panels show a low-z QSO at $z = 0.60$ (left) and a high-z QSO at $z = 2.59$ (right). Computed according to noise model 11.

about hyperparameter (HP) tuning of these models are described in Appendix D.

A DT is a structure where the algorithm makes predictions by splitting the data set based on constraints imposed in terms of the features. Each decision rule is encoded in a node of the tree. The algorithm establishes which feature will be evaluated at each node by measuring the worth of a split based on each of the features. This is quantified by the information gain, which measures the expected decrease in some impurity function. This function can be either *entropy* or the *gini impurity*. The features that lead to the highest increase in the gain are then allocated to the corresponding node.

3.2.1 *Random forests*

RFs have been widely used for many tasks related to astrophysical data, including source classification (Nakoneczny et al. 2019; Baqui et al. 2021; Nakazono et al. 2021). The method consists of combining multiple DTs to avoid overfitting and build a powerful classifier.

We implemented the RF model with the `scikit-learn` (Pedregosa et al. 2011) PYTHON package. Each tree is built with a subsample of the data, using the bootstrap aggregating (*bagging*; Breiman 1996) technique.

The number of features to consider when looking for the best split is by default set as the square root of the total number of features. The mechanism of combining independent trees using the bagging strategy makes RF robust to overfitting, and is usually not necessary to limit the growth of each individual tree.

The size of the subsample, the number of features, and the maximum depth of the trees are examples of RF HPs. The chosen values of the HPs from `scikit-learn RandomForestClassifier` are specified in Table D2.

3.2.2 *LightGBM*

Gradient boosting decision tree (GBDT) is another type of DT ensemble method that has also proved to be an excellent tool for a variety of problems, including astrophysical source classification (Nakoneczny et al. 2019). As opposed to RF, the trees are not grown independently. Instead, each tree is built to reduce the error of the previous one. This is an iterative method that uses gradient descent to minimize the loss function, which we chose to be the categorical cross-entropy – see equation (2).

We implemented GBDTs with LGBM (Ke et al. 2017). There are several well-succeeded frameworks to implement GBDTs, for example XGBoost (Chen & Guestrin 2016). LGBM was developed

to accelerate the training, but it often presents similar (or even better) performance compared with XGBoost. However, due to LGBM's leaf-wise growth scheme, it might be susceptible to overfitting, so we limit the growth rate and the maximum number of leaves of the trees (see Table D1).

4 PERFORMANCE IN THE MOCK TEST SETS

We start analysing the performance of the CNN1 (with and without the errors), CNN2, RF, and LGBM classifiers when they are applied to the mock test samples. The results when applying those methods to real data will be shown in the next section.

4.1 Evaluation metrics

In order to build a high-quality quasar catalogue, we need to find the best possible balance between *completeness* and *purity*, i.e. we want to recover the highest fraction of quasars possible, but in a such a way that our sample remains as free from contaminants as possible. With that in mind, we evaluate the performance of the classifiers by computing both purity ('precision') and completeness ('recall'):

$$\text{purity} = \frac{\text{TP}}{\text{TP+FP}}\,, \tag{3}$$

$$\text{completeness} = \frac{\text{TP}}{\text{TP+FN}}\,, \tag{4}$$

where TP, FP, and FN are true positive, false positive, and false negative, respectively. In order to find the ideal balance between completeness and purity, it is useful to define the F_1 score, which combines both scores into a single number:

$$F_1 = 2 \times \frac{\text{purity} \times \text{completeness}}{\text{purity} + \text{completeness}}\,. \tag{5}$$

All ML models employed in this work return a score associated with each class, which can be interpreted as a proxy for the probability that an object belongs to that class. The scores of all classes add up to 1. We have the freedom to choose different thresholds for these classification scores (the 'probabilities') in order to improve the final classification. Depending on that choice, one may obtain a more complete or more pure sample; i.e. the F_1 score depends on the threshold. By default, the chosen class k corresponds to the class with the highest score, according to the argmax rule:

$$y_i = \underset{k}{\text{argmax}}\; f^k(\mathbf{x}_i)\,, \tag{6}$$

where $\mathbf{x}_i$ and y_i are the input data and predicted class of instance i, respectively, and f is some function that assigns probabilities to each class k. This means that, when we apply some trained ML model to classify an instance i, it returns a probability associated with each class k and the final class corresponds to k with highest score.

Another useful metric is the *receiver operating characteristic* (ROC) curve, because it shows the quality of a classifier before choosing a specific threshold by computing the TP rate *versus* FP rate. Moreover, the *area under the ROC curve* (ROC-AUC) is a useful summary statistic of the ROC curve to measure the quality of a classifier. Since we are working with multiple classes, we computed the one-*versus*-all ROC-AUC score.

Finally, in order to compare the overall performance of a classifier by considering the performance over all classes, it is useful to compute the unweighted, or *macro*-averaged score, defined as

$$\bar{S} = \frac{1}{K}\sum_{k}^{K} S_k\,, \tag{7}$$

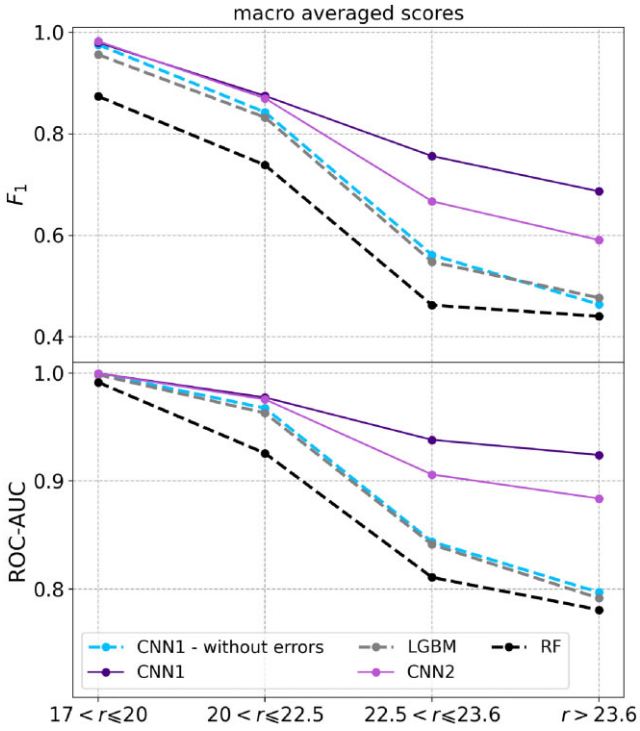

Figure 4. Performance of the ML models when applied to the balanced test set, in terms of the macro-averaged F_1 score (top), and the ROC-AUC (bottom), for the different *r-magnitude* bins.

where S is some score or metric, k labels the individual classes, and K is the total number of classes. This metric does not take into account the imbalance of classes, and thus avoids biasing the analysis over more frequent types.

Fig. 4 shows the macro-averaged F_1 and ROC-AUC scores obtained with the classifiers in the balanced test set, in multiple intervals of r magnitude – see Appendix A for the complete confusion matrices. It is striking how much the performance of the CNN1 classifier improves when the information about the errors is included, in particular for the fainter objects where SNR is even more crucial. That performance is similar using CNN2, which employs an entirely different architecture for the input data but that, like CNN1, also uses the convolutional layers to incorporate the errors in the context of their corresponding measurements. The fact that both DT-based methods (specially LGBM), which do not take the errors into account, attain a performance that is similar to CNN1 without errors indicates that the reason for the improvement in the classification seen in the two CNN methods with errors is in fact due to the additional information contained in the uncertainties. We also see from Fig. 4 that the performances of all the classifiers degrade as the samples become fainter, which is expected since those objects are increasingly noisier and therefore harder to identify. We used the same magnitude bins as Martinez-Solaeche et al. (in preparation), which verified a similar behaviour. Due to its superior performance, we will focus on the results obtained with CNN1 for the remainder of this section, unless noted otherwise – but we emphasize that the results outlined here are qualitatively consistent between all classifiers.

4.2 Results

In this section, we present the results when we apply the CNN1 method to the two mock test sets: the balanced test set (with 10^4

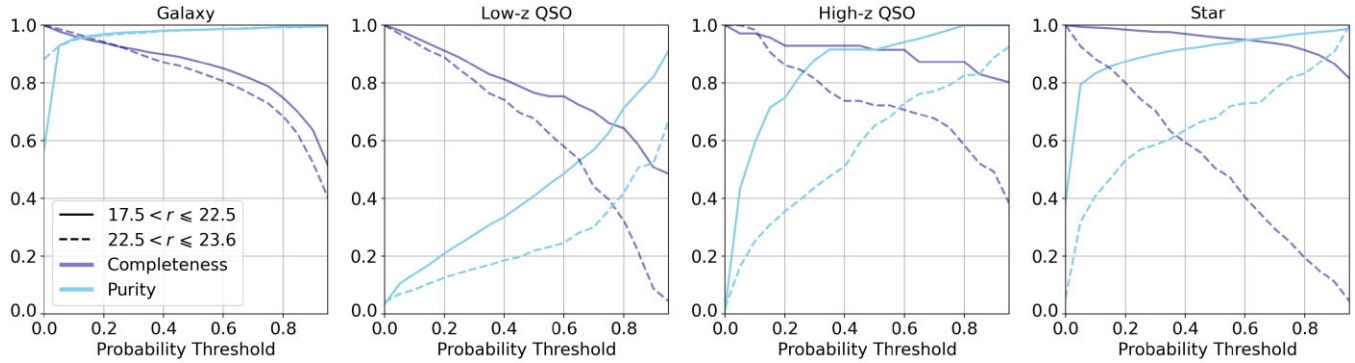

Figure 5. Completeness and purity of the CNN1 method for the mock 1-deg^2 test, as a function of the probability threshold, for each class. Brighter (fainter) objects are shown in solid (dashed) lines.

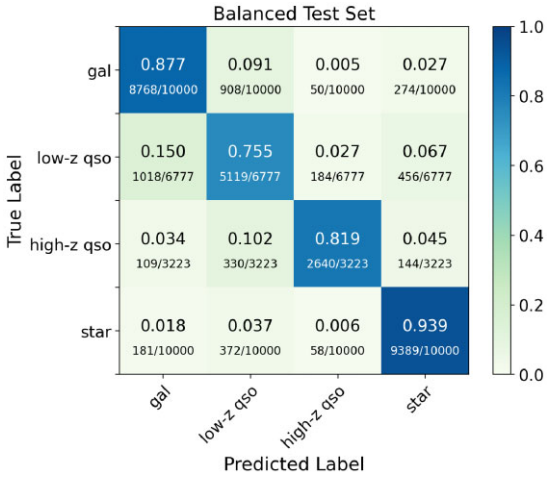

Figure 6. Confusion matrix computed with CNN1 for the mock balanced test.

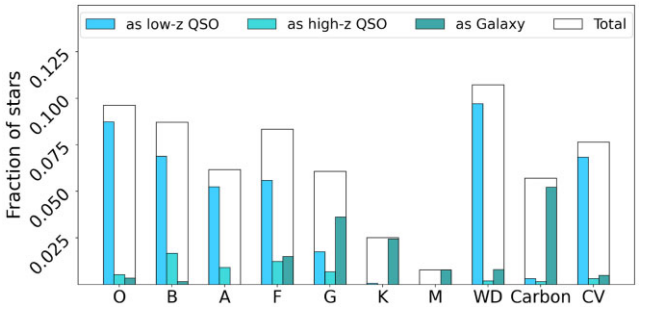

Figure 7. Fraction of stellar types that were incorrectly classified by CNN1 in the balanced test set.

objects in each class), and the 1-deg^2 test set, which is perhaps a more realistic representation of the miniJPAS sample. For much of this analysis, it is more revealing to evaluate the predictions in terms of the balanced test set, just because it is the largest one and we can thus work with more reliable statistics. However, evaluating the proper choice of threshold using the balanced test set can be misleading, since we want to estimate the purity and completeness in a realistic scenario, with the expected fraction of objects of each class. Therefore, we start by showing, in Fig. 5, the purity and completeness as a function of the probability threshold in the 1-deg^2 test set. We split the sample in two bins of r magnitude, $17.5 < r \leq 22.5$ and $22.5 < r \leq 23.6$, because the optimal choice for the cut might depend on how bright the object is: fainter objects are much noisier, so we expect a classifier to be less confident in this regime. Based on this analysis, we define the '1-deg^2 threshold criteria' to select candidates in the miniJPAS catalogue – one value for bright and one for faint sources, according to the magnitude bins shown in Fig. 5. It corresponds to the value of threshold that leads to highest F_1 score in the 1-deg^2 test sample, and it must be at least equal to 0.5 to ensure that the probability associated with some class is greater than the sum of the others. Notice, however, that not all objects are assigned a class according to this criterion.

For the remaining of the analysis in this section, we work with the balanced test set, for which the best choice of threshold is in good approximation of the 'argmax' criterion, defined in equation (6).

Fig. 6 shows the confusion matrix computed with CNN1. We are able to distinguish between low-z and high-z QSOs satisfactorily, and the main source of confusion is between low-z QSOs and galaxies, which is in agreement with the results of Martinez-Solaeche et al. (in preparation). In Appendix A, we show the confusion matrices split into the same r-magnitude bins as in Fig. 4, for all ML methods.

As a complementary analysis, we trained CNN1 in a binary classification scheme, by labelling low-z and high-z QSOs as one single class, and stars together with galaxies as another class. The results of that analysis are nearly identical with the numbers shown in Fig. 6 when we combine the low-z and high-z QSOs in one class, and the stars and galaxies in the other class.

Fig. 7 shows the fractions of stellar types that were incorrectly classified in the mock balanced test set. We show this result in terms of fractions to avoid biasing the analysis over more frequent stellar types; i.e. we take the ratio between the number of incorrectly classified stars of a given stellar type and the total number of stars of the corresponding type. White dwarfs (WDs) and O-type stars show the highest fraction of incorrect classifications, which are often classified as low-z QSOs. The steep blue continuum of the WD spectra can be easily mistaken for the blue and featureless continuum of the QSO spectra at low redshifts (Richards et al. 2002; Myers et al. 2015). The A, F, and M stellar types also have colours and/or spectral features similar to those from QSOs. However, we did not detect significant confusion of these types as compared to the others. In particular, M stars are classified as galaxies. Stars of types O, B, and F are featureless, which might explain the significant confusion with low-z QSOs. In fact, low-z QSOs are the class of objects that are most affected by contamination from stars, which is a well-known problem for broad-band classification in the optical range (Richards et al. 2009), which still persists even with narrow-band data. For the

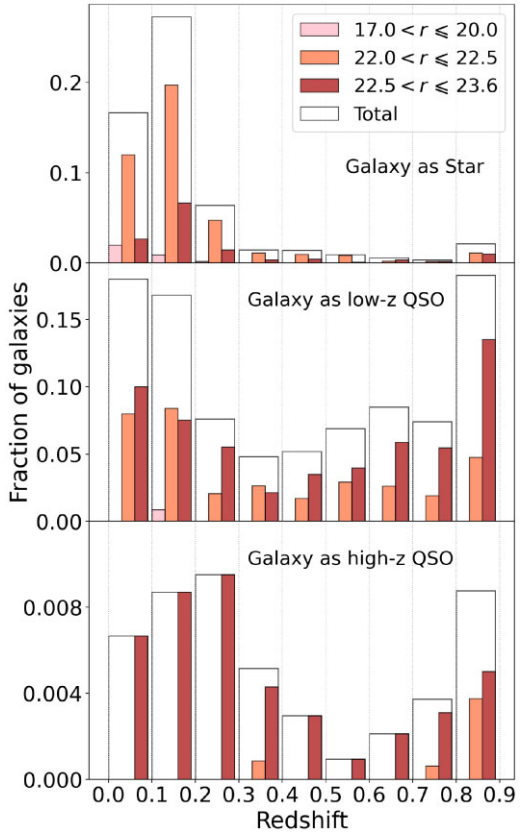

Figure 8. Redshift of the galaxies that were incorrectly classified as stars (top), low-z QSOs (middle), and high-z QSOs (bottom) by CNN1 in the balanced test set.

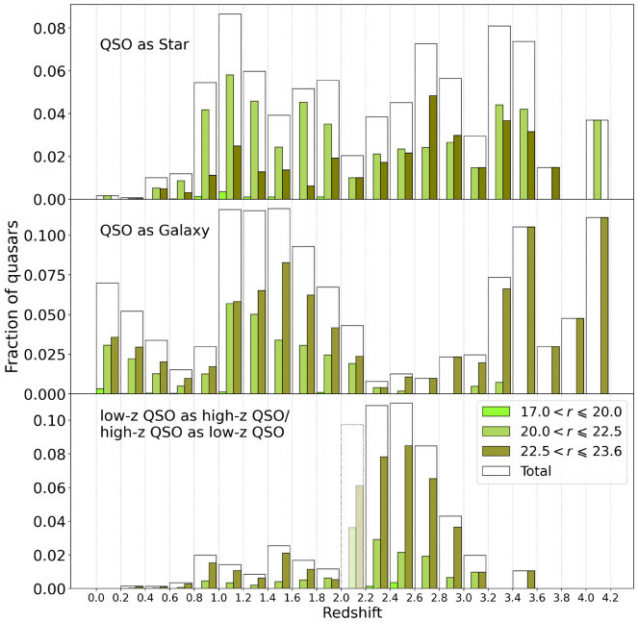

Figure 9. Redshifts of the quasars from the balanced test set that were incorrectly classified by CNN1. The top and middle panels show the quasars that were classified as stars and galaxies, respectively, and the bottom panel shows the quasars that were classified as quasars in the wrong redshift interval. The bars of the histograms cover a redshift range of $\Delta z = 0.2$ and were split according to the r magnitude. The bin containing the pivot value $z = 2.1$ that separates low-z QSOs from high-z QSOs is shown with different transparency.

few stars that end up classified as high-z QSOs, most are of types B, A, and F as well as some G stars, although some stellar types outside the main sequence (WD, Carbon, and CV) can also contaminate that sample. In general, redder stars (G, K, and M) are more often confused with galaxies, while bluer stars (O, B, A, and F) are more often confused with QSOs, and Carbon stars are the type most often confused with galaxies.

In Fig. 8, we show the redshifts of the galaxies for each magnitude bin that are confused with stars (upper), low-z QSOs (middle), and high-z QSOs (lower panel). Since galaxies and quasars are typically not as bright as Milky Way stars, we split the samples into bins of magnitude in the r band in order to check the dependence of the classification on the brightness of these sources. Once again, we compute the fraction of galaxies, now in each redshift bin. There is very little leaking of galaxies to high-z quasars and it is dominated by fainter objects. From the confusion matrix in Fig. 6, we see that there are only 50 galaxies classified as high-z QSOs. The galaxies that are classified as low-z QSOs (and also those classified as stars) have typically lower redshifts, although we see similarly high confusion of galaxies within $0.8 < z < 0.9$ and low-z QSOs.

In Fig. 9, we show the redshifts of the QSOs that were incorrectly classified. Similarly to what happened for the incorrectly classified galaxies, the confusion as a function of redshift is partially related to the fainter magnitudes of these objects. The top and middle panels show the QSOs that were classified as stars and galaxies, respectively. The bottom panel shows the low-z/high-z QSOs that were classified as high-z/low-z QSOs.

QSOs classified as galaxies are typically fainter, while those classified as stars are similarly distributed in the bright and faint ends. This reflects the fact that, on the one hand, faint objects are harder to classify, and we thus expect a higher mixing at this regime. On the other hand, stars are most abundant in the bright end, and therefore are expected to be the most frequent contaminants.

The quasar population within $z \in [0.6, 2.0]$ has a scarcity of emission lines, which could explain the confusion with stars and galaxies. For $z < 0.6$, we see a higher contamination of the galaxy sample that does not happen for stars. This might be due to the fact that the strongest QSO emission lines in this redshift range, such as Hα, are also commonly found in galaxies.

From Fig. 6, we see that 10 per cent of the high-z QSOs are classified as low-z QSOs, while only 2.7 per cent of low-z QSOs are classified as high-z QSOs. The bottom panel of Fig. 9 shows the redshift of those objects. The redshift cut at $z = 2.1$, which differentiates the two populations of QSOs, blurs the distinction between the two classes in the $z \in [2.0, 2.2]$ range (the bin with higher transparency). The number of incorrectly high-z QSOs that are classified as low-z QSOs starts dropping for $z > 2.2$, faster for bright objects and slower for faint objects, indicating some level of confusion between the Lyα break and other spectral features of the low-z quasars.

4.3 Robustness tests

We tested the robustness of the ML classification by modifying the composition of the training set in different ways. After training the ML model with the modified samples, we evaluated the performance in the balanced test set, which we kept unchanged.

The first test consists in using only 50 per cent of the original numbers of stars and keeping the number of galaxies and QSOs

from the original training set (*half stars* test). Since the stars were removed randomly, we expect the completeness of the star sample and, as a consequence, the purity of the other classes, to decrease to some extent. The second exercise is the *double stars* test, which is complementary to the previous one: we exclude 50 per cent of the galaxies and 50 per cent of the quasars (once again, randomly), while maintaining all the stars of the original training set.

The idea is that, by changing the proportion of classes in the training set, we expect the models to show a drop in their performances, in particular for the less represented types. If the classification is very sensitive to small changes in the exact mixes of populations in the training set, then the model is not robust. If, on the other hand, the performance of the classifier drops by only a small amount after a significant change to the training set, then the ML model has converged to a nearly stationary regime.

Fig. 10 shows the scores (completeness and purity) as a function of the probability threshold for the different training sets. As expected, the completeness of stars drops in the *half stars* test. Although the purity of stars increases, it does not compensate the loss in the completeness, which also translates into a lower purity of galaxies and quasars (especially at high redshift). The same reasoning works for the *double stars* test: the completeness of quasars and galaxies drops by a small amount, but there is no significant gain in the purity because of the mixing between these two classes.

These results reflect a well-known feature of ML techniques, which are unable to reliably identify objects that are poorly represented in the training set. Nevertheless, we verified that the performance of our classifiers is relatively insensitive to significant changes in the training sample, which indicates that our ML models are robust in that sense.

5 MINIJPAS POINT-LIKE SOURCE CLASSIFICATION

In this section, we discuss the predictions in the miniJPAS point-like source subsample, which contains 11 419 objects. Considering only the magnitude range of $17.5 \leq r < 23.6$, we are left with 7468 sources. We have spectroscopic confirmation for some of the objects in this data set, obtained by cross-matching the miniJPAS catalogue with the SDSS DR12 Superset (see Section 2.1). The confusion matrix obtained for that sample is shown in Fig. 11. The typical magnitude range covered by this sample is $17.5 \leq r \leq 22.5$ (see Fig. 1). Therefore, in order to see the degradation of the results on real data relative to the mocks, one should compare Fig. 11 with the first two bins from Fig. A1. The completeness of all classes is higher than 0.8, which is a good indication that the models trained with the mocks translate fairly well to real data predictions. In particular, we see, once again, that the main source of confusion is between low-z QSOs and galaxies. Regarding high-z QSOs, of the 30 objects of the cross-match sample, 3 were incorrectly classified as galaxies, 2 as low-z QSOs, and only 1 as a star.

Fig. 12 shows the number of objects found in the point-like source catalogue within $r \in [17.5, 23.6]$, as a function of the ML score ('probability') threshold. Coloured lines show the models that include the uncertainties (CNN1 and CNN2) and, for comparison, the grey lines show LGBM and CNN1 without errors. The choice of threshold for CNN1 can be guided by Fig. 5. Once again, we split the sample into brighter and fainter objects ($17.5 < r \leq 22.5$ and $22.5 < r \leq 23.6$, respectively) in order to evaluate how many sources of each class are found in these two regimes, and to evaluate how confident the models are when facing brighter and fainter objects. The dotted curves show a more dramatic decrease for higher probabilities, which means that the models are less confident when presented with fainter objects. According to the ML classifiers, stars (galaxies) are the most abundant objects in the bright (faint) end. The numbers predicted by the classifiers are very similar for bright objects. CNN1 and CNN2, however, find a significantly higher number of faint high-z QSOs as compared to LGBM and CNN1 without errors.

Fig. 13 shows the number of objects classified by CNN1 with and without errors, along with the number predicted by the corresponding LFs (see section 2.2 and also Queiroz et al. 2022). We show both the classification obtained using the argmax rule, equation (6), which assigns a class to all sources in the catalogue, and the classification using the 1-deg^2 threshold criteria and a very restrict threshold of 0.9. Adding up all the objects predicted by the LFs results in ∼4000 objects with $r \in [17.5, 23.6]$. However, the number of instances from the miniJPAS catalogue within that interval is 7468. Therefore, we should not expect the numbers to agree perfectly with the LFs even when applying the argmax threshold. On the other hand, the total number of objects using the threshold of 0.9 in CNN1 is 4420.

6 CONCLUSION

In this work, which is part of the effort to identify quasars in the miniJPAS survey, we applied several ML models (CNN1, CNN2, LGBM, and RF) to classify miniJPAS point-like sources as stars, galaxies, and low-z ($z < 2.1$) and high-z ($z \geq 2.1$) quasars, employing only photometry-based pseudo-spectra. In order to train and validate the models, we used mock data catalogues developed by Queiroz et al. (2022). The final miniJPAS quasar catalogue will be produced by combining the predictions from several classifiers (Pérez-Ràfols et al., in preparation), among them the ML models presented in this work, as well as those presented in Martínez-Solaeche et al. (in preparation) and Pérez-Ràfols et al. (in preparation).

In this paper, we have constructed and tested five different ML models designed to be applied to miniJPAS data. We have focused on CNNs because of their potential to extract local features from the input data (the pseudo-spectra), and their ability to incorporate the information about errors in the data (Rodrigues et al. 2021). We also applied well-established DT-based models as a baseline, and in order to complement the CNN approach. We tested the robustness of the training sets by varying the populations of stars and retraining the models on modified samples, finding very small variations in the purity and completeness when training with these different data sets and applying to a fixed test set. We have also checked, using permutation feature importance, that bluer filters are particularly relevant to correctly classifying high-z QSOs (see Appendix B).

We evaluated the performance of the classifiers in terms of the purity and completeness of the predicted samples, and analysed the confusion between the four classes. We also investigated in more detail the sources of misclassification in terms of their luminosities, stellar types, and redshifts. We verified that, as a general rule, the main source of confusion is between galaxies and low-z QSOs in the faint end. Stars are more often confused with low-z QSOs as well, specially bluer types (O, B, A, and F), cataclysmic variables, and WDs. The redshift range of QSOs that were most often classified as galaxies is $z \in [1.0, 1.6]$. The performances of the classifiers decrease as the objects become fainter and noisier. We verified that the predictions with a mock test set are indeed consistent with our previous knowledge about quasars, stars, and galaxy features, which reinforces the quality of the mock data and also of the ML models developed in this work.

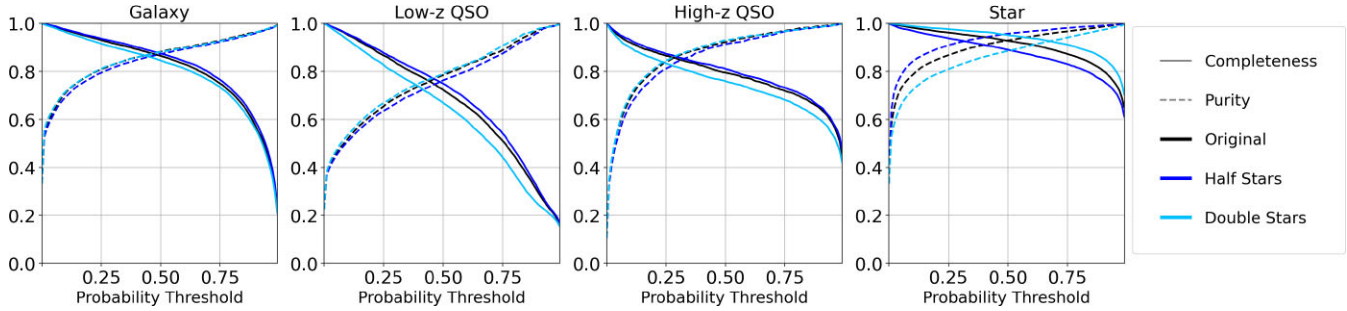

Figure 10. Robustness tests performed with CNN1 in the balanced test set. Completeness (solid lines) and purity (dashed lines) as a function of the probability threshold for each class. Different colours represent different training sets.

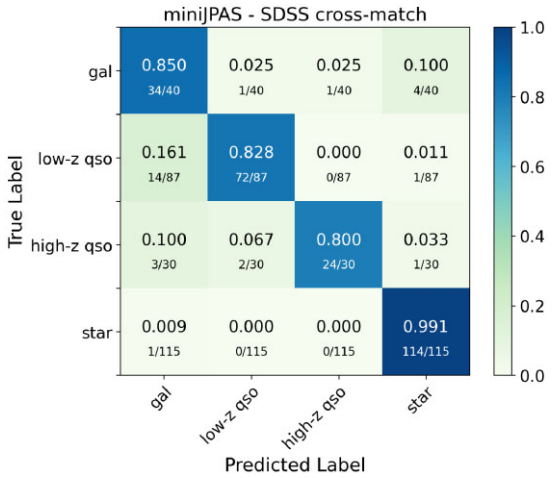

Figure 11. Confusion matrix obtained with the method CNN1 for the cross-match of the miniJPAS point sources with the SDSS DR12 Superset.

After validation, the ML models were finally applied to the miniJPAS data. For the few objects with spectroscopic confirmation of their classes, we obtained results consistent with the mock test sets (QSO completeness ∼0.8 and purity ∼0.95). Of the 7468 point-like sources in miniJPAS that lie in the magnitude range $17.5 < r \leq 23.6$, we found 2309 stars, 3827 galaxies, 118 low-z QSOs, and 547 high-z QSOs with CNN1 (noise model 11) and the 1-deg^2 threshold criteria – 667 objects did not have a type assigned with sufficient confidence to pass the thresholds specified in Section 4.2.

When applying proper choices of probability thresholds to select the quasar candidates, the models underestimate the number of low-z QSOs and overestimate the number of high-z QSOs, specially in the very faint end, as compared to the LF from Palanque-Delabrouille et al. (2016). Taken at face value, our results seem more consistent with the LF from Croom et al. (2009), which expects a higher number of faint high-redshift QSOs as compared to Palanque-Delabrouille et al. (2016).

This paper is another milestone in the J-PAS effort to map quasars at all redshifts with a minimal selection bias. These quasars will be useful for a variety of applications: first, to study large-scale structure at high and intermediate redshifts, using both the QSOs themselves as tracers (Ata et al. 2018), the Lyα forest from their lines of sight (Bautista et al. 2017), which will be measured by the WEAVE instrument (Pieri et al. 2016), and their cross-correlations (du Mas des Bourboux et al. 2017); secondly, to determine with higher accuracy both the quasar LF (Croom et al. 2009; Palanque-Delabrouille et al. 2016) and the black hole mass function (Chaves-Montero et al. 2022), revealing the history of formation of those objects; and finally, in the long run J-PAS should also be able to make a census of the QSOs including different subtypes that may be less represented in spectroscopic surveys due to the traditional targeting strategies.

ACKNOWLEDGEMENTS

This paper has gone through internal review by the J-PAS collaboration. NR acknowledges financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001. RA was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). CQ acknowledges financial support from FAPESP (grants 2015/11442-0 and 2019/06766-1) and CAPES – Finance Code 001. IPR was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowskja-Curie grant agreement number 754510. MPP and SSM were supported by the Programme National de Cosmologie et Galaxies (PNCG) of CNRS/INSU with INP and IN2P3, co-funded by CEA and CNES, the A∗MIDEX project (ANR-11-IDEX-0001-02) funded by the 'Investissements d'Avenir' French Government program, managed by the French National Research Agency (ANR), and by ANR under contract ANR-14-ACHN-0021. GMS, RMGD, and LADG acknowledge 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 the project PID2019-109067-GB100. JCM and SB acknowledge financial support from Spanish Ministry of Science, Innovation, and Universities through the project PGC2018-097585-B-C22. AFS acknowledges support from the Spanish Ministerio de Ciencia e Innovación through project PID2019-109592GB-I00 and the Generalitat Valenciana project PROMETEO/2020/085. RAD acknowledges partial support from CNPq grant 308105/2018-4. AE acknowledges the financial support from the Spanish Ministry of Science and Innovation and the European Union – NextGenerationEU through the Recovery and Resilience Facility project ICTS-MRR-2021-03-CEFCA. LSJ acknowledges support from CNPq (304819/2017-4) and FAPESP (2019/10923-5). JV acknowledges the technical members of the UPAD for their invaluable work: Juan Castillo, Tamara Civera, Javier Hernández, Ángel López, Alberto Moreno, and David Muniesa.

This study is based on observations made with the JST250 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

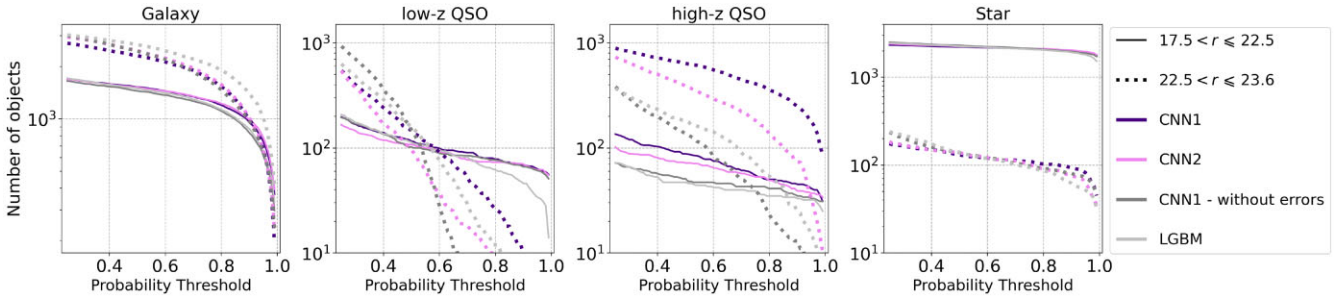

Figure 12. Number of objects predicted by different classifiers as a function of the probability threshold. We compare models that do (coloured lines) and do not (grey lines) include the uncertainties: CNN1 (purple), CNN2 (pink), CNN1 without errors (dark grey), and LGBM (light grey). Solid and dotted lines represent objects in different *r* bins.

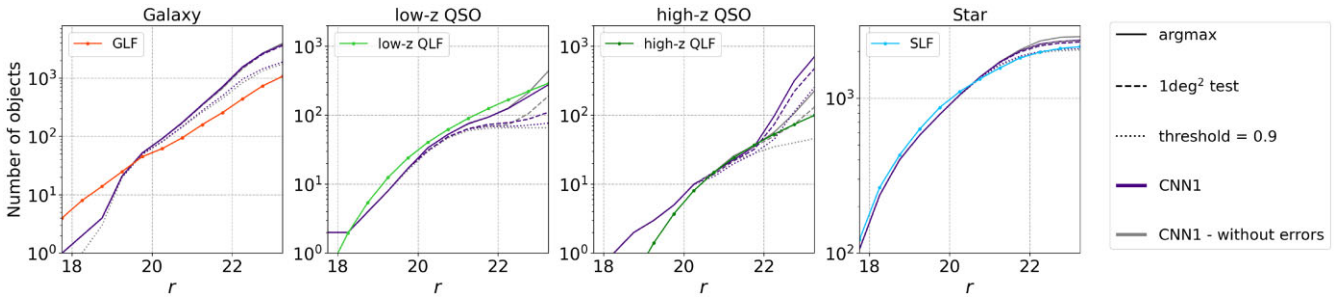

Figure 13. Number of objects predicted by CNN1 with (purple) and without (grey) errors as a function of the *r* magnitude. We compare the obtained numbers when imposing different probability threshold criteria: argmax (solid lines), 1 deg^2 (dashed lines), and a very strict choice of threshold = 0.9 (dotted lines). The LFs of each type are shown as coloured solid-dotted lines for comparison.

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 Aragonese Government through the Research Groups E96, E103, E16_17R, and E16_20R; 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 and FCDD13-4E-2685). Funding for the J-PAS Project has also been provided by the Brazilian agencies FINEP, FAPESP, and FAPERJ and by the National Observatory of Brazil, with additional funding provided by the Tartu Observatory and by the J-PAS Chinese Astronomical Consortium.

Funding for the SDSS-III/IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-III/IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

DATA AVAILABILITY

The final quasar catalogue will be generated with the combined code, described in the final article of the series (Pérez-Ràfols et al., in preparation).

REFERENCES

Alam S. et al., 2021, Phys. Rev. D, 103, 083533
Ata M. et al., 2018, MNRAS, 473, 4773
Baqui P. O. et al., 2021, A&A, 645, A87
Bautista J. E. et al., 2017, A&A, 603, A12
Benitez N. et al., 2014, preprint (arXiv:1403.5237)
Blake C. et al., 2012, MNRAS, 425, 405
Bonoli S. et al., 2021, A&A, 653, A31
Breiman L., 1996, Mach. Learn., 24, 123
Breiman L., 2001, Mach. Learn., 45, 5
Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth International Group, Belmont, CA
Burke C. J., Aleo P. D., Chen Y.-C., Liu X., Peterson J. R., Sembroski G. H., Lin J. Y.-Y., 2019, MNRAS, 490, 3952
Busca N., Balland C., 2018, preprint (arXiv:1808.09955)
Cabayol L. et al., 2018, MNRAS, 483, 529

Chaves-Montero J. et al., 2017, MNRAS, 472, 2085
Chaves-Montero J. et al., 2022, A&A, 660, A95
Chen T., Guestrin C., 2016, in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Min. KDD'16. ACM, New York, NY, p. 785
Chollet F. et al., 2015, Keras, available at https://github.com/fchollet/keras
Cole S. et al., 2005, MNRAS, 362, 505
Cooper M. C. et al., 2011, The Astrophysical Journal Supplement Series, 14
Croom S. M. et al., 2009, MNRAS, 399, 1755
Dalton G., 2016, in Skillen I., Balcells M., Trager S., eds, ASP Conf. Ser. Vol. 507, Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields. Astron. Soc. Pac., San Francisco, p. 97
Davis M. et al., 2007, ApJ, 660, L1
Dawson K. S. et al., 2013, The Astronomical Journal, 10
Dawson K. S. et al., 2016, AJ, 151, 44
Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in 2009 IEEE Conf. Comput. Vis. Pattern Recognit. IEEE Computer Society, Los Alamitos, CA, USA, p. 248
DES Collaboration et al., 2021, preprint (arXiv:2105.13549)
du Mas des Bourboux H. et al., 2017, A&A, 608, A130
Dwelly T. et al., 2017, MNRAS, 469, 1065
Hikage C. et al., 2019, PASJ, 71, 43
Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015, MNRAS, 450, 305
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jansen F. et al., 2001, A&A, 365, L1
Johnson J. M., Khoshgoftaar T. M., 2019, J. Big Data, 6, 27
Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., 2017, Adv. Neural Inf. Process. Syst., 30, 3146
LeCun Y., Boser B. E., Denker J. S., Henderson D., Howard R. E., Hubbard W. E., Jackel L. D., 1989, Neural Comput., 1, 541
López-Sanjuan C. et al., 2019, A&A, 631, A119
Lovell C. C., Acquaviva V., Thomas P. A., Iyer K. G., Gawiser E., Wilkins S. M., 2019, MNRAS, 490, 5503
Marín-Franch A. et al., 2012, in Navarro R., Cunningham C. R., Prieto E., eds, Proc. SPIE Conf. Ser. Vol. 8450, Modern Technologies in Space- and Ground-Based Telescopes and Instrumentation II. SPIE, Bellingham, p. 84503S
Martí P., Miquel R., Castander F. J., Gaztañaga E., Eriksen M., Sánchez C., 2014, MNRAS, 442, 92
Morganson E. et al., 2015, ApJ, 806, 244
Myers A. D. et al., 2015, ApJS, 221, 27
Nair V., Hinton G. E., 2010, in Proc. 27th Int. Conf. Mach. Learn. ICML'10. Omnipress, Madison, WI, p. 807
Nakazono L. et al., 2021, MNRAS, 507, 5847
Nakoneczny S. J. et al., 2021, A&A, 649, A81
Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, A&A, 624, A13
Newman J. A. et al., 2013, The Astrophysical Journal Supplement Series, 5
Palanque-Delabrouille N. et al., 2016, A&A, 587, A41
Pâris I. et al., 2017, A&A, 597, A79
Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pérez-Ràfols I., Pieri M. M., Blomqvist M., Morrison S., Som D., 2020, MNRAS, 496, 4931
Pieri M. M. et al., 2016, in Reylé C., Richard J., Cambrésy L., Deleuil M., Pécontal E., Tresse L., Vauglin I., eds, SF2A-2016: Proc. Annu. Meeting French Soc. Astron. Astrophys., WEAVE-QSO: A Massive Intergalactic Medium Survey for the William Herschel Telescope, p. 259
Qu H., Sako M., Möller A., Doux C., 2021, AJ, 162, 67
Qu H., Sako M., Moller A., Doux C., 2022, preprint(arXiv:2207.09440)
Queiroz C. et al., 2022, preprint (arXiv:2202.00103)
Ramachandra N., Chaves-Montero J., Alarcon A., Fadikar A., Habib S., Heitmann K., 2021, MNRAS, 515, 1927
Reis I., Baron D., Shahaf S., 2018, AJ, 157, 16
Richards G. T. et al., 2002, AJ, 123, 2945
Richards G. T. et al., 2009, ApJS, 180, 67
Robin A. C., Reylé C., Derrière S., Picaud S., 2003, A&A, 409, 523
Rodrigues N. V. N., Abramo L. R., Hirata N. S., 2021, preprint (arXiv:2108.04742)
Sharma K., Kembhavi A., Kembhavi A., Sivarani T., Abraham S., Vaghmare K., 2019, MNRAS, 491, 2280
Sharma K., Kembhavi A., Kembhavi A., Sivarani T., Abraham S., Vaghmare K., 2020, MNRAS, 491, 2280
Shy S., Tak H., Feigelson E. D., Timlin J. D., Babu G. J., 2022, AJ, 164, 6
Simonyan K., Zisserman A., 2014, preprint (arXiv:1409.1556)
Takada M. et al., 2014, PASJ, 66, R1
Villacampa-Calvo C., Zaldivar B., Garrido-Merchán E. C., Hernández-Lobato D., 2021, J. Mach. Learn. Res., 22, 1
Wolf C., Meisenheimer K., Rix H. W., Borch A., Dye S., Kleinheinrich M., 2003, A&A, 401, 73
Wright E. L. et al., 2010, AJ, 140, 1868

APPENDIX A: ADDITIONAL RESULTS

Fig. A1 shows the confusion matrices in bins of *r* magnitude. These results are summarized in the plots from Fig. 4.

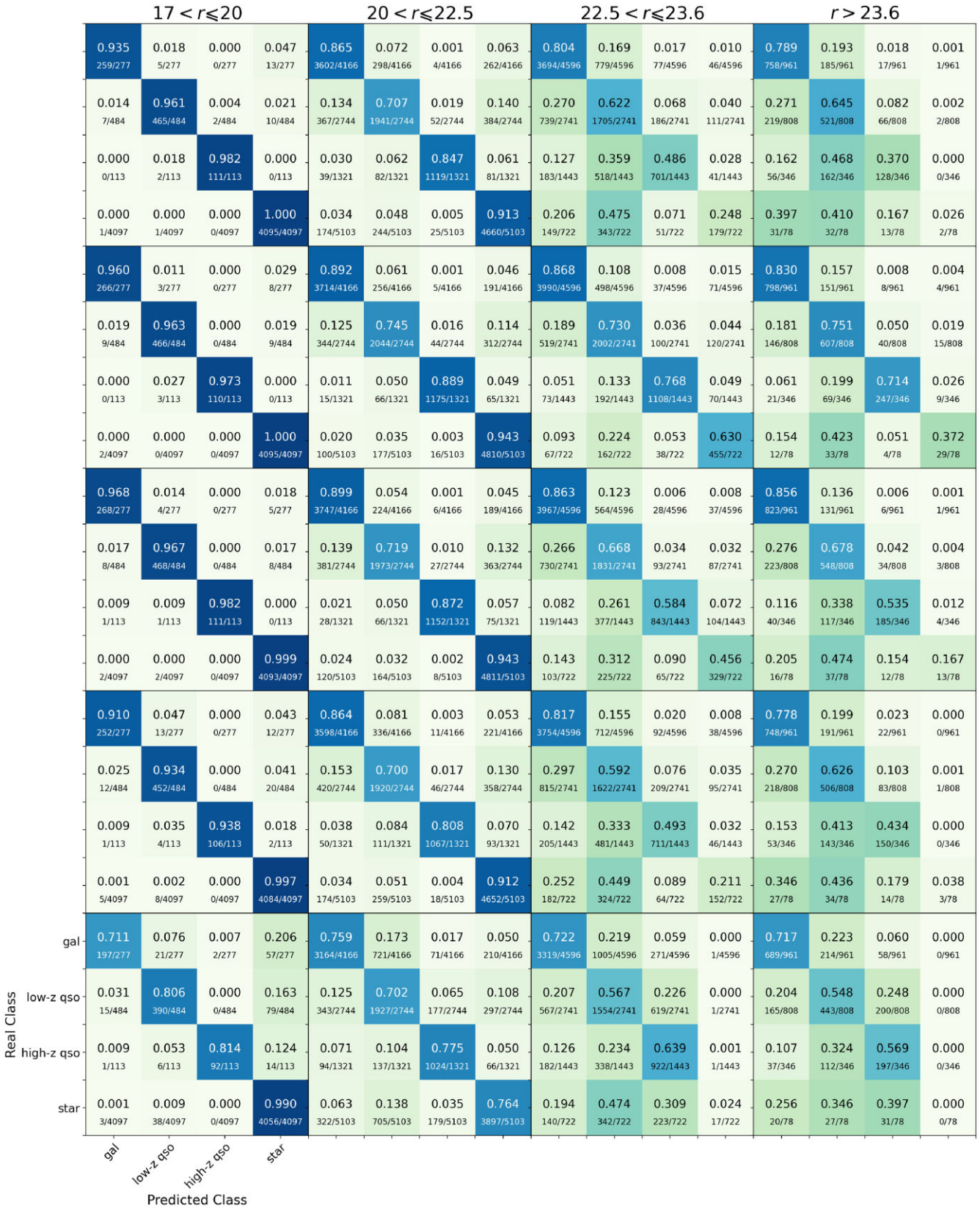

Figure A1. Confusion matrices of the classifiers in different *r*-band magnitude bins. From top to bottom: CNN1 without errors, CNN1, CNN2, LGBM, and RF.

APPENDIX B: FEATURE IMPORTANCE

We performed a permutation feature importance analysis in the balanced test set to explore which features are more relevant for the models to make the predictions. We implemented this with the `eli5` package. The procedure is the following: we exclude one filter at time and evaluate how the F_1 score of each class decreases with this missing filter. By 'exclusion' of the filter we mean that the value of the filter becomes a random number, computed by combining the values of the features. Missing filters that lead to higher decrease in the performance are more important.

Fig. B1 shows the result of the permutation feature importance with LGBM in the balanced test set. We evaluate how much the F_1 score decreases as we remove each of the features. We see that the exclusion of redder filters leads to a higher decrease in the F_1 score of low-z QSOs, while for high-z QSOs the bluer filters are more important. The Ly α and C IV emission lines are important features that characterize high-z QSOs. In the redshift range of $2.1 \leq z \leq 4.0$, the Ly α line falls within $3780\,\text{Å} < \lambda < 6080\,\text{Å}$ and C IV falls within $4800\,\text{Å} < \lambda < 7745\,\text{Å}$, which could explain the importance of the filters that cover these wavelengths.

We retrained LGBM excluding the 10 least important filters according to Fig. B1 for each of the four classes. The results remained very similar, indicating that, although their contribution to the overall classification performance seems small, there is no clear advantage in excluding those features.

APPENDIX C: CNN SETTINGS

In this section, we describe the construction of CNN2 input data matrices (illustrated in Fig. 3). The parameters of the matrices are the number of columns, number of rows, and the values to set the upper and lower boundaries (`n_cols`, `n_rows`, `up_bound`, `low_bound` – see Rodrigues et al. 2021). A matrix is created by getting the mean value of the normalized fluxes (see equation 1) of the object and by establishing the upper and lower values (the boundaries of the matrix) with `low_bound` and `up_bound`. In other words, if an object has the mean value equal to $\bar{m}$, the matrix will cover the range of $[\bar{m} - \texttt{low\_bound}, \bar{m} + \texttt{up\_bound}]$. The number of columns can be simply set as the number of attributes $\texttt{n\_cols} = 60$, since in our problem there is no uncertainty between the filters; i.e. a measurement certainly belongs to the given filter. The other parameters must be chosen more carefully to ensure that the matrix covers the complete J-spectra ranges and to ensure that the resolution of the pixels is large enough to properly resemble the probability distribution. Each filter has a specific probability distribution defined according to noise model 11. We set $\texttt{n\_rows} = 90$, $\texttt{up\_bound} = 0.6$, and $\texttt{low\_bound} = 0.3$. The resolution of the pixels is given by $(\texttt{up\_bound} + \texttt{low\_bound})/\texttt{n\_rows} \approx 0.01$, which means that the probability distribution of the normalized fluxes is binned with intervals of ≈0.01.

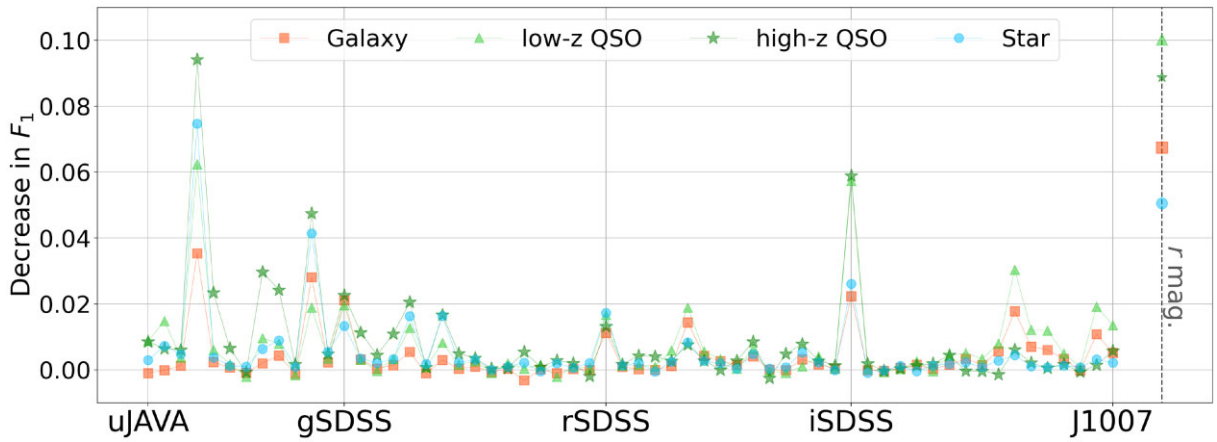

Figure B1. Permutation feature importance analysis in the balanced test set with LGBM. It computes the decrease in the F_1 score of each class when the measurement of a given filter is not available. The input features of LGBM (shown in the horizontal axis) are the normalized fluxes and the magnitude in the r band (see Section 3.2).

Table D1. LGBM HP settings. Parameters not shown were set as default.

Hyperparameter	Value
`objective`	'Multiclass'
`num_class`	4
`boosting`	'GBDT'
`learning_rate`	0.1
`num_leaves`	31
`max_depth`	6
`early_stopping_rounds`	200

Table D2. RF HP settings. Parameters not shown were set as default.

Hyperparameter	Value
`n_estimators`	100
`criterion`	'gini'
`max_depth`	None
`min_samples_split`	100
`min_samples_leaf`	20
`max_features`	'Auto'
`max_samples`	None
`bootstrap`	True
`random_state`	2
`class_weight`	{0:1, 1:1, 2:1.47, 3:3.11}

This paper has been typeset from a TEX/LATEX file prepared by the author.

APPENDIX D: HYPERPARAMETER TUNING

This section describes the hyperparameter (HP) setting of the DT-based models. There are automated ways to set HPs, e.g. with grid search, but these might be computationally expensive. In this work, we performed a simple manual selection, by varying a few HPs that we consider relevant to monitor overfitting and underfitting. The best set of HPs was chosen according to the performance in the validation set. Parameters not shown were set as default.

For LGBM, we tried varying the boosting type to search for better performance and computational gains, and the HPs shown in Table D1 to monitor overfitting, such as the number of leaves and maximum depth of a tree. The number of trees (`n_iterations`) is conditioned to `early_stopping_rounds`, which interrupts training after 200 iterations without improving the loss in the validation set.

For RF, we tried different values for the parameters shown in Table D2. Although we do not limit the depth of each tree (`max_depth` = None), we avoid overfitting by: (i) using the bagging strategy to create a tree; i.e. we set `bootstrap` = True, draw a sample (with replacement) equally sized to the training set (`max_samples` = None), and sample $\sqrt{n}$ features (`max_features` = 'auto'), where n is the total number of features; and (ii) increasing the required number of instances to perform a split and to create a leaf in the trees (`min_samples_split` and `min_samples_leaf`, respectively). We also find an improvement by weighting the two types of quasars to match the proportion of stars and galaxies.