Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **507,** 4425–4444 (2021) https://doi.org/10.1093/mnras/stab2142
Advance Access publication 2021 July 24

Galaxy morphological classification catalogue of the Dark Energy Survey Year 3 data with convolutional neural networks

Ting-Yun Cheng [1,2]★ Christopher J. Conselice,[2,3] Alfonso Aragón-Salamanca,[2] M. Aguena [4,5] S. Allam,[6] F. Andrade-Oliveira,[5,7] J. Annis,[6] A. F. L. Bluck,[8,9] D. Brooks,[10] D. L. Burke,[11,12] M. Carrasco Kind,[13,14] J. Carretero [15] A. Choi,[16] M. Costanzi,[17,18,19] L. N. da Costa,[5,20] M. E. S. Pereira,[21] J. De Vicente,[22] H. T. Diehl,[6] A. Drlica-Wagner,[6,23,24] K. Eckert,[25] S. Everett,[26] A. E. Evrard [21,27] I. Ferrero [28] P. Fosalba [29,30] J. Frieman,[6,24] J. García-Bellido,[31] D. W. Gerdes,[21,27] T. Giannantonio,[9,32] D. Gruen [11,12,33] R. A. Gruendl,[13,14] J. Gschwend,[5,20] G. Gutierrez,[6] S. R. Hinton [34] D. L. Hollowood [26] K. Honscheid,[16,35] D. J. James,[36] E. Krause,[37] K. Kuehn,[38,39] N. Kuropatkin,[6] O. Lahav,[10] M. A. G. Maia,[5,20] M. March [25] F. Menanteau,[13,14] R. Miquel,[15,40] R. Morgan,[41] F. Paz-Chinchón,[13,32] A. Pieres,[5,20] A. A. Plazas Malagón,[42] A. Roodman,[11,12] E. Sanchez,[22] V. Scarpine,[6] S. Serrano,[29,30] I. Sevilla-Noarbe,[22] M. Smith [43] M. Soares-Santos,[21] E. Suchyta [44] M. E. C. Swanson,[13] G. Tarle,[21] D. Thomas [45] and C. To[11,12,33]

Affiliations are listed at the end of the paper

Accepted 2021 July 21. Received 2021 July 2; in original form 2021 April 25

ABSTRACT

We present in this paper one of the largest galaxy morphological classification catalogues to date, including over 20 million galaxies, using the Dark Energy Survey (DES) Year 3 data based on convolutional neural networks (CNNs). Monochromatic *i*-band DES images with linear, logarithmic, and gradient scales, matched with debiased visual classifications from the Galaxy Zoo 1 (GZ1) catalogue, are used to train our CNN models. With a training set including bright galaxies ($16 \leq i < 18$) at low redshift ($z < 0.25$), we furthermore investigate the limit of the accuracy of our predictions applied to galaxies at fainter magnitude and at higher redshifts. Our final catalogue covers magnitudes $16 \leq i < 21$, and redshifts $z < 1.0$, and provides predicted probabilities to two galaxy types – ellipticals and spirals (disc galaxies). Our CNN classifications reveal an accuracy of over 99 per cent for bright galaxies when comparing with the GZ1 classifications ($i < 18$). For fainter galaxies, the visual classification carried out by three of the co-authors shows that the CNN classifier correctly categorizes discy galaxies with rounder and blurred features, which humans often incorrectly visually classify as ellipticals. As a part of the validation, we carry out one of the largest examinations of non-parametric methods, including ∼100 ,000 galaxies with the same coverage of magnitude and redshift as the training set from our catalogue. We find that the Gini coefficient is the best single parameter discriminator between ellipticals and spirals for this data set.

Key words: methods: data analysis – methods: observational – catalogues – galaxies: structure.

1 INTRODUCTION

Galaxy morphology is linked to the stellar populations of galaxies, providing essential clues to their formation history and evolution. Visual morphological classification was pioneered by Hubble (1926). His system initially had two broad galaxy morphological types, early-type galaxies (ETGs) and late-type galaxies (LTGs), based on their appearance in optical light. These two broad categories connect galaxy morphology with a variety of stellar and structural properties. For instance, ETGs are dominated by older stellar populations and have no spiral structure, while LTGs usually contain a younger stellar population and often have spiral arms. These differences in stellar properties indicate that galaxies with different morphologies are at different evolutionary stages and evolution paths. Therefore, the availability of galaxy morphologies for very large samples is of great importance when studying the formation and evolution of galaxies.

Conventionally, visual assessment is the main method of galaxy morphological classification (e.g. de Vaucouleurs 1959, 1964; Sandage 1961; Fukugita et al. 2007; Nair & Abraham 2010; Baillard et al. 2011). Since around 2000, there has been a significant growth in the size of imaging data sets and increasingly complex ones from e.g. the *Hubble Space Telescopes*. Due to this and the development of computational capacity, non-parametric methods

★ E-mail: ting-yun.cheng@durham.ac.uk

were developed such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), the Gini coefficient, and the M20 parameter (Abraham, van den Bergh & Nair 2003; Conselice 2003; Lotz, Primack & Madau 2004; Law et al. 2007). There are good indications that these parameters, which make no assumptions about the galaxy, are largely free from subjective biases. However, even these computational methods become challenging to apply when the astronomical data become too large and we have to use Big Data techniques and machine learning. We are now in this era with the extensive imaging now provided by the Dark Energy Survey[1] (DES; Abbott et al. 2018) that has imaged over hundreds of millions of galaxies. This is just the first of many upcoming imaging surveys that will be carried out in the coming decade, including from the Vera Rubin Observatory and *Euclid*.

Another successful approach for carrying out large-scale morphological analyses is the 'Galaxy Zoo' projects (Lintott et al. 2008, 2011; Willett et al. 2013), designed initially for classifying galaxies in the Sloan Digital Sky Survey (SDSS). This Galaxy Zoo is such that amateurs classify galaxies by answering a series of questions based on galaxy images through an online interface. Studies resulting from Galaxy Zoo show the usefulness of the input from non-professionals in morphological classification of galaxies. With many volunteers, this process accelerates the classification procedure by including the general public rather than limiting these efforts to experts. However, the size of astronomical data generated by large-scale surveys such as DES and future surveys such as the Vera Rubin Observatory Legacy Survey of Space and Time and the *Euclid Space Telescope* has increased to the stage that it would take of the order of >100 yr to classify with Galaxy Zoo. Therefore, machine learning techniques are critical for analysing large-scale astronomical data set, such as galaxy images.

The concept of machine learning in computational science started from Fukushima (1975, 1980) and Fukushima, Miyake & Ito (1983). For the past decades, machine learning techniques have been widely used in astronomical studies, such as star–galaxy separation (e.g. Odewahn et al. 1992; Weir, Fayyad & Djorgovski 1995), strong lensing identification (e.g. Jacobs et al. 2017; Petrillo et al. 2017; Lanusse et al. 2018; Cheng et al. 2020b), and finding galaxy mergers (e.g. Bottrell et al. 2019; Ferreira et al. 2020), among many other applications. Since these early papers, the computational capability and machine learning methodologies have made a remarkable improvement and machine learning is becoming a standard tool in astronomical investigations. Specifically within galaxy morphological classifications, there are a slew of studies applying different supervised machine learning approaches (e.g. Huertas-Company et al. 2008, 2009, 2011; Shamir 2009; Dubath et al. 2011; Polsterer, Gieseke & Kramer 2012; Miller et al. 2017; Beck et al. 2018; Sreejith et al. 2018), neural networks (e.g. Maehoenen & Hakala 1995; Naim et al. 1995; Lahav et al. 1996; Ball et al. 2004; Banerji et al. 2010), and convolutional neural networks (CNNs; e.g. Dieleman, Willett & Dambre 2015; Huertas-Company et al. 2015, 2018; Domínguez Sánchez et al. 2018; Cheng et al. 2020a; Ghosh et al. 2020; Hausen & Robertson 2020; Walmsley et al. 2020).

In this new study, we apply the CNN set up and calibration investigated and assembled in Cheng et al. (2020a, hereafter, C20) to predict probabilities of binary galaxy morphological classification for the DES Year 3 GOLD data (hereafter, the DES Y3 data; Sevilla-Noarbe et al. 2020). This project allows us to build one of the largest catalogues of galaxy morphological classification to date, which includes ∼20 million resolved galaxies, along with the companion DES catalogue produced in Vega-Ferrero et al. (2021). Both studies use the DES imaging data; however, there is only an ∼60 per cent overlapping in samples between the two due to different initial sample selection criteria applied. Their approach involves simulating bright galaxies to a fainter magnitude for training, and uses multiband images, while we use single-band images of bright galaxies and include linear, gradient, and logarithmic images to emphasize different shapes and light distribution of galaxies for training. Our paper is therefore based on single-band apparent morphologies, similar to how visual estimates have been carried out for the past 100 yr. The two works use different methodologies and training set-ups as well. The comparison of the two studies is ongoing and will provide a solid validation in morphological classification of the overlap samples using the different approaches. This will give an insight for future deep learning applications in galaxy morphological classification, but this type of detail is beyond the scope of this catalogue paper. Since it is a large amount of work to compare the two catalogues, which includes more than 20 million galaxies each, a detailed comparison of the two studies is separate from this paper.

The arrangement for this paper is as follows. The data sets are described in Section 2, and we introduce the CNN used in the paper in Section 3. Other catalogues used for validating our CNN predictions are introduced in Section 4. The examination of the predictions is shown in Section 5, while the content of our classification catalogue is presented in Section 6. Finally, we summarize this study in Section 7.

2 DATA SETS

The DES (DES Collaboration 2005, 2016) is a wide-field optical imaging survey covering 5000 square deg (∼1/8 sky; Neilsen et al. 2019) that partially overlaps with the survey area of the SDSS, but has a better imaging quality and deeper depth than the SDSS images. The Dark Energy Camera (Flaugher et al. 2015) is used in DES that has a high quantum efficiency in the red wavebands (>90 per cent from ∼650 to ∼900 nm), and gives images with a good image quality for imaging observations of distant objects compared with previous surveys with the spatial resolution of 0.263 arcsec per pixel and the single epoch depth of $i = 22.51$ (Abbott et al. 2018). An individual DES survey exposure has more than 500M pixels. Each coadd (tile) image covers 1/2 square deg and has a size of $10\,000 \times 10\,000$ pixels. To create the galaxy stamps for this study, we follow the guideline in C20 (details in Section 2.1) and apply the same pre-processing procedure used in the paper to both the training set (Section 2.2) and the DES Y3 data (Section 2.4). In the next subsections, we give an overview of how we prepare our data for analysis from the DES imaging.

2.1 Pre-processing

The data preparation we use closely follows the procedure described in C20. There are two main parts of the data preparation: (1) stamp creation and (2) image processing. Fig. 1 shows the pre-processing procedure used in this study. Using the DES GOLD catalogues, we cut the original coadd images, which have a size of $10\,000 \times 10\,000$ pixels, into many different 'postage stamp' images – creating millions of galaxy stamps with sizes of 50×50 pixels (approximately 13 arcsec $\times$ 13 arcsec). When a galaxy size, as given in the DES catalogue, is larger than the size threshold (30×30 pixels), a larger 200×200 pixel stamp is cut from the images, and then re-sampled to produce a 50×50 pixel image by calculating the mean value in

[1] https://www.darkenergysurvey.org/

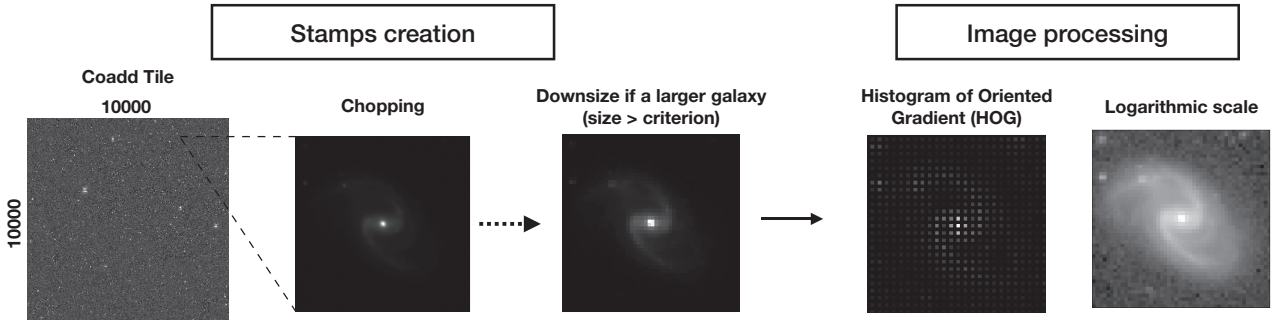

Figure 1. Pre-processing procedure pipeline that is used to prepare our galaxy sample for analysis (see details in Section 2.1). This shows the chopping, resizing processes (if needed), and imaging processing we utilize, including HOG and using a logarithmic scaling as input.

4×4 pixel blocks. This is done for a very small fraction of the galaxy sample since over 99 per cent of all DES galaxies are smaller than 25×25 pixels. Additionally, when creating stamps for the training set, each image is rotated by different angles to increase the number of training images (see Section 2.2).

In the second step, we create two extra images that are both included in training our CNN models. One is an image with gradient features that we obtain by a feature extraction technique called the histogram of oriented gradient (HOG; Dalal & Triggs 2005). The HOG, as a feature extractor, is a well-known technique within pattern recognition studies, e.g. human detection, face recognition, and handwriting recognition (e.g. Dalal & Triggs 2005; Shu, Ding & Fang 2011; Kamble & Hegadi 2015, etc.). In astronomy, it has already been used in a few of studies such as spectral lines observation (Soler et al. 2019), gravitational lensing detection (Avestruz et al. 2019), and galaxy morphological classification such as our previous work (Cheng et al. 2020a).

The key feature of HOG is to characterize the local appearance and the shape of objects based on local intensity gradients (Dalal & Triggs 2005). This technique calculates the gradients of the horizontal (x) and vertical (y) directions of stamps. The magnitude and orientation of the gradient are calculated as below

$$|G| = \sqrt{G_x^2 + G_y^2}, \qquad \theta = \arctan\left(\frac{G_y}{G_x}\right), \tag{1}$$

where $|G|$ is the gradient magnitude of each pixel, G_x is the gradient magnitude measured in the x-direction, G_y is measured in the y-direction, and θ is the orientation of the gradient for each pixel in the images. It then measures the contribution of gradients from each pixel in the cell with a size of 2 by 2 pixels, and describes these using a histogram of different orientation angles. We rescale the HOG output images so that their pixel values are between 0 and 1 (hereafter, HOG images), and use them as one of the inputs to train our CNN models.

In addition to the HOG images, the other input we use is the image itself within a logarithmic scale (hereafter, log images). In C20, we tested the impact of using log images to train the CNN algorithms. We show in C20 that the improvement rate by using log images is positive, but decreased when the number of training data is increased. Therefore, there might not be a significant improvement provided by log images in our case. However, in order to completely consider different significant features in our images, we decide to include the log images with rescaled pixel values between 0 and 1 when training the final CNN models for the task of catalogue construction.

2.2 Training data – DES Y1 data

The training data used throughout are described in C20, which is the subset of the first year DES GOLD data (DES Y1 data), the DES observation of SDSS stripe 82 (Drlica-Wagner et al. 2018) and matched with the visual binary classifications from the Galaxy Zoo 1 project (GZ1[2]; Lintott et al. 2008, 2011; Section 2.3). In this paper, the morphological classification catalogue is built based on monochromatic i-band images only, due to the limitation of our computational resources and the cost of computational time to generate the pre-processed images and the memory storage of the enormous size of the DES Y3 data.

We directly used the visual classification (with over 80 per cent vote agreements) provided in Lintott et al. (2011, 'Flags' from their table 2; '*morphological flags*' hereafter), giving us 2862 galaxies with classification labels in total to train our machine. The intrinsic ratio between the number of spirals and ellipticals in this catalogue is $\sim$3. The magnitude range of the overlap data ranges from $16 \leq i < 18$, and their redshifts are all at $z < 0.25$. However, in C20, we show how to correct the labels for $\sim$2.5 per cent of our sample galaxies that are found to be mislabelled in GZ1 by comparing to the DES data, which has a better resolution and deeper depth than the SDSS data, in which a few galaxies are mislabelled due to the debias process carried out in GZ1 for creating the *morphological flags* (details in Section 2.3). Additionally, $\sim$0.56 per cent of galaxies that we cannot confirm the classification for according to our test in C20 are excluded from our final training set. The final number of galaxies in our initial training set is 2846, with a ratio of $\sim$3 between the number of spirals and ellipticals.

The training set is prepared following the pipeline shown in Fig. 1. Considering we have a limited amount of labelled data, to prevent from overfitting during the training process, an extra process of rotating images is performed to increase the number of the training data. An extra amount of Gaussian noise (mean = 0, variance = 1E-8) is also added, which is negligible towards causing any impact to the signal-to-noise ratio, the visual appearance, and the structure of galaxies, but shows a detectable change of pixel values (Dieleman et al. 2015; Huertas-Company et al. 2015). We do this to increase the variation of pixel values while increasing the number of training sets. Finally, we retain the balance between the numbers of elliptical (E) and spiral galaxies (S) in the training set that is proved to be an important factor in C20. This rotational operation increases the number of data for training purposes (including training and

[2]https://data.galaxyzoo.org/

validation) to 54 133 galaxy stamps with the ratio of number of types held to $E/S \sim 1$.

2.3 The GZ1 catalogue

The Galaxy Zoo projects are among the most successful attempts using citizen science to obtain large numbers of galaxy morphological classifications. A set of questions are asked to the volunteers for each galaxy image. Based on the answers from the volunteers, the GZ1 statistically provides the morphological classification of ∼900 000 galaxies. Of these, ∼670 000 galaxies with spectroscopic redshifts have been bias corrected (Bamford et al. 2009).

In this study, we use three main pieces of classification information from GZ1: *raw votes*, *debiased votes*, and *morphological flags*. The *raw votes* are the likelihood calculated directly from the volunteers' votes for each image. The *debiased votes* and *morphological flags* are derived after applying bias corrections based on different assumed ellipticals/spirals ratio (*E*/*S* ratio; Bamford et al. 2009; Lintott et al. 2011).

In GZ1, a correction factor is necessary to account for a classification bias that depends on the apparent brightness and size of each galaxy. For example, when viewing a spiral galaxy at higher redshift, its decreasing apparent brightness and size make it more difficult to appreciate morphological details such as spiral arms, resulting in an increased likelihood of it being classified as an elliptical galaxy. The corrections needed to account for this bias are calculated by assuming that the morphological mix does not evolve significantly in the narrow redshift range covered by GZ1 (Bamford et al. 2009). This assumption has been shown to be a reasonable one (Conselice, Blackburne & Papovich 2005).

In order to perform this bias correction, GZ1 use two different values for the *E*/*S* ratio, one to obtain the *morphological flags* and a different one to estimate the *debiased votes*. The *morphological flags* provided by Lintott et al. (2011) are determined using the *E*/*S* values that only take into account the classifications with at least a 0.8 morphological vote fraction. On the other hand, the *debiased votes* provided by GZ1 are based on *E*/*S* ratios that use the raw likelihood.

In C20, we note that some galaxies with less accurate *morphological flags* after the bias correction from the GZ1 still showed a questionable label when comparing with our CNN predictions. Therefore, through repeated tests of our CNN and visual assessment, we correct the labels for ∼2.5 per cent of our sample galaxies, and excluded ∼0.56 per cent galaxies that we cannot confirm the classification for based on our tests in C20. The corrected labels in C20 are shown to better correspond to the classification based on the *debiased votes* in Lintott et al. (2011, 'Debiased votes' in their table 2), which is debiased based on the *E*/*S* ratio using the likelihoods directly (Bamford et al. 2009). The *debiased vote*, as stated on the website of the GZ1, corrects well the bias that existed in GZ1 visual assessment, and provides a more accurate classification than the *morphological flags*. Therefore, although our CNN model is trained with the *corrected morphological flags* based on the DES imaging data, we validate our CNN predictions with the classification based on *debiased votes* (Section 5.1).

2.4 DES Year 3 data

The data used to build the catalogue presented in this work are from the DES Y3 GOLD catalogue (Sevilla-Noarbe et al. 2020). We initially use the images that are selected with the flags shown in Table 1 and within a magnitude range of $16 \leq i \leq 22$. The top two flags guarantee that astronomical objects selected using these flags are most likely to be galaxies. This is such that the galaxy sample, as defined by these flags, has a rate of contamination of point sources less than ∼2 per cent at fainter magnitudes, as derived by comparing with the HSC-SSP DR2 catalogue (Aihara et al. 2018). The bottom four flags ensure that objects have a consistency with the Y3 GOLD footprint, denote quality selection for clean samples, and are used to select the data with a reliable SExtractor (Bertin & Arnouts 1996) analysis.

Table 1. The flags used to select data in the DES Y3 GOLD catalogue. The first two flags guarantee that the astronomical objects that are the most likely to be a galaxy are selected, and the last four flags indicate the samples are clean, consistent with the Y3 GOLD footprint, and with a reliable analysis from the SExtractor (Bertin & Arnouts 1996).

Selection flags	
EXTENDED_CLASS_COADD	= 3
EXTENDED_CLASS_WAVG	= 3
FLAGS_FOOTPRINT	= 1
FLAGS_FOREGROUND	= 0
bitand(FLAGS_GOLD,120)	= 0
bitand(FLAGS_BADREGIONS,1)	= 0

This selection provides over 50 million galaxies for the initial task, with the redshift distribution of the selected data peak at $z \sim 0.4$ with over 99.9 per cent of the galaxies at $z \leq 1.2$. The number of galaxies in each magnitude bin increases exponentially when going fainter. A pre-processing procedure described in Section 2.1 is also applied to the selected data.

The selected data are separated into six magnitude bins from $i = 16$ to 22 for further analysis (Section 5). After an examination carried out in Section 5.2, our final catalogue includes over 20 million galaxies with the magnitude range of $16 \leq i < 21$.

The galaxies in the final catalogue have a wider range of magnitudes and redshifts than those in the training set – the training set galaxies are, typically, brighter and have lower redshifts (Section 2.2). Therefore, in our study, we also investigate how the confidence of our CNN predictions might be impacted when applied to galaxies at fainter magnitudes and higher redshifts (Section 5.3). Alternatively, approaches such as using simulated data can help us to build a training set that reaches fainter magnitudes or higher redshifts, e.g. the companion DES morphological catalogue presented in Vega-Ferrero et al. (2021).

3 CNNs

CNNs (Lecun et al. 1998) are a type of neural network that includes convolutional layers used to extract strongly weighted features from input images for a given classification problem. The architecture of the CNN used throughout this paper is shown in Fig. 2. This design is inspired by the best performing architecture used in Dieleman et al. (2015), but with fewer convolutional layers and parameters. The dimension of the inputs is $50 \times 50 \times 3$, with the depth including the linear images, HOG images, and log images. Three convolutional layers with kernel sizes of 3, 3, and 2, respectively, are used in this study, and each of them is followed by a max-pooling layer with a size of 2. The max-pooling layer is also referred to as a 'downsampling' layer, which is used to reduce the spatial size and the numbers of parameters involved in the architecture. After the third convolutional layer, two dense layers with 1024 hidden units for each layer follow. In addition, dropouts (=0.5) are applied to reject irrelevant parameters and prevent overfitting in training the

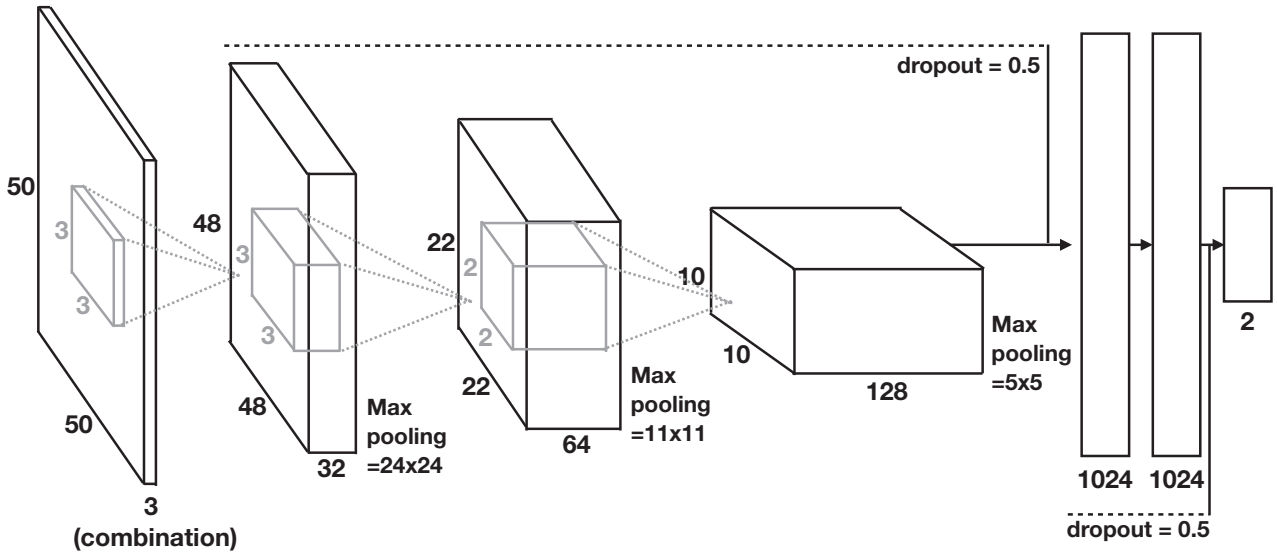

Figure 2. The schematic overview of the CNN architecture used throughout. The architecture starts from an input of dimension $50 \times 50 \times 3$, and is followed by three convolutional layers with kernel sizes of 3, 3, and 2 and channel sizes of 32, 64, and 128, respectively, plus a max-pooling layer after each. Two dense layers with 1024 hidden units are following the third convolution layer. A dropout ($p = 0.5$) is applied after the third convolutional layer and after the second dense layer. Probabilities for two classes are predicted in the final output of our CNN, 'Ellipticals' and 'Spirals'.

CNN. A dropout follows the third convolutional layer (max-pooling layer), and the other one comes after the two dense layers.

The activation function used in the convolutional layers and the dense layers is the Rectified Linear Unit (`ReLu`; Nair & Hinton 2010) such that $f(z) = 0$ if $z < 0$, while $f(z) = z$ if $z \geq 0$. Finally, the `softmax` function (Bishop 2006), $f(z) = \exp(z)/\sum \exp(z^j)$, is applied to the output layer and provides the probability distribution of each type. For the CNN training, we apply *Adam* optimizer, *Nesterov momentum*, and set *momentum* = 0.9 according to Dieleman et al. (2015). The learning rate is set to 0.001, and the maximum number of iterations is 500, with an early-stopping mechanism that triggers when the validation set hits the local minimal loss.

A CNN has the technical advantage of not requiring the pre-processing procedure commonly used in artificial neural networks. However, in C20, we have proven that combining pre-processed images such as HOG images and log images qualitatively improves the performance of our CNN and reaches a final accuracy of over 0.99. Accuracy here is defined as the number of matched classifications by CNN and GZ1 from the total overlapped samples (equation 2). In this study, we independently train the CNN five times with the data sets described in Section 2.2, which is then randomly separated into training and validation sets with a fraction of 0.9 and 0.1 of the total, respectively, in each run. Doing this avoids using exactly same batches for training each run. We then apply these pre-trained models to predict morphological classifications for the DES Y3 data (Section 2.4). The final morphological prediction is obtained by averaging the predicted probabilities of these five independent CNN models for each type – 'Ellipticals' and 'Spirals'.

4 CATALOGUES FOR CROSS-VALIDATION

Once we have the morphological predictions from the CNN for millions of galaxies, it is of great importance to validate the reliability of these classifications. In this study, we compare our CNN predictions with four different resources: (1) the GZ1 catalogue using the galaxies that were not present in the training set (Section 2.3); (2) visual classifications carried out by TC, CC, and AAS[3] (Section 4.1); (3) VIPERS (VIMOS Public Extragalactic Redshift Survey) unsupervised spectral classification (Siudek et al. 2018; Section 4.2); and (4) non-parametric methods using the structural measurements from Tarsitano et al. (2018) (Section 4.3). In DES Y3 GOLD catalogue, a quantity with 'FRACDEV' (Everett et al. 2020), which indicates the fraction of the fitted galaxy profile represented by a de Vaucouleurs model (de Vaucouleurs 1948), may be used to compare with our classification. However, there are some priors used in the assignment of this fraction that might need a further examination for its reliability. Therefore, in this work, we do not use this quantity to compare with our CNN classification. The validation between them could possibly be further investigated in the future work.

4.1 Visual classification of randomly selected subsamples

Visual classification (hereafter, VIS) was carried out by three of the co-authors (TC, CC, and AAS[3]) for a reasonably large number of galaxies. We randomly selected 500 galaxies per magnitude bin from the DES Y3 data set for galaxies with $16 \leq i \leq 22$. For the brightest bins ($16 \leq i < 18$), only galaxies in GZ1 were included. In doing so, we covered the whole magnitude range of the DES sample with a significant overlap with GZ1 for cross-validation.

The classification system we use is displayed in Table 2. We classify galaxies into six categories: Ellipticals (0), Early Spirals (1), Late Spirals (2), Edge-on Spirals (3), Irregulars (4), and Unknown (5). To compare with our CNN predictions, which provides probabilities for binary classification, for the ellipticals and spirals, we merge three subcategories of spiral galaxies into one – Spirals (1), and others retain the original label. The label with the most *combined votes* (Table 2) from our visual classifiers is set as the final visual type of a galaxy. This is the morphological type that is picked by at least two out of three of the classifiers. Those galaxies without a dominant label

[3]TC: Ting-Yun Cheng; CC: Christopher J. Conselice; AAS; Alfonso Aragón-Salamanca.

Table 2. The classification system we applied in the visual classification carried out by TC, CC, and AAS[3]. Galaxies are classified into six categories (*primary votes*) that are then merged into four categories, i.e. Ellipticals, Spirals, Irregulars, and Unknown (*combined votes*; see the text).

Labels	Primary votes	Combined votes
0	Ellipticals	Ellipticals
1	Early spirals	Spirals
2	Late spirals	Spirals
3	Edge-on spirals	Spirals
4	Irregulars	Irregulars
5	Unknown	Unknown

are categorized into the class of 'Unknown'; the relative fraction of these 'unknown' types increases with magnitude. The distribution of each visual type in each magnitude bin is shown in Fig. 3.

In order to validate the VIS, we compared the classifications of brighter galaxies ($i < 18$) with the GZ1 classifications based on the *debiased votes* and *raw votes* (Fig. 4). The *raw votes* directly reflect the votes from the volunteers of the GZ1. The *debiased votes*, as described in Section 2.3, are bias corrected using the *E*/*S* ratio measured directly from the raw likelihood. We apply a threshold of 0.8 to both votes to decide the morphology type with a higher confidence.

In Fig. 4, our VIS classifications show apparently better agreement with the raw votes from the GZ1 volunteers when comparing with the GZ1 debiased votes. The majority of the mismatched cases when comparing with the labels based on the debiased votes occur when a galaxy is classified as Elliptical by our visual classifications. This indicates that our judgement for galaxy morphology is also biased by the size, magnitude, and redshift of the galaxies. This gets worse when a galaxy is fainter, which is shown in Fig. 3. It is clear that significantly more galaxies are visually classified as Ellipticals.

Although our visual classification suffers from the same type of biases as GZ1, unfortunately we cannot perform a bias correction similar to the one they carried out. There are several reasons for this. First, the broader redshift range of our sample makes the assumption of unevolving morphological mix unreliable. Secondly, the number of galaxies we have been able to classify is too small to provide reliable correction statistics. Thirdly, the lack of spectroscopic redshifts would render any redshift-dependent correction highly uncertain. Therefore, additional factors such as Sérsic index (Sérsic 1963, 1968) and colour will be considered to validate the CNN predictions (Section 5.2).

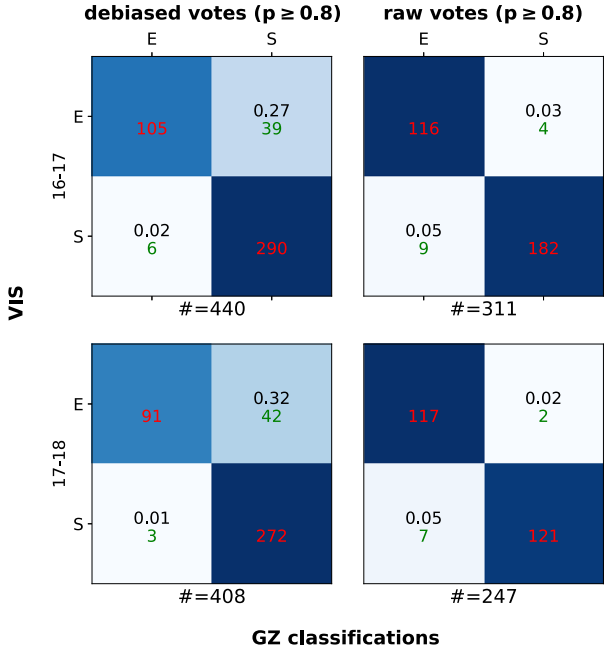

Figure 4. The confusion matrices between our visual classifications (VIS) and the GZ1 classifications based on the *debiased votes* (first column) and *raw votes* (second column) (Lintott et al. 2011). A threshold of 0.8 is applied to both votes here to select high confidence classifications. Rows are separated by different magnitude bins: $16 \leq i < 17$ (first row) and $17 \leq i < 18$ (second row).

4.2 Unsupervised spectral classification

With the known correlation between spectral classification of galaxies and galaxy morphology (e.g. Morgan & Mayall 1957; Bershady 1995; Zaritsky, Zabludoff & Willick 1995; Kennicutt 1998; Baldry et al. 2004, etc.), we compare our predictions within a fainter magnitude ($i \geq 18$) with an unsupervised spectral classification presented in Siudek et al. (2018) from the VIPERS. This provides a different way to examine the robustness of our CNN predictions, although with some caveats. These spectral classifications employ a Fisher Expectation-Maximization (FEM) unsupervised algorithm to categorize galaxies with redshifts $z \sim 0.4$–1.3 into 12 classes based on 12 rest-frame magnitude and spectroscopic redshift. Except for the class 12, which is the class of broad-line active galactic nuclei, other classes can be classified into three main categories: passive

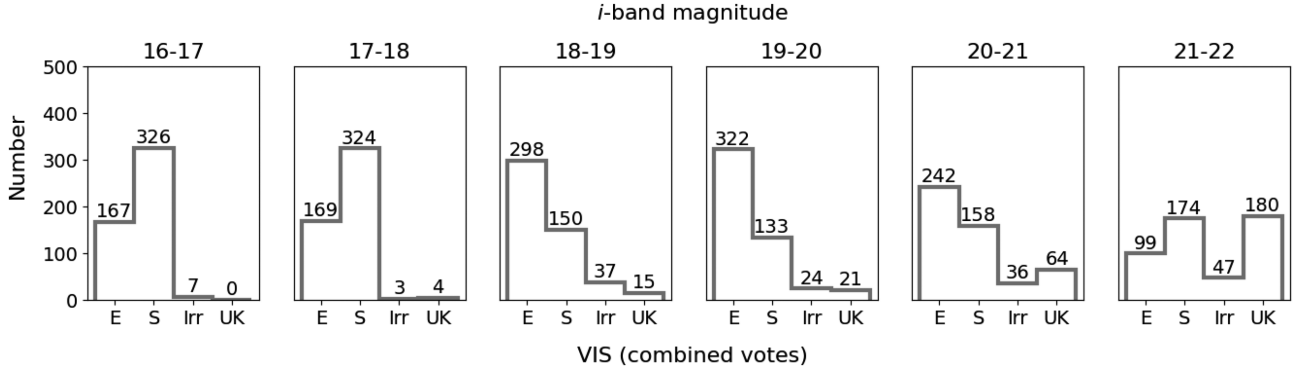

Figure 3. The frequency distribution of each visual classification in each magnitude bin. On the *x*-axis, 'E', 'S', 'Irr', and 'UK' are short for Ellipticals, Spirals, Irregulars, and Unknown, respectively. The number above each bar represents the number of galaxies visually classified in that category. The range of the magnitude in the *i* band is shown at the top of each panel.

(class 1–3), intermediate (class 4–6), and star-forming galaxies (class 7–11).

4.3 DES Y1 catalogue of morphological measurements

To obtain a reliable analysis of the quality of our CNN labels, parametric factors such as the Sérsic index and non-parametric coefficients such as *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), Gini coefficient, and M20 are used in this study. Tarsitano et al. (2018) included 45 million objects selected from the first year DES data, and provided the largest structural catalogue to date for galaxies. The selected samples from this catalogue cover the magnitude range of $i \leq 23$. According to the suggestions from the paper, we apply an initial cut as follows:

(i) MAG_AUTO_I ≤ 21.5
(ii) SN_I > 30
(iii) SG > 0.005,

where *MAG_AUTO_I* represents the cut in *i*-band apparent magnitude and *SN_I* is the signal-to-noise ratio in the *i* band. The *SG* is used for optimizing the separation between stars and galaxies while maintaining the completeness. The cut ($SG > 0.005$) recommended in Tarsitano et al. (2018) is the optimal compromise between the completeness and purity of the galaxy sample. These selections provide 12 million galaxies with 90 per cent completeness in Sérsic measurements and 99 per cent completeness in non-parametric measurements in the *i* band.

The parameters provided from the single Sérsic fits (e.g. Sérsic index, ellipticity, etc.) are measured with GALFIT (Peng et al. 2010). We then apply a further cut suggested in Tarsitano et al. (2018) to select the galaxies that are successfully validated and calibrated. The calibration is made based on four parameters: size, magnitude, Sérsic index, and ellipticity using simulated galaxies generated with these parameters (Tarsitano et al. 2018):

(i) FIT_STATUS_I $= 1$.

On the other hand, the non-parametric parameters (CAS parameters, Gini, and M20) are measured using the Zurich Estimator of Structural Types (ZEST+; Scarlata et al. 2007a, b). The calibration is applied with the same procedure as the parameter fit but uses concentration instead of Sérsic index for non-parametric parameters, and the validation is discussed on the Gini–M20 plane as a function of other morphological measurements such as concentration (C), asymmetry (A), and clumpiness (S) (Tarsitano et al. 2018). One criterion is applied in non-parametric coefficients to select the objects with successfully validated and calibrated measurements.

(i) FIT_STATUS_NP_I $= 1$

5 VALIDATION AND DISCUSSION

In this section, we carry out the cross-validation of our CNN predictions using multiple sources listed in Section 4. Included among this, we also discuss the confidence levels assigned to the predictions with a probability threshold of 0.5 and the uses of this catalogue with these confidence assignment; that is, we explain how to use our catalogue for determining galaxy morphologies.

Some quantities are used to examine the performance of our CNN classifications such as accuracy, precision (Prec), recall (R), true positive rate (TPR; the same definition as R), and false positive rate (FPR). Accuracy is defined as the number of correct classifications compared to the 'true' labels from the total samples. In equation (2),

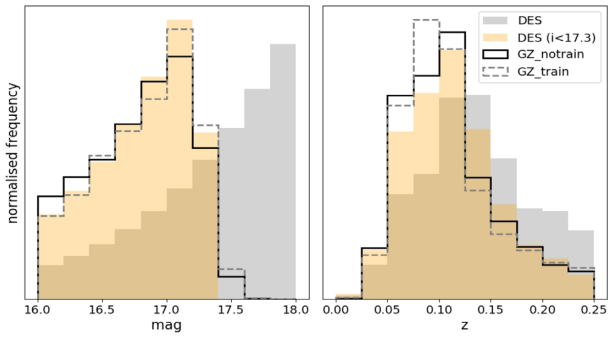

Figure 5. The magnitude and redshift distribution of the DES Y3 data with the same coverage as the training set (Section 2.2). The grey and yellow shadings represent the DES Y3 data without and with a cut at $i = 17.3$, respectively. The solid lines show the overlap region with the GZ1 catalogue, excluding the training set, while the dashed lines show only the training set.

'T' and 'F' represent 'True' and 'False' while 'P' and 'N' denote 'Positive' and 'Negative', respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \quad (2)$$

Precision and recall are defined as follows:

$$R = \frac{TP}{TP + FN}; \quad Prec = \frac{TP}{TP + FP}. \quad (3)$$

Finally, TPR (same definition as R) and the FPR are defined as below:

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}. \quad (4)$$

5.1 GZ1 catalogue

To validate our CNN predictions, first we compare our CNN classifications with the GZ1 labels based on the *debiased votes* (Section 2.3). We do this by matching our DES Y3 data with the GZ1 catalogue that provides us with ∼2700 additional galaxies that are not within the training set. These additional samples are used to examine our CNN predictions in this section. The distribution of the DES Y3 data for this test is in the same magnitude and redshift range as the training set (Section 2.2) as shown in Fig. 5. Note that there are significantly fewer faint galaxies at $i > 17.3$ in our sample with overlapping GZ1 classifications. Therefore, a cut of $i = 17.3$ is applied when carrying out the analysis in this subsection. We then discuss the performance of the CNN predictions below and above this magnitude limit in later sections.

First, in Fig. 6, we show the change in accuracy when applying different likelihood thresholds to the GZ1 debiased votes. The first two panels are separated by the magnitude cut $i = 17.3$, and the third panel contains all overlapping data between GZ1 and DES Y3 data used in this paper. The accuracy of the training set is represented by the black lines. One applies a probability threshold of $p = 0.5$ to our CNN predictions (dashed line), while the other applies a threshold of $p = 0.8$ (solid line). The comparison of the results using different likelihood thresholds at various GZ1 debiased votes is shown by the blue lines. The line styles reflect the same meaning as the black lines. Meanwhile, the second *y*-axis is used for the shading bars that show the number of galaxies under the likelihood threshold. Both Prec and R, as defined in equation (3), of each data point in Fig. 6 are very high. For example, both Prec and R are ≥0.97 at each data point when CNN uses $p = 0.5$, while the two values are ≥0.98 with $p = 0.8$ on CNN.

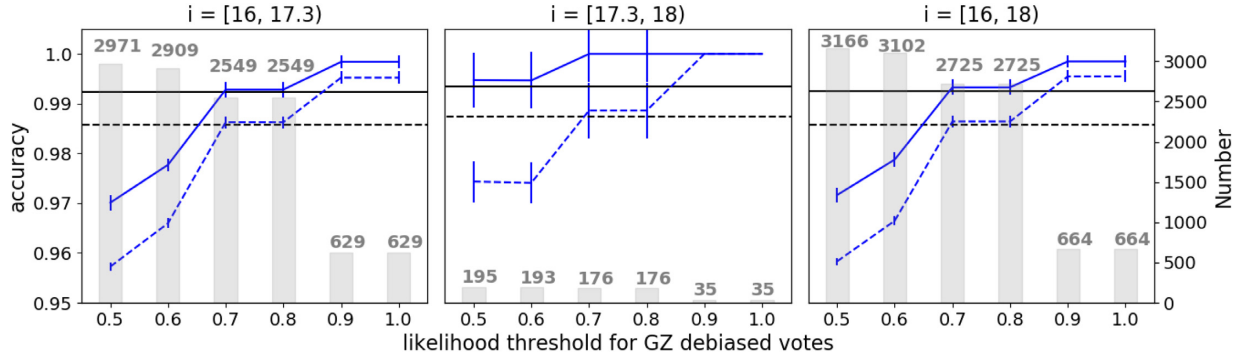

Figure 6. The comparison between the accuracy of the CNN predictions when applying different likelihood thresholds to the GZ1 debiased votes. A magnitude cut $i = 17.3$ is applied in the first two panels, and the third panel includes total data in the first two panels. The black lines are the accuracy of the training sets when applying a probability threshold, $p = 0.5$ (dashed line) and $p = 0.8$ (solid line), to our CNN predictions. The blue lines represent the change of accuracy when comparing our CNN predictions, based on a probability cut $p = 0.5$ (dashed line) or $p = 0.8$ (solid line), with the GZ1 classifications based on different likelihood thresholds. The second y-axis on the right reflects the height of the shading bars, which gives the number of data points left after applying a likelihood threshold. The error bars represent the standard deviation of the accuracy obtained within five models.

Table 3. Content of the *confidence_flag* (column 7) shown in Table 4. The 'superior confidence' flag is for classifications within the same magnitude and redshift ranges as the training set. The details of other levels are described in Section 5.3. The total number of classifications provided in this catalogue is 21 119 107.

Labels	Representation	Number of galaxies
4	Superior confidence	672 927
3	High confidence	3409 459
2	Confidence	9230 182
1	Less confidence	4347 472
1∗	Less confidence (for Spirals only)	2599 656
0	No confidence	859 411
Total		21 119 107

Table 4. Content of the catalogue published with this paper. Columns 9 to 10 are quantities that are taken directly from the DES Y3 GOLD catalogue, and the corresponding column names are highlighted and placed within brackets in the description.

Col.	Keyword	Description
1	DES_Y3_ID	DES Y3 ID
2	RA	Right ascension
3	DEC	Declination
4	pE	Probability of being ellipticals
5	pS	Probability of being spirals
6	MORPH_FLAG	CNN predictions with a probability threshold of 0.8
7	confidence_flag	Confidence level for predictions
8	frac_n	Fraction of predictions that satisfy Sérsic index criteria
9	MAG_I	i-band magnitude *(MAG_AUTO_I)*
10	ZMEAN	Photometric redshift *(DNF_ZMEAN_MOF)*

We note that the accuracy of our CNN predictions compared with the GZ1 classifications based on a debiased likelihood threshold of 0.8 shows a good consistency with the accuracy of the training set (the first panel in Fig. 6). In the second panel ($17.3 \leq i < 18$), the CNN shows a slightly better performance than the brighter range ($16 \leq i < 17.3$) and training set. However, the scatter for the CNN predictions is larger because there are significantly fewer samples in the second plot. When taking the scatter into account, the performance of our CNN predictions in this magnitude range also shows a good consistency with the training set. Therefore, based on Fig. 6, we interpret that there is a 'superior confidence' level to the CNN predictions within the brighter magnitude range of $16 \leq i < 18$ and redshift range of $z < 0.25$. Additionally, in later analysis, we apply a likelihood threshold of 0.8 to the GZ1 debiased votes to determine the GZ1 classifications for comparison. In our catalogue, we provide a classification flag with a probability threshold of 0.8 (89 per cent of the total samples; *MORPH_FLAG* in Table 4) to reject samples with low predicted probabilities from the CNN model. The prediction with a probability lower than 0.8 is labelled as 'uncertain (-1)' in our catalogue (Section 6). Several reasons might result in the low predicted probabilities, which has been discussed in C20. One of the possible reasons is caused by stellar contamination in the DES galaxy sample at $i < 18$.

In the first three panels of Fig. 7, we show confusion matrices within a certain magnitude range as listed above the graph. The x-axis indicates the CNN predictions with a probability threshold of 0.8, while the y-axis shows the GZ1 classifications with a vote threshold of 0.8. We focus on classifying galaxies into two types, namely Ellipticals (E) and Spirals (S). The numbers at the bottom of the confusion matrices show the number of galaxies that are not classified as uncertain type within the ranges in each column. For the first three plots, we exclude the training set, and compare the performance with the training set in the last panel. In this figure, we notice that the two labels (GZ1 and CNN) match well, and the majority of mismatches occur in the case where the CNN classification is Spiral, but the debiased GZ1 classification disagrees.

Fig. 8 showcases the galaxies that are classified as Spirals by the CNN but Ellipticals by the GZ1. Some galaxies in this category show discy structures (e.g. [1], [3], [5], [14], and [15]) or asymmetric features (e.g. [8] and [9]) in the DES imaging data. In C20, we proved that the higher quality DES imaging data reveal detailed structures that were not detected in the data from SDSS. This condition explains the mismatched classifications happened in Fig. 8.

Ideally, we would have liked to examine the Sérsic index distribution of these mismatched galaxies to determine whether these misclassified galaxies have particular structural properties. However,

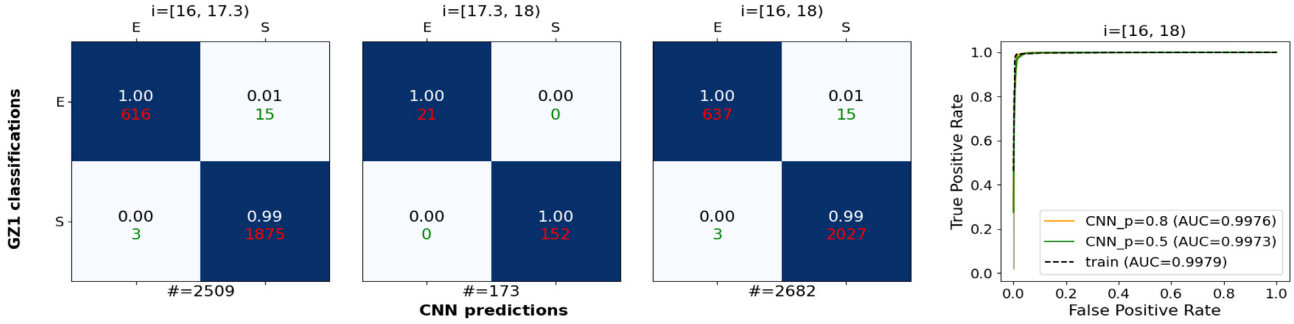

Figure 7. The combined graph of the confusion matrices and the ROC curve comparing our CNN predictions with the GZ1 labels defined by the debiased votes with a threshold of 0.8. The first three panels show the confusion matrices within certain magnitude ranges: $16 \leq i < 17.3$, $17.3 \leq i < 18$, and $16 \leq i < 18$, where the CNN predictions use a probability threshold of 0.8. The red or green colour in each quadrant represents the number of galaxies that agree with the classifications derived through CNN predictions and GZ1 classifications within each quadrant. The number above it indicates the fraction of these galaxies within each certain type decided by our CNN classifier. The number of galaxies within each magnitude range is shown below each graph.
The last panel shows the ROC curve where the *y*-axis represents the TPR, and the *x*-axis is the FPR. The orange and green lines show the curve from our CNN predictions with probability thresholds of 0.8 and 0.5, respectively, while the black dashed line is from the training set.

in this case, there are fewer than three overlapping galaxies with mismatched labels in Tarsitano et al. (2018). The mismatched test sample is far too small for any statistically meaningful analysis. Therefore, we leave this additional cross-validation to future work, when more structural measurements are obtained for the DES Y3 data. We note that, although we cannot carry out this additional test, we are confident of the excellent performance of our CNN predictions within the magnitude and redshift range covered by the GZ1 training set. Based on the discussions above and, in particular, the confusion matrices shown in Fig. 7, we conclude that in this magnitude and redshift range, which includes ∼670 000 galaxies, our CNN classifier has an accuracy of over 99 per cent.

The last panel in Fig. 7 shows a Receiver Operating Characteristic curve (ROC curve; Fawcett 2006; Powers 2011) that is used to examine the performance of our machine learning technique by comparing the probabilities predicted by the machine with the true labels. On an ROC curve, the *y*-axis is the TPR and the *x*-axis is the FPR (equation 4); therefore, the closer an ROC curve gets to the corner (0, 1), the better the performance is.

Another important indicator on the ROC curve is the 'area under the curve', which has a larger value for a better performance of a machine learning model. From the ROC curve, both CNN predictions with probability thresholds of 0.5 (green line) and 0.8 (orange line) within the coverage of the training sets in magnitude ($16 \leq i < 18$) and redshift ($z < 0.25$) show an excellent consistency with the results of the training set. This result doubly confirms our confidence on these predictions. Therefore, the CNN predictions within this range are labelled as 'superior confidence (4)' in the *confidence_flag* (Table 3) in the catalogue, Table 4.

5.2 Visual classification

To allow us to test the quality of the CNN classifications of fainter galaxies, we carried out a visual classification of 500 randomly picked galaxies in each *i*-band apparent magnitude bin with an interval of 1 mag using the DES imaging data. The first five panels in Fig. 9 show the confusion matrices in each magnitude range, and the ROC curve is shown on the last panel. We apply a probability threshold of 0.8 to our CNN predictions in this figure to reject samples with low predicted probabilities. The performance quality of our CNN method drops with magnitude when comparing with the visual classifications.

There are two main reasons responsible for this decreasing performance. First of all, through the confusion matrices, we notice that the majority of mismatches happened in the cases where our CNN method classified a galaxy as a spiral galaxy but we visually classified it as an elliptical galaxy. This situation is caused by the fact that our CNN is trained with the corrected debiased GZ1 classifications (Section 2.3); however, the visual classification used here is a raw classification. In Section 4.1, we pointed out that our visual classifications suffer from a similar classification bias compared with the raw GZ1 classifications that are influenced by the magnitude, size, and redshift of the targets. Therefore, in Fig. 10, we combine the Sérsic index and colour of each galaxy to cross-validate our results. The colour information is obtained using apparent magnitude, which is measured in an elliptical aperture determined by the Kron radius, from the DES Y3 GOLD catalogue. In this work, we use apparent colour for our validation instead of the colour with absolute magnitude. It is due to the large uncertainties in redshift estimation for the DES galaxies that can have a strong effect on absolute magnitude derivation. The Sérsic index is from the DES Y1 morphological measurements (Tarsitano et al. 2018) selected based on the suggested flags (described in Section 4.3). Due to the applied cut in magnitude up to $i = 21.5$ used in Tarsitano et al. (2018), the last panel in Fig. 10 only shows galaxies within the magnitude range of $21 \leq i \leq 21.5$.

In Fig. 10, the central contour shows the density distribution of the Sérsic index and the $(g - i)$ colour at each magnitude. The histograms at the top and the right show their respective normalized frequency distribution. The bottom and left histograms show the misclassified samples colour labelled by the visual classifications. From this figure, it is clear that the majority of misclassified galaxies labelled as Ellipticals by our visual assessment are in fact discier and bluer. Since our CNN is self-debiased by training with the corrected debiased GZ1 labels (Section 2.3), it shows a more sensible classification of the images than humans have difficulty to classify correctly; that is, our CNN classifications are more likely to be correct than the visually based ones.

We remind the reader that our CNN classifier is trained with monochromatic *i*-band images, without any colour information. Therefore, the strong colour segregation between CNN-classified Ellipticals and Spirals is reassuring: The connection between CNN morphology and colour is independent, and not based on the training process – colour and galaxy morphology are linked through galaxy

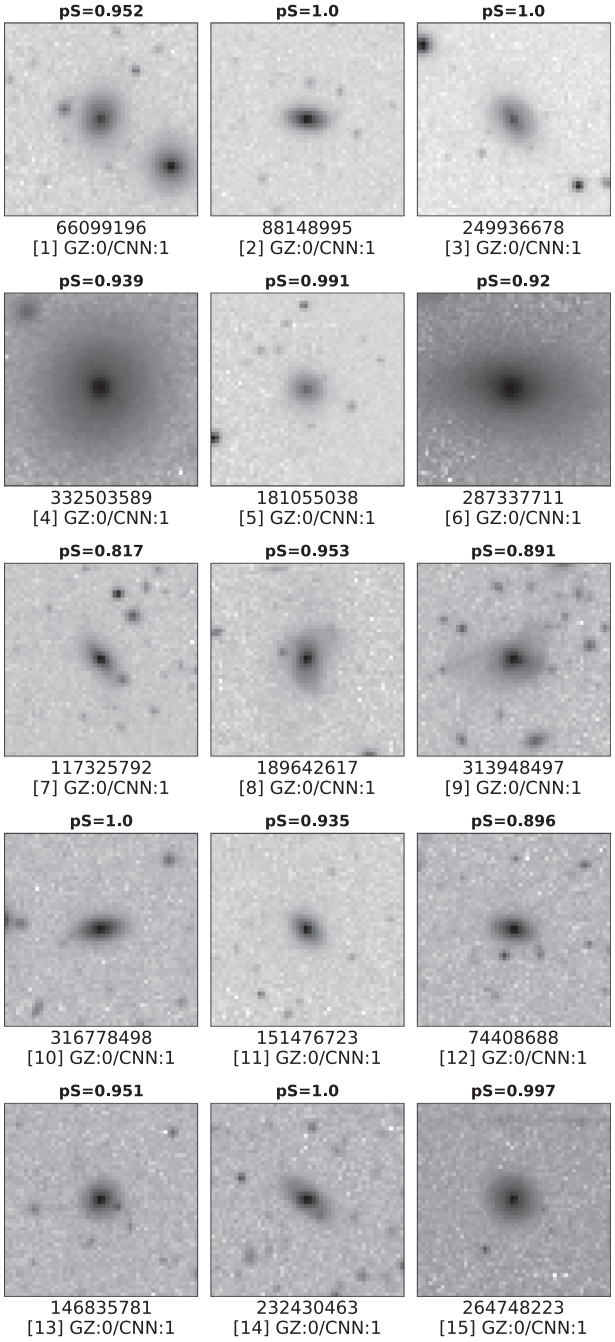

Figure 8. Examples of galaxies that our CNN classified as Spirals while the GZ1 labelled as Ellipticals. The predicted probability of being Spirals from the CNN is shown above each stamp (**pS**).

formation and evolution processes, and are not strongly the result of classification biases.

Secondly, in addition to the bias in the visual classification, another potential reason for the decreasing performance in our CNN classifier is caused by the fact that we train our CNN with brighter galaxies ($16 \leq i < 18$) that are at a lower redshift ($z < 0.25$), and then use this model to predict galaxies in different domains (i.e. galaxies with a fainter magnitude and a higher redshift). Combining with the 'self-bias correction' feature shown in our CNN, we expect that our CNN classifies more disc galaxies at fainter magnitudes. For example, for the faintest magnitude range in our study ($i \geq 21$), the 'self-bias correction' of our CNN classifier is overapplied due to the very low signal-to-noise ratio compared with the training set. This overdone bias correction gives us an artificially low number of Ellipticals classified by the CNN. The ratio of the CNN-classified Ellipticals to Spirals in this magnitude range is $\sim 6 \times 10^{-5}$ for total samples and $\sim 8.4 \times 10^{-5}$ for overlapped samples shown in Fig. 10. The evolution of the *E*/*S* ratio strongly depends on the methods used to classify galaxy morphology, in particular at a high redshift. However, there is not a significant evolution in morphology mix within the redshift range in our sample (over 99.9 per cent of the galaxies at $z \leq 1.2$; Section 2.4), shown in previous studies (e.g. Cassata et al. 2005; Conselice et al. 2005). Therefore, the significant difference in the number of our CNN-classified Ellipticals and Spirals is likely caused by the reason discussed above.

This is shown in both the confusion matrix and the colour–Sérsic diagram: No visually classifiable Ellipticals are picked out by our CNN classifier (Fig. 9), and there is not a clear separation between Ellipticals and Spirals in the Sérsic index distribution (Fig. 10).

Machine learning is sensitive to image qualities such as the signal-to-noise ratios and resolution. In our case, the apparent magnitude of a galaxy, which is influenced by the redshift, affects the signal-to-noise ratio of the galaxy, which can affect how easily structure can be seen. Additionally, due to the effects of distance, a galaxy at a higher redshift shows less detailed structure; i.e. the resolution of the galaxy images decreases. However, there is a certain level of tolerance for variations within these effects, which is still a popular topic to investigate in computational science using images for topics such as object identification, face recognition, etc. (e.g. Amirshahi, Pedersen & Yu 2016; Dodge & Karam 2016; Karahan et al. 2016; Zhou, Song & Cheung 2017; Prakash & Karam 2019, etc.).

For galaxy morphology, a few specific features such as light distribution, spiral arms, disc structures, etc. are dominant when visually classifying galaxies. This gives the possibility of using visual classification in galaxy morphology with images of a low quality. Similar to visual classification, our CNN shows a capability to classify galaxies based on the feature of light distribution and disc structure in Fig. 10, which even shows a likely better classification than human opinion.

Additionally, we note that using monochromatic images we are sampling different rest-frame morphologies at different redshifts. An *i*-band image for an object at a higher redshift, e.g. $z = 1$ (the upper limit in our final catalogue), examines the morphology at ~400 nm. One can debate whether this fact helps or obscures machines in classifying galaxy morphologies at a higher redshift. The different distributions presented in different rest-frame morphologies due to the redshift effect challenge the machine to adapt the domain learned in the training set to a different domain. However, bluer rest-frame morphologies emphasize the feature of spiral arms (location of young stars or star-forming regions), which is one of the dominant features in separating Ellipticals and Spirals. This fact might help the machine to distinguish Spirals even though the image resolution drops at a higher redshift.

Therefore, we statistically investigate the confidence of our CNN predictions at fainter magnitudes and higher redshifts by comparing the quality of our morphologies with our visual assessments, structural measurements such as the Sérsic profile, and galaxy properties such as colour. This analysis also investigates the limit of our CNN classifier, which is trained with bright galaxies at low redshift, on classifying galaxies within different ranges of magnitude and redshift. The detailed discussion of this is in Section 5.3.

As a side note, within the faintest magnitude bin in Fig. 10, even though the CNN-classified Ellipticals are rare and do not have

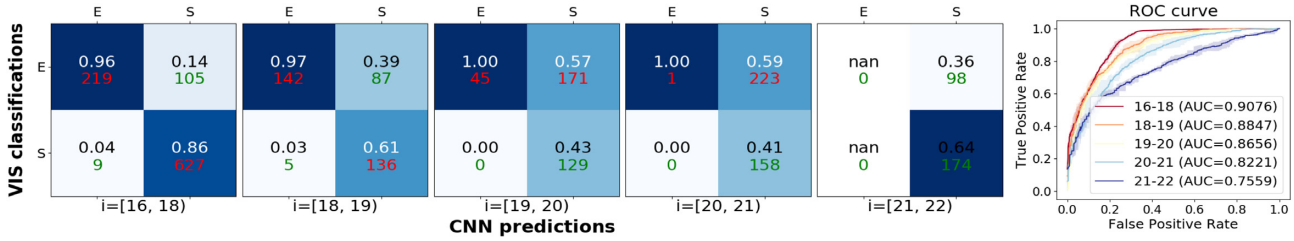

Figure 9. The confusion matrices and the ROC curve of different magnitude ranges. The *x*-axis of the confusion matrices is the CNN predictions with a probability threshold of 0.8 and the *y*-axis is our visual classifications. On the ROC curve, the *x*-axis is the FPR while the *y*-axis represents the TPR. Different colours indicate different magnitude ranges.

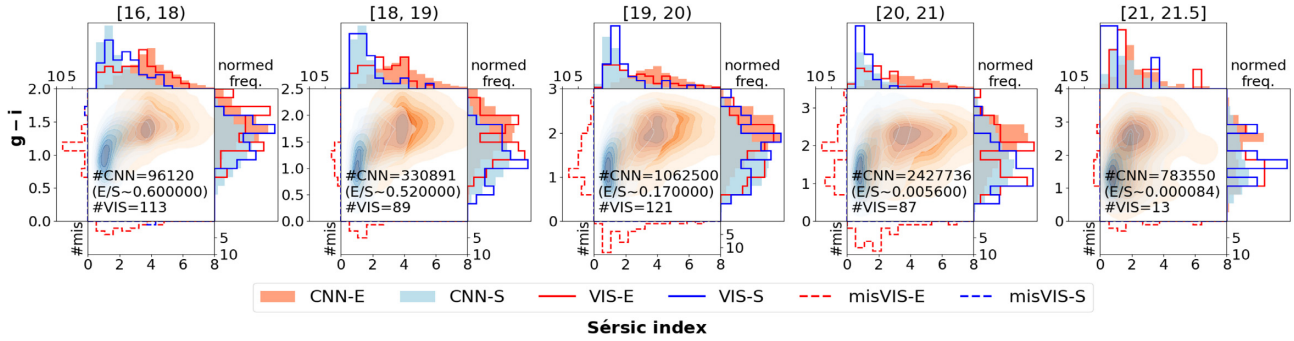

Figure 10. The diagram shows the colour and the Sérsic index distribution of our samples. The *y*-axis presents the colour $g - i$ and the *x*-axis shows the Sérsic index. The central contour plot shows the two-dimensional density distribution, while the histograms at the top and the right show the normalized frequency distribution of each quantity, Sérsic index and $g - i$ colour, respectively. The histograms at the bottom and the left show the samples with mismatched labels between the CNN and visual classifications for the Sérsic index and the colour, respectively. The red/orange colour represents the Ellipticals, while the bluish colour is for the spiral galaxies. The shadings represent the CNN predictions, and the solid lines show the distribution labelled by the visual classifications. Finally, the dashed lines show the misclassified samples with the labels from the visual classifications.

the expected Sérsic index distribution, we still find a fairly good separation in their colour distribution. This indicates that the CNN-classified Ellipticals with $21 \leq i \leq 21.5$ share some similarities among themselves. Therefore, this particular class of galaxies might have a different formation history from other Ellipticals, resulting in a relatively discy structure but redder colours. It would be interesting to test this hypothesis with multicolour data in the future.

Nonetheless, based on the analysis in this section, we exclude the CNN classifications in the magnitude range ($i \geq 21$) from our final catalogue due to the strong imbalance of the CNN classifications between the two types and the poor division in the colour–Sérsic diagram.

5.3 Further investigation into fainter galaxies

With the predicted probabilities provided by our CNN classifier, users can simply use these labels to allocate a classification with a higher predicted probability to a galaxy. However, as discussed above, the machine is trained with bright galaxies ($16 \leq i < 18$) at low redshift ($z < 0.25$), one might thus consider that the predictions for galaxies with similar properties as the training set are more robust. Therefore, in this section, we provide two additional quantities that can be used to select the CNN classifications that users might have more faith in based on different presumptions. We carry out the examination by further exploring our CNN predictions using Sérsic index and colour ($g - i$). In this section, we statistically assess our CNN classifications with a probability threshold of 0.5 for each magnitude and redshift range.

One way to determine the quality of our classifications is based on using only the Sérsic index, which is a fairly good indicator for galaxy structure even at high redshift. The predictions might be more robust if a CNN-classified Elliptical has a Sérsic index larger than 2.5 or a CNN-classified Spiral has a Sérsic index smaller than or equivalent to 2.5. Therefore, we statistically examine how well the CNN classifications do within a certain magnitude and redshift range to satisfy the corresponding Sérsic index as discussed above. Equation (5) is then used to provide 'frac_n' in Table 4, where N represents the total number of samples with Sérsic index measurements and n means the Sérsic index. The pE and pS are the predicted probabilities of being Ellipticals and Spirals by CNN, respectively.

$$frac\_n = \frac{N\left[(pE > 0.5) \wedge (n > 2.5)\right] + N\left[(pS > 0.5) \wedge (n \leq 2.5)\right]}{N} \quad (5)$$

The other quantity we consider is called 'confidence_flag' (Table 4) in this work, which is determined by comparing the distributions of both Sérsic index and colour ($g - i$) in each magnitude and redshift bin shown in Fig. 12 to the distributions of the reference samples. The confidence scheme is listed in Table 3. From the discussion in Section 5.1, CNN classifications for galaxies with $16 \leq i < 18$ and $z < 0.25$ have our higher confidence class – 'superior confidence'. In addition, they are the reference for the others. Note that this analysis is strongly based on a presumption that the machine has a better performance when classifying galaxies that are in a similar observed 'condition' (e.g. distance, magnitude, and size) to the training set. However, this notion may or may not be true, which needs more investigation to be confirmed, and will be the topic of a forthcoming paper.

In Fig. 12, we further carry out statistical analyses to determine the confidence level for galaxies with $18 \leq i < 21$ by subdividing the galaxies in each magnitude bin into 0.25-wide redshift bins. Galaxies are excluded from the catalogue if the number of galaxies with a given morphology type falls below 30 in a given bin since we do not have the necessary statistics to assess their reliability. The excluded galaxies are generally at the highest redshifts in their magnitude bins. Examples within each magnitude and redshift bin are shown in Fig. 11. Using our CNN classifications and Sérsic index information, we present examples of probable classes of Ellipticals (left) and Spirals (right) within different magnitude and redshift bins. Each row and column shows a range of magnitude and redshift, respectively.

Each row in Fig. 12 shows the diagrams within a given magnitude range, while each column presents them in a different redshift bin. We use the 'superior confidence' classifications, top-left diagram, as reference to assess the confidence level of other ranges; i.e. the closer the distribution is to the reference, a higher value in the confidence scheme is assigned. We note that colour may actually not be a good criterion at the higher redshifts for the reasons described above. However, we use colour as it is one of the criteria that separates galaxy types quite cleanly at the lowest redshifts for high-mass galaxies. What we require is specifically includes (1) a clear distinction in both quantities between the two galaxy types; (2) the peaks of the Sérsic index distribution must be at similar locations for both morphologies when comparing with the reference; i.e. the Sérsic index distributions should peak between 1 and 2 for Spirals and ∼4 for Ellipticals; (3) the median values of the Sérsic index for both types are similar to the one in the reference within 1σ (median absolute deviation); and (4) no unusual features should be apparent in any of the single distributions (e.g. no bimodal or messy distributions). In Table 3, a 'high confidence' is assigned when all the four criteria are satisfied. A 'confidence' label and a 'less confidence' label are given when one or two of the criteria are missing, respectively. When more than three criteria are not satisfied, 'no confidence' is allocated to the classifications for the galaxies within the corresponding magnitude and redshift ranges. Examples of the CNN classification with different confidence levels are shown in Fig. 13.

5.3.1 Magnitude bins: 16 ≤ i < 18

In Section 5.1, we established the excellent performance of our CNN predictions for galaxies in the same magnitude and redshift ranges as the training set ($16 \leq i < 18$ and $z < 0.25$, respectively). On the first column of the first row in Fig. 12, we show this robust conclusion again using a parametric morphology indicator, the Sérsic index, and a generic galaxy property – its colour. The distributions of both in this range are used as reference to determine the confidence level of other ranges. The median value of Sérsic index for Spirals and Ellipticals is 1.61 ± 0.60 and 3.75 ± 0.92, respectively, in this magnitude and redshift range.

We notice that the *frac_n* is relatively small in this region. In section 5.2, we discussed that our CNN classifies the class of Spirals mainly based on the presence of disc structure. Therefore, the small value of *frac_n* in this region is due to the constraint on Sérsic index in equation (5). The fraction of CNN-classified Spirals with Sérsic index between 2.5 and 4 is similar to the fraction of Ellipticals in this magnitude and redshift range. This indicates that the class of lenticular galaxies that is not well defined in our training set and has ambiguous structure could possibly confuse our CNN classifier (Cheng et al. 2021).

Next, we extend this examination to higher redshift but remain within the same magnitude range (second column at the first row in Fig. 12). A clear distinction between two CNN predicted types in the Sérsic index distribution can be seen within this redshift range, $0.25 \leq i < 0.5$, and the peaks of both types are located in a sensible region. However, the CNN-classified Spirals have a broader distribution compared with the reference sample such that the median value is 2.66. Additionally, their colour distribution shows an apparent overlap with the CNN-classified Ellipticals. This suggests two possibilities: (1) our CNN classifier is being less accurate within this range, and/or (2) there are a fair number of galaxies with the structural features of Spirals but which are red in colour, particularly within $g - i$. Overall, we label the CNN predictions within this range as 'less confidence'.

5.3.2 Magnitude bins: 18 ≤ i < 19

In the magnitude range $18 \leq i < 19$, we have three redshift bins that include more than 30 galaxies with morphological measurements within each type: $z < 0.25$, $0.25 \leq z < 0.5$, and $0.5 \leq z < 0.75$. In the first plot, we notice a good differentiation between the features of Ellipticals and Spirals such that the median value of the Sérsic index for Spirals and Ellipticals is 1.32 and 3.18, respectively. However, the peak of the Sérsic index distribution for Ellipticals is not located at ∼4. Therefore, we label the predictions of this range as those with 'confidence'.

The second diagram ($0.25 \leq z < 0.5$) has a slightly broader distribution of Sérsic indices for CNN-classified Spirals with a median value of 2.25 compared to the reference. Except for this, a distinguishable separation in Sérsic index distribution and colour distribution is presented. Hence, we recognize the CNN-classified labels in this range as 'confidence'.

Finally, the last panel shows an apparently overlapping colour distribution between Ellipticals and Spirals. In addition, a slightly bimodal structure is presented in the colour distribution. However, the median value of the Sérsic indices for Spirals and Ellipticals is 1.88 and 3.52, respectively, which is acceptable when comparing with the values of the reference. However, the peak of Sérsic index distribution for Spirals is off compared with the reference, i.e. not located between 1 and 2. We thus conservatively label the classifications in this range as 'no confidence'. However, we cannot discount that these classifications are reliable given the range and distributions in Sérsic indices.

5.3.3 Magnitude bins: 19 ≤ i < 20

In this fainter magnitude range, we observe an interesting result: A good consistency for our CNN predictions compared to the reference is found in the two higher redshifts bins, $0.25 \leq z < 0.5$ and $0.5 \leq z < 0.75$, than in the lower one. In these ranges, the median values of Sérsic indices for each morphology type are within the ranges provided by the reference, and the peaks of the Sérsic index distributions are reasonable. However, for the category of galaxies with redshifts $0.5 \leq z < 0.75$, a clearly bimodal distribution is presented. We therefore give the morphological classifications for galaxies in these redshift ranges a 'high confidence' and 'confidence' label, respectively.

The low redshift interval ($z < 0.25$; first column) shows a worse performance. We find a flat Sérsic index distribution for the CNN-classified Ellipticals that peaks at roughly $n \sim 2$ with a median value of 2.43. Additionally, although there is a separation in the colour distribution between the two types, the CNN-classified Ellipticals

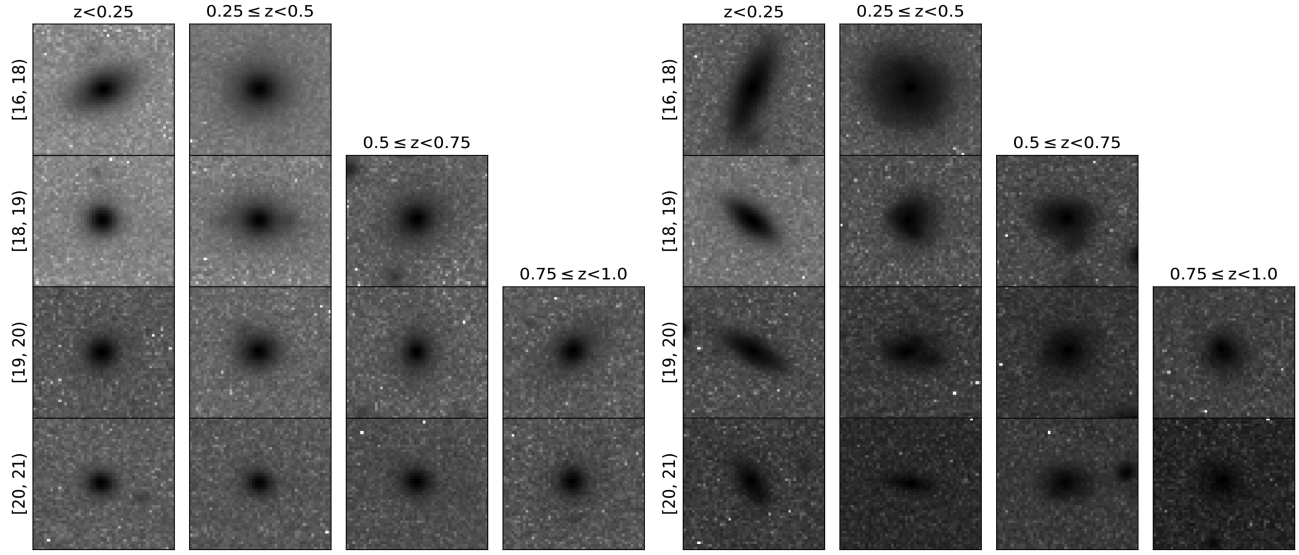

Figure 11. Examples within different magnitude and redshift bins. CNN-classified Ellipticals are shown at the left side while the right side showcases Spirals (disc galaxies). Each row and column represents a range of magnitude and redshift, respectively.

show a bimodal colour distribution that partially overlaps with the CNN-classified Spirals. Although the performance for Ellipticals in this redshift range is clearly worse, the behaviour for Spirals is significantly better: There is a fairly good discrimination in both the Sérsic index and the colour distributions. This means that in this redshift range, our CNN-classified spiral sample has a high purity but not a high completeness. We therefore label the classifications made in this range as 'less confidence' but with a '$*$' mark (Table 3). The '$*$' indicates that this confidence level is only defined for CNN-classified Spirals, and the classified Ellipticals are labelled as 'no confidence'. Clearly, this sample cannot be used to find all Spirals, but we do have some confidence in the morphologies for the ones it does classify. In addition, there are a much larger number of CNN-classified Spirals than Ellipticals; therefore, the high *frac_n* in this region supports the confidence assignment.

It seems counter-intuitive that a better performance is found for higher redshift galaxies than for lower redshift ones at these faint magnitudes. However, the reason is that the fainter galaxies in the training set tend to be low-luminosity galaxies or are systems at higher redshifts. Therefore, there is a somewhat better overlap in the properties of faint higher redshift galaxies than there is for faint lower redshift ones between the general DES Y3 sample and the training set. This issue is also discussed in computational science as an interesting issue called 'The Elephant in the Room (Rosenfeld, Zemel & Tsotsos 2018)'. They proposed one of the reasons for this situation – '*Out of Distribution Examples*'. In our case, we interpret the distribution presented in faint lower redshift galaxies as less likely to occur under our distributions of training sets.

Finally, for this magnitude range we give a 'no confidence' label to the highest redshift range ($0.75 \leq z < 1.0$). This is due to the messy galaxy property distributions reflected in the bimodal colour distributions for both morphological types, a significantly higher Sérsic index than expected for the CNN-classified Ellipticals, and a relatively low Sérsic index for the CNN-classified Spirals. Interestingly, despite the relatively anomalous Sérsic index, a fairly sharp differentiation between both types is shown in the Sérsic index distributions. For the CNN-classified Ellipticals, this suggests a class of red galaxies that has a higher concentration and a more peaked surface brightness distribution than expected. This is an interesting conclusion from our CNN classification analysis that deserves to be explored further in future work.

5.3.4 Magnitude bins: $20 \leq$ i < 21

As we get to fainter magnitudes, using our CNN methodology to classify galaxies becomes more of a challenge. From Figs 9 and 10, we notice that there are also significantly fewer galaxies classified as Ellipticals by our CNN set up in this range, such that the CNN-classified *E/S* ratio is ∼0.0030 for total samples and ∼0.0056 for overlapping samples with morphological measurements, while the ones in other brighter ranges have a ratio over 0.1. This indicates that the bias self-correction by our CNN classifier might be overdone in this range compared to the brighter ranges. However, unlike the result shown in the range $21 \leq i \leq 21.5$ in Fig. 10, a better and clearer separation between both types in Sérsic index and colour is presented. Hence, we carry out a further investigation within different redshift bins for this range.

In the first plot on the bottom row in Fig. 12, the distributions of CNN-classified Spirals are fairly reasonable with a median value of 1.04. However, a differing peak assignment of the Sérsic index occurs within the CNN-classified Ellipticals. Therefore, we decided to assign a class of 'less confidence' with '$*$' for this range, where '$*$' means that this confidence label is for the Spirals (no confidence to the classification of Ellipticals). The *frac_n* reflects the same support as the plot above within $19 \leq i < 20$.

The second plot for this magnitude range in Fig. 12 shows a good separation between the two types of galaxies in Sérsic index and colour space. Although a strong imbalance in the number of Ellipticals and Spirals still exists here, the differentiation in two types proves a certain degree of confidence to our CNN predictions. The median value of the Sérsic index for Spirals and Ellipticals is 1.22 and 3.31, respectively, in this redshift range. However, the peak of the Sérsic index distribution for the Ellipticals is off compared with the reference. Hence, we label the predictions in this range as 'confidence'. The reason for this good separation in this significantly fainter magnitude range is also due to the effect discussed in Section 5.3.3 that the galaxies in this magnitude (20

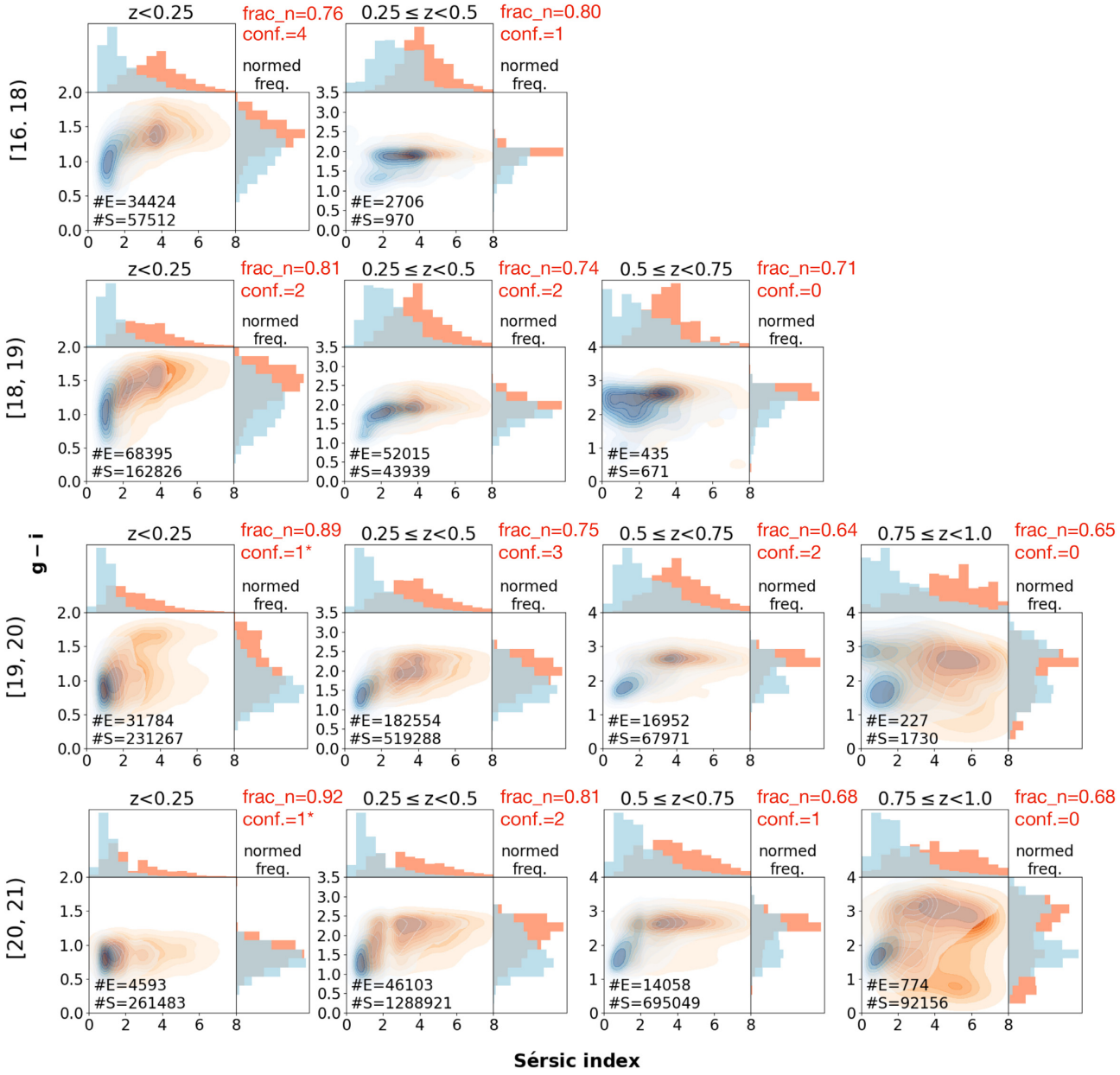

Figure 12. The colour–Sérsic diagrams of different redshift bins for each of the magnitude ranges. The histograms at the top and the right of each diagram show the normalized frequency distribution of Sérsic index and colour $g - i$, respectively. The red shading represents the Ellipticals classified by our CNN with a probability threshold of 0.5, while the blue shading shows the CNN-classified Spirals with a probability threshold of 0.5. The magnitude range is shown at the left of each row while the redshift range is presented above each graph. The textual information in the diagrams shows the number of Ellipticals (E) and Spirals (S) classified by our CNN and with the DES Y1 morphological measurements from Tarsitano et al. (2018). The red text at the top right corner of each plot indicates (1) the fraction of CNN classifications that satisfies Sérsic index criteria (frac_n; equation 5) and (2) the confidence level (conf.) for each magnitude and redshift bin.

$\leq i < 21$) and redshift ($0.25 \leq z < 0.5$) range have similar galaxy features and galaxy properties to the reference samples. The shift in magnitude for these galaxies is due to the change in redshifts.

This situation is also demonstrated within the third plot of Fig. 12 ($20 \leq i < 21$ and $0.5 \leq z < 0.75$) whereby both types are distinguished in Sérsic index distribution with a median value of 1.78 and 3.77 for Spirals and Ellipticals, respectively. However, CNN-classified Spirals have a relatively flat colour distribution that shows an indication of a small bimodal distribution and prevents a clear separation. Hence, a class of 'less confidence' is assigned to this range. Finally, the last diagram shows a messy distribution. Therefore, we simply label this range as a 'no confidence' class.

5.4 VIPERS spectral classification

After Sections 5.2 and 5.3, we finalize the number of galaxy classifications in our final catalogue. In this section, we compare our CNN predictions with the spectral classification from VIPERS presented in Siudek et al. (2018, Section 4.2). The number of overlapping samples between their spectral classification catalogue and our final catalogue is 10 254, of which 9459 galaxies have a

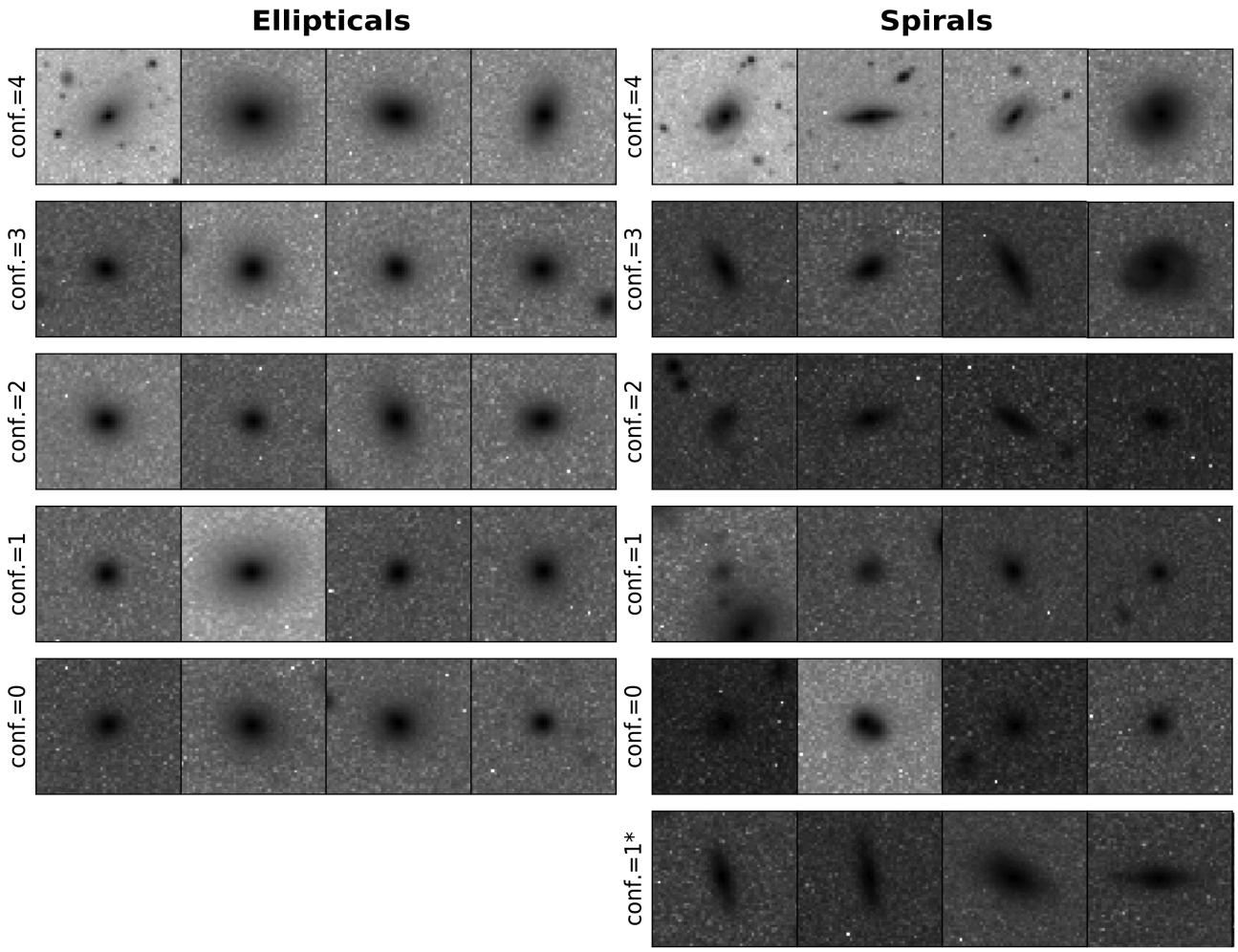

Figure 13. Examples of each confidence level (see Table 3) for the prediction of the two morphological types. The top row represents the most confident classification. CNN-classified Ellipticals are shown on the left-hand side, while the right-hand panel presents CNN-classified Spirals. Confidence '1∗' is for Spirals only as explained in Table 3.

high class membership probability in their catalogue. This is enough galaxies to test how our classifications agree with those based on spectroscopy.

Three main classes of spectral-types for galaxies, namely passive (P), intermediate (I), star forming (SF), are defined in this catalogue that we use to examine our CNN classifications. Fig. 14 shows that Ellipticals labelled by our CNN are mostly passive, such that we find that fractions of 0.75, 0.81, and 0.82 are passive CNN-classified Ellipticals from $p = 0.5$ and 0.8 to $p = 0.8$ with conf. $\geq$2, but CNN-labelled Spirals show a mixture of three classes. We further examine the Sérsic index distribution of the CNN-labelled Spirals in the last panel of Fig. 14. This figure shows that the CNN-labelled Spirals at the passive and intermediate spectral stages are mostly discy, i.e. those galaxies with Sérsic index distribution at $<$4. The fraction of intermediate CNN-labelled Spirals with Sérsic indices smaller than 4 is 0.87, 0.88, and 0.93 ($<$3 is 0.75, 0.76, and 0.85) of the total sample in each row, from $p = 0.5$ and 0.8 to $p = 0.8$ with conf. $\geq$2, while the fraction of passive CNN-labelled Spirals is 0.60, 0.63, and 0.57 ($<$3 is 0.42, 0.44, and 0.38).

Since galaxies with fainter magnitudes and at higher redshift might not have a clear visual spiral arm (Fig. 11), in addition to the possibility of passive Spirals (Masters et al. 2010), our CNN is likely to classify passive disc galaxies, such as lenticulars, into the class of Spirals.

5.5 Non-parametric methods and galaxy properties

Another examination is carried out using non-parametric methods such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), Gini coefficient, and M20. In this study, the non-parametric measurements are from Tarsitano et al. (2018) using the *i*-band images, and we use the measurements after applying the selection criteria described in Section 4.3.

Furthermore, this validation can work in both directions. We can use the non-parametric measurements to check the robustness of our CNN-based morphological classifications, while also use our most reliable morphological classifications (those with 'superior confidence') to assess the ability of non-parametric methods to separate the Ellipticals and the Spirals (Fig. 15). Such an analysis of non-parametric measurements as proxies for morphology has never before been carried out with samples as large as ours. In this work, we include over 100 000 galaxies in the 'superior confidence' category from our DES Y3 morphological classifications.

In Fig. 15, we show the pair plots of six different parameters: concentration (C), asymmetry (A), clumpiness (S), Gini, M20, and

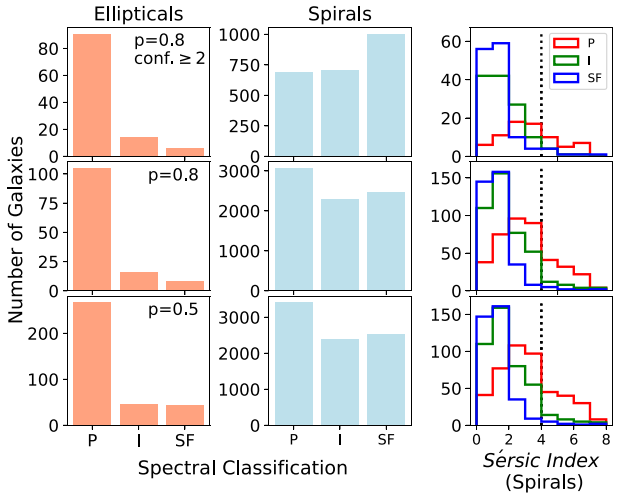

Figure 14. Comparison between our CNN predictions and the VIPER unsupervised spectral classification. From the bottom to the top row, a probability threshold of 0.5, 0.8, and 0.8 with the confidence level greater than 2 is applied in each row. The P, I, and SF at the *x*-axis of the first two panels represent the spectral classifications of passive, intermediate, and star-forming galaxies, respectively. The last panel shows the Sérsic index distribution of Spirals labelled by our CNN. Different colours, red, green, and blue, represent galaxies with different spectral classifications, passive (P), intermediate (I), and star forming (SF), respectively. The vertical dotted line indicates where the Sérsic index equals to 4.

Sérsic index. For the A, S parameters, we only showcase the data with values smaller than 0.2 to focus on 'typical galaxies'. The Sérsic index is used as a comparison to the non-parametric methods, and it is one of the main features used to define the confidence level (Section 5.3). It shows a clear separation between the two morphological types here. In addition to this, we note that only the Gini coefficient shows a consistently distinguished difference between the two types in the histogram.

The Gini coefficient (G) reflects the inequality of the flux distributed among the pixels of a given galaxy; if $G = 1$, the light is concentrated in one pixel, while conversely, $G = 0$ means that the light is uniformly distributed to every pixel. Therefore, the Gini coefficient is somewhat analogous to the concept of concentration, and Ellipticals generally have a higher value than Spirals. Nevertheless, the concentration does not show a separation as good as the one for the Gini coefficient. A slight shift between the peaks of the two morphological types is shown in the histogram of the concentration; however, a large overlapping area is also shown. Additionally, the difference of the mean concentration values between both types is relatively small compared with previous studies (Conselice 2003; Hernández-Toledo et al. 2008; Hambleton et al. 2011). On the other hand, both asymmetry and clumpiness also fail to show a consistent distinction between the two morphological types in our analysis.

Finally, the M20 histogram does not show a clear separation between the two morphological types either. However, a clean separation does show itself in the contour of the Gini coefficient and M20. The black dashed line indicates a cut used to separate Ellipticals and Spirals and described in Lotz et al. (2008) such that

$$G = 0.14M_{20} + 0.8. \quad (6)$$

Thus, we find that the Gini coefficient is a possible better tracer of the overall structure of a galaxy than any other non-parametric morphological quantities such as C, A, S, and M20 (Zamojski et al. 2007) when separating Ellipticals from Spirals.

6 GALAXY MORPHOLOGICAL CLASSIFICATION CATALOGUE

In this paper, with the CNN trained with the subset of the DES Y1 data with the GZ1 labels corrected in C20 (Sections 2.2 and 2.3), we provide one of the largest catalogues to date with galaxy morphological classifications for over 20 million galaxies from the DES Y3 data ($16 \leq i < 21$ and $z < 1.0$; Section 2.4; along with the companion catalogue produced by Vega-Ferrero et al. 2021). As mentioned in Section 1, an extensive comparison of the these two catalogues is ongoing, the result of which will be published in a future paper.

The items provided in our catalogue of morphological types are listed in Table 4. The average predicted probabilities from the five individual CNN models (Section 3) are used as the final probabilities of being Ellipticals (pE) and Spirals (pS). With this quantity, users can apply a probability threshold, which determines the tolerance of the accuracy of the morphological classification, to fit with their scientific goals. In this catalogue, we provide the classification label based on a threshold of 0.8 (*MORPH_FLAG*) for the user's convenience.

Our CNN classifier is trained with bright galaxies (magnitudes $16 \leq i < 18$) at low redshifts ($z < 0.25$). Therefore, for users who are more comfortable with our machine's predictions when applied to galaxies with similar condition to the training set, we carried out a statistical analysis (Section 5.3) to investigate the impact when the target data have a worse image quality than the training set due to faintness and redshift effects. Within this analysis, we provide a confidence flag for every galaxy (Table 3) within our CNN classification final catalogue. In addition, another flag, *frac_n*, which is defined as the fraction of predictions that satisfy the Sérsic index criteria (equation 5), serves a similar purpose. Overall, over 20 million CNN classifications with an assigned confidence level are included in our final catalogue, of which ∼670 000 galaxies have a 'superior confidence', ∼3.4 millions of galaxies are assigned as a 'high confidence' classification, and ∼9 million galaxies have a 'confidence' label. Finally, in columns 9 and 10 in Table 4 we provide magnitude and redshift information directly from the DES Y3 GOLD catalogue to allow customized magnitude/redshift cut when applying our predictions.

7 SUMMARY

We present in this paper one of the largest galaxy morphological classification catalogues produced to date (along with the other DES catalogue presented in Vega-Ferrero et al. (2021), using the DES Y3 data with over 20 million galaxies. We carry out these classifications using CNNs trained with the subset of the DES Y1 data. The *corrected* debiased labels, which are initially from the GZ1 catalogue and corrected in C20, are used to label our training set (Section 2.2). With a combination of three different types of inputs, including linear images, log images, and HOG images (Section 2.1), our CNN classifier reaches an accuracy of over 99 per cent when compared with the GZ1 classifications (i-band magnitude <18 and redshift <0.25). The majority of mismatches occur in the case when a galaxy is classified as a Spiral by our CNN but as Ellipticals by GZ1. The reason behind this mismatch is likely to be the better resolution and deeper depth of the DES imaging data, which reveals unnoticeable structure in the data used in GZ1 from the SDSS (more discussion in Cheng et al. 2020a). Additionally, training with the *corrected* debiased labels, our CNN classifier is shown to be self-debiased and more accurate in classifying disc galaxies, while human

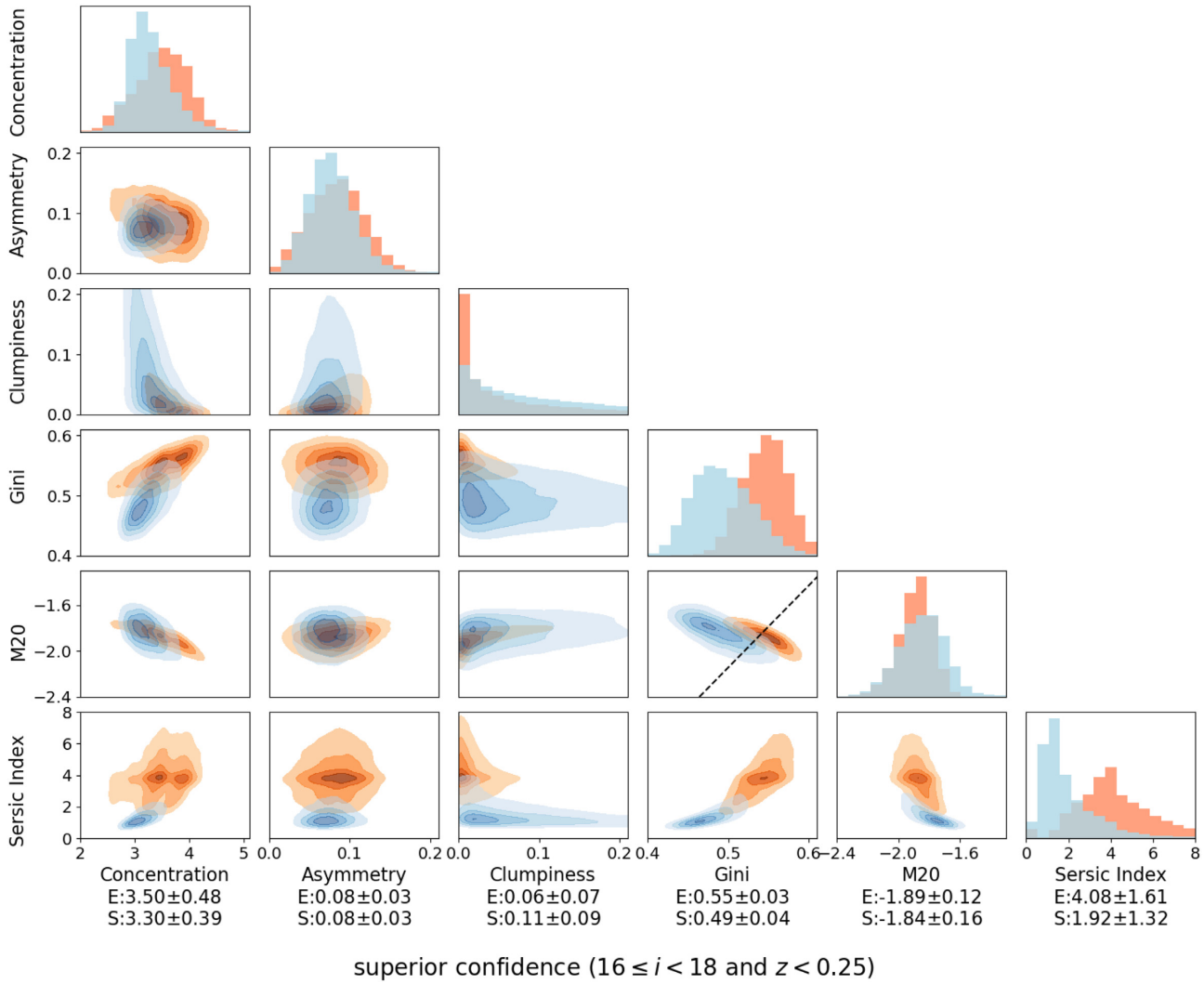

Figure 15. The pair plots of six morphological parameters: concentration, asymmetry, clumpiness, Gini, M20, and Sérsic index labelled by the CNN classifications with 'superior confidence'. The colour shadings represent the CNN classifications. The red/orange and blue colour are for Ellipticals (E) and Spirals (S), respectively. The mean value of each parameter for both types with the standard deviation is shown below each column. The black dashed line shows a cut from Lotz et al. (2008) to separate Ellipticals and Spirals based on the M20 and the Gini coefficients.

visual classifications have difficulty detecting at faint magnitudes down to $i \sim 21$ (see Section 5.2).

Trained with bright galaxies at low redshift, our CNN classifier is statistically assessed for its performance when used to predict morphologies for fainter galaxies at higher redshift. This assessment provides an investigation about how well a machine trained within one domain can be applied to the conditions in different domains; in our case, we applied the machine trained with bright galaxies ($i < 18$) at low redshift ($z < 0.25$) to fainter galaxies ($i \geq 18$). Using a cross-validation with the Sérsic index and galaxy colour ($g - i$), we provide a confidence evaluation scheme to our CNN classifications (Table 3) through a statistical analysis of data in different magnitude and redshift bins (Section 5.3). We define six confidence levels by comparing with the Sérsic indices and colour distributions of the data within the same coverage as the training set. In this assessment, we find that a better confidence is assigned to faint galaxies at higher redshift compared to galaxies with fainter magnitudes, but at lower redshift. For example, the confidence of predictions for galaxies with $19 \leq i < 20$ at $0.25 \leq z < 0.5$ is higher than the one at $z < 0.25$ at the same magnitude range. Faint galaxies in the training set are generally at higher redshift. A faint galaxy at relatively low redshift is an anomaly, in the sense that they do not exist in our training domain, in the machine's view. Thus, the machine gives a better prediction for fainter galaxies at higher redshift than systems at lower redshift, even though the magnitude and redshift ranges of these galaxies are beyond the ranges of the training set. Finally, we conclude that over 13 million galaxies (over 60 per cent of the total classifications) have at least a 'confidence' level as defined in our work.

As a part of the validation, we carry out a large examination of non-parametric methods such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), the Gini coefficient, and M20 using over 100 000 classifications with structural measurements from Tarsitano et al. (2018). From this, we conclude that the Gini coefficient shows the most significant distinction, as a single parameter, between Ellipticals and Spirals within all parameters tested. Additionally, with a combination of the M20 index, a straight line (Lotz et al. 2008) can be drawn to separate these two types (Fig. 15).

In addition, we compare our CNN predictions with spectral classification from VIPERS presented in Siudek et al. (2018). The result shows that the CNN-classified Ellipticals are mostly passive with a passive fraction of over 0.75. On the other hand, the CNN-classified Spirals show a mixture of passive, intermediate, and star-forming classes, but the majority have disc-like structures (Sérsic index <3). In addition to the possibility of passive Spirals, lenticulars are also responsible for the fraction of passive CNN-labelled Spirals in our case.

In this work, we used only observed data (bright galaxies at low redshift) to train our CNN, which limits potential applications to very faint galaxies. However, through the analysis carried out in this work, we notice that our machine classifies discy galaxies with round and blurred structure to the class of Spirals, while humans usually misclassify these systems as Ellipticals (Section 5.2). This supports the usefulness of our machine classification for fainter galaxies. Users can straightly utilize the predicted probabilities (*pE* and *pS* in Table 4) to obtain galaxy morphology predictions. The *MORPH_FLAG* provided in Table 4 uses a probability threshold of 0.8 to define Spirals (1) and Ellipticals (0) for users' convenience.

Our new morphological catalogue allows a variety of new approaches towards understanding galaxy properties and evolution that involve morphology that could not be carried out before. For example, non-parametric analysis methods of galaxy structure can be assessed using an unprecedented sample not only in size but also in quality. Our catalogue can also be used to cross-validate other classification methods, and to explore galaxy properties and environment as a function of morphology with superb statistics. Future papers will examine these features of the galaxy population and galaxy evolution with morphology using our classifications.

Scientifically, there are of course a myriad of other uses for our catalogue, as morphology is one of the fundamental properties of galaxies. For the time being, this will remain one of the largest sets of morphological classifications available for analysis for any survey done to date (along with the companion classification catalogue produced by Vega-Ferrero et al. 2021). Our methodology is also scalable and meant to be of use for applications to future imaging data sets such as the ones that will eventually be created from the *Euclid Space Telescope* and the Vera Rubin Observatory Legacy Survey of Space and Time, among others.

ACKNOWLEDGEMENTS

TYC acknowledges the support of the Vice-Chancellor's Scholarship from the University of Nottingham and STFC grants ST/T000244/1 and ST/P000541/1 at the Durham University. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovaçâo, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions in the DES.

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

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

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

DATA AVAILABILITY

This DES Y3 morphological classification catalogue is currently not publicly available but can be shared on request to the corresponding author. It will be available soon in the Dark Energy Survey Data Management (DESDM) system.

REFERENCES

Abbott T. M. C. et al., 2018, ApJS, 239, 18
Abraham R. G., van den Bergh S., Nair P., 2003, ApJ, 588, 218
Aihara H. et al., 2018, PASJ, 70, S4
Amirshahi S. A., Pedersen M., Yu S. X., 2016, J. Imaging Sci. Technol., 60, 604101
Avestruz C., Li N., Zhu H., Lightman M., Collett T. E., Luo W., 2019, ApJ, 877, 58
Baillard A. et al., 2011, A&A, 532, A74
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Ball N. M., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R. J., 2004, MNRAS, 348, 1038
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Banerji M. et al., 2010, MNRAS, 406, 342
Beck M. R. et al., 2018, MNRAS, 476, 5516
Bershady M. A., 1995, AJ, 109, 87

Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bishop C. M., 2006, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin
Bottrell C. et al., 2019, MNRAS, 490, 5390
Cassata P. et al., 2005, MNRAS, 357, 903
Cheng T.-Y. et al., 2020a, MNRAS, 493, 4209
Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020b, MNRAS, 494, 3750
Cheng T.-Y., Huertas-Company M., Conselice C. J., Aragón-Salamanca A., Robertson B. E., Ramachandra N., 2021, MNRAS, 503, 4446
Conselice C. J., 2003, ApJS, 147, 1
Conselice C. J., Blackburne J. A., Papovich C., 2005, ApJ, 620, 564
Dalal N., Triggs B., 2005, 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR'05), Vol. 1, IEEE, San Diego, CA, USA, p. 886
de Vaucouleurs G., 1948, Ann. Astrophys., 11, 247
de Vaucouleurs G., 1959, Handbuch Phys., 53, 275
de Vaucouleurs G., 1964, AJ, 69, 561
DES Collaboration, 2005, preprint (astro-ph/0510346)
DES Collaboration, 2016, MNRAS, 460, 1270
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Dodge S., Karam L., 2016, 2016 8th Int. Conf. Qual. Multimedia Exper., QoMEX 2016. Inst. Electr. Electron. Eng. Inc., Lisbon, Portugal
Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, MNRAS, 476, 3661
Drlica-Wagner A. et al., 2018, ApJS, 235, 33
Dubath P. et al., 2011, MNRAS, 414, 2602
Everett S. et al., 2020, preprint (arXiv:2012.12825)
Fawcett T., 2006, Pattern Recognit. Lett., 27, 861
Ferreira L., Conselice C. J., Duncan K., Cheng T.-Y., Griffiths A., Whitney A., 2020, ApJ, 895, 115
Flaugher B. et al., 2015, AJ, 150, 150
Fukugita M. et al., 2007, AJ, 134, 579
Fukushima K., 1975, Biol. Cybern., 20, 121
Fukushima K., 1980, Biol. Cybern., 36, 193
Fukushima K., Miyake S., Ito T., 1983, IEEE Trans. Syst. Man Cybern., 13, 826
Ghosh A., Urry C. M., Wang Z., Schawinski K., Turp D., Powell M. C., 2020, ApJ, 895, 112
Hambleton K. M., Gibson B. K., Brook C. B., Stinson G. S., Conselice C. J., Bailin J., Couchman H., Wadsley J., 2011, MNRAS, 418, 801
Hausen R., Robertson B. E., 2020, ApJS, 248, 20
Hernández-Toledo H. M., Vázquez-Mata J. A., Martínez-Vázquez L. A., Avila Reese V., Méndez-Hernández H., Ortega-Esbrí S., Núñez J. P. M., 2008, AJ, 136, 2115
Hubble E. P., 1926, ApJ, 64, 321
Huertas-Company M., Rouan D., Tasca L., Soucail G., Le Fèvre O., 2008, A&A, 478, 971
Huertas-Company M. et al., 2009, A&A, 497, 743
Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, A&A, 525, A157
Huertas-Company M. et al., 2015, ApJS, 221, 8
Huertas-Company M. et al., 2018, ApJ, 858, 114
Jacobs C., Glazebrook K., Collett T., More A., McCarthy C., 2017, MNRAS, 471, 167
Kamble P. M., Hegadi R. S., 2015, Procedia Comput. Sci., 45, 266
Karahan S., Kilinc Yildirum M., Kirtac K., Rende F. S., Butun G., Ekenel H. K., 2016, 2016 Int. Conf. Biometrics Spec. Interest Group (BIOSIG). IEEE, Darmstadt, Germany, p. 1
Kennicutt R. C., Jr, 1998, ARA&A, 36, 189
Lahav O., Naim A., Sodré L. J., Storrie-Lombardi M. C., 1996, MNRAS, 283, 207
Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, MNRAS, 473, 3895
Law D. R., Steidel C. C., Erb D. K., Pettini M., Reddy N. A., Shapley A. E., Adelberger K. L., Simenc D. J., 2007, ApJ, 656, 1
Lecun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lintott C. et al., 2011, MNRAS, 410, 166
Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163
Lotz J. M. et al., 2008, ApJ, 672, 177
Maehoenen P. H., Hakala P. J., 1995, ApJ, 452, L77
Masters K. L. et al., 2010, MNRAS, 405, 783
Miller A. A., Kulkarni M. K., Cao Y., Laher R. R., Masci F. J., Surace J. A., 2017, AJ, 153, 73
Morgan W. W., Mayall N. U., 1957, PASP, 69, 291
Naim A., Lahav O., Sodre L. J., Storrie-Lombardi M. C., 1995, MNRAS, 275, 567
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
Nair V., Hinton G. E., 2010, Proc. 27th Int. Conf. Mach. Learn., ICML'10. Omnipress, Haifa, Israel, p. 807
Neilsen E. H., Jr, Annis J. T., Diehl H. T., Swanson M. E. C., D'Andrea C., Kent S., Drlica-Wagner A., 2019, preprint (arXiv:1912.06254)
Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A., 1992, AJ, 103, 318
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2010, AJ, 139, 2097
Petrillo C. E. et al., 2017, MNRAS, 472, 1129
Polsterer K. L., Gieseke F., Kramer O., 2012, Astronomical Society of the Pacific Conference Series, 462, 561
Powers D. M. W., 2011, J. Mach. Learn. Technol., 2, 37
Prakash C. D., Karam L. J., 2019, preprint (arXiv:1912.01707)
Rosenfeld A., Zemel R., Tsotsos J. K., 2018, preprint (arXiv:1808.03305)
Sandage A., 1961, The Hubble Atlas of Galaxies. Carnegie Institution of Washington publication, Washington, DC, USA
Scarlata C. et al., 2007a, ApJS, 172, 406
Scarlata C. et al., 2007b, ApJS, 172, 494
Sérsic J. L., 1963, Bol. Asociacion Argentina Astron. La Plata Argentina, 6, 41
Sérsic J. L., 1968, Atlas de Galaxias Australes. Cordoba, Argentina: Observatorio Astronomico, Cordoba, Argentina
Sevilla-Noarbe I. et al., 2020, preprint (arXiv:2011.03407)
Shamir L., 2009, MNRAS, 399, 1367
Shu C., Ding X., Fang C., 2011, Tsinghua Sci. Technol., 16, 216
Siudek M. et al., 2018, A&A, 617, A70
Soler J. D. et al., 2019, A&A, 622, A166
Sreejith S. et al., 2018, MNRAS, 474, 5232
Tarsitano F. et al., 2018, MNRAS, 481, 2018
Vega-Ferrero J. et al., 2021, MNRAS, 506, 1927
Walmsley M. et al., 2020, MNRAS, 491, 1554
Weir N., Fayyad U. M., Djorgovski S., 1995, AJ, 109, 2401
Willett K. W. et al., 2013, MNRAS, 435, 2835
Zamojski M. A. et al., 2007, ApJS, 172, 468
Zaritsky D., Zabludoff A. I., Willick J. A., 1995, AJ, 110, 1602
Zhou Y., Song S., Cheung N., 2017, 2017 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). IEEE, New Orleans, LA, USA, p. 1213

[1]*Centre of Extragalactic Astronomy, Durham University, Stockton Road, Durham DH1 3LE, UK*
[2]*School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK*
[3]*Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK*
[4]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP 05314-970, Brazil*
[5]*Laboratório Interinstitucional de e-Astronomia – LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil*
[6]*Fermi National Accelerator Laboratory, PO Box 500, Batavia, IL 60510, USA*
[7]*Instituto de Física Teórica, Universidade Estadual Paulista, São Paulo, 01140-070, Brazil*
[8]*Cavendish Laboratory Astrophysics Group, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[9]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[10]*Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

[11]*Kavli Institute for Particle Astrophysics & Cosmology, Stanford University, PO Box 2450, Stanford, CA 94305, USA*
[12]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[13]*Center for Astrophysical Surveys, National Center for Supercomputing Applications, 1205 West Clark Street, Urbana, IL 61801, USA*
[14]*Department of Astronomy, University of Illinois at Urbana–Champaign, 1002 W. Green Street, Urbana, IL 61801, USA*
[15]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona), Spain*
[16]*Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[17]*Astronomy Unit, Department of Physics, University of Trieste, Via Tiepolo 11, I-34131 Trieste, Italy*
[18]*INAF – Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[19]*Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34014 Trieste, Italy*
[20]*Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil*
[21]*Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[22]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, 28040, Spain*
[23]*Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[24]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[25]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[26]*Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
[27]*Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[28]*Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029 Blindern, NO-0315 Oslo, Norway*
[29]*Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain*
[30]*Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain*
[31]*Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, E-28049 Madrid, Spain*
[32]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[33]*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[34]*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
[35]*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
[36]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[37]*Department of Astronomy/Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA*
[38]*Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia*
[39]*Lowell Observatory, 1400 Mars Hill Road, Flagstaff, AZ 86001, USA*
[40]*Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
[41]*Physics Department, University of Wisconsin–Madison, 2320 Chamberlin Hall, 1150 University Avenue Madison, WI 53706-1390, USA*
[42]*Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA*
[43]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[44]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[45]*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK*

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