Improving galaxy morphologies for SDSS with Deep Learning

H Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D Tuccillo, J.L. Fischer

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **476,** 3661–3676 (2018)
Advance Access publication 2018 February 9

doi:10.1093/mnras/sty338

Improving galaxy morphologies for SDSS with Deep Learning

H. Domínguez Sánchez,[1,2]★ M. Huertas-Company,[1,2,3] M. Bernardi,[1] D. Tuccillo[2,4] and J. L. Fischer[1]

[1]*Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104, USA*
[2]*LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, F-75014 Paris, France*
[3]*University of Paris Denis Diderot, University of Paris Sorbonne Cité (PSC), F-75205 Paris, Cedex 13, France*
[4]*Mines ParisTech, 35 rue Saint Honoré, F-77305 Fontainebleau Cedex, France*

Accepted 2018 January 31. Received 2017 December 30; in original form 2017 September 7

ABSTRACT

We present a morphological catalogue for ∼670 000 galaxies in the Sloan Digital Sky Survey in two flavours: T-type, related to the Hubble sequence, and Galaxy Zoo 2 (GZ2 hereafter) classification scheme. By combining accurate existing visual classification catalogues with machine learning, we provide the largest and most accurate morphological catalogue up to date. The classifications are obtained with Deep Learning algorithms using Convolutional Neural Networks (CNNs). We use two visual classification catalogues, GZ2 and Nair & Abraham (2010), for training CNNs with colour images in order to obtain T-types and a series of GZ2 type questions (disc/features, edge-on galaxies, bar signature, bulge prominence, roundness, and mergers). We also provide an additional probability enabling a separation between pure elliptical (E) from S0, where the T-type model is not so efficient. For the T-type, our results show smaller offset and scatter than previous models trained with support vector machines. For the GZ2 type questions, our models have large accuracy (>97 per cent), precision and recall values (>90 per cent), when applied to a test sample with the same characteristics as the one used for training. The catalogue is publicly released with the paper.

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

1 INTRODUCTION

Since the beginning of the last century, it is well known that galaxies exhibit a wide variety of morphologies. The first classification was done by Hubble (1926, 1936), dividing the galaxies into two broad types: galaxies with a dominant bulge component (also known as early-type galaxies, ETGs) and galaxies with a significant disc component (late-type or spiral galaxies). The spiral galaxies are further divided into barred (with the presence of a bar shaped central structure) or unbarred, and ordered according to their spiral arms strength. The intermediate type between elliptical and spiral galaxies are called S0, while there is also a population of galaxies with irregular or distorted shapes. According to this visual classification, a number can be assigned to each type of galaxy, which is known as the T-type (de Vaucouleurs 1963).

Interestingly, morphology is very closely related to the stellar properties of the galaxies: in the local Universe most elliptical galaxies show redder colours, larger masses, higher velocity dispersions, and older stellar populations than spiral galaxies, which are mostly gas rich star-forming systems with high rotation velocities (e.g. Roberts & Haynes 1994; Blanton & Moustakas 2009; Pozzetti et al. 2010 and references therein). It is also well known that both the structural and intrinsic properties of galaxies undergo a significant evolution across cosmic time (e.g. Wuyts et al. 2011, Huertas-Company et al. 2013, Huertas-Company et al. 2015, Barro et al. 2017). Understanding how morphology relates to all these other properties and in which way they affect galaxy assembly is one of the major challenges of present-day astronomy.

It is, therefore, crucial to have accurate galaxy morphological classifications for large samples. Morphological classification has traditionally been done by eye. However, this presents two major problems: first, it is not obvious how to categorize galaxies into one of each subclass, since there is a smooth transition between each T-type. This effect is even more evident at high redshift where, in addition to the poorer image quality, important structural changes, and transitions between morphological types are taking place (e.g. Huertas-Company et al. 2015). Secondly, visual classification is an incredible time-consuming task. This is an enormous disadvantage in the era of big data, when extremely large surveys (such as SDSS, Eisenstein et al. 2011, Dark energy Survey, Dark Energy Survey Collaboration et al. 2016 or EUCLID, Racca et al. 2016) release images for millions of galaxies. Visual classification does become a real impossible task.

One smart way to overcome the problem of visual classification for large amounts of data was the Galaxy Zoo project[1] (Linott

★ E-mail: helenado@sas.upenn.edu

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

et al. 2008), where 'science citizens' volunteered to classify galaxies through a user friendly web interface. The first approach was a very simple classification into three types (ETGs, spirals or mergers) but, given the success of the project, a more complex classification system, GZ2, was proposed in Willett et al. 2013. However, galaxy classifications made by amateur astronomers, which is a difficult task even for professionals, has its caveats. For example, features such as bars are only selected when the bar is obvious and the volunteers tend to choose intermediate options when available (e.g. prominence of bulge, roundness, etc.). There are also a large number of galaxies with uncertain classifications caused by the disagreement between classifiers.

Automated classifications using a set of parameters that correlate with morphologies (e.g. concentrations, clumpiness, asymmetries, Gini coefficients, etc.) have also been attempted (Abraham et al. 1996; Conselice, Bershady & Jangren 2000; Lotz et al. 2008). A generalization of that approach, using an n-dimensional classification with optimal non-linear boundaries in the parameter space, was proposed in Huertas-Company et al. (2011).

A natural step forward is to take advantage of the recently popular Deep Learning algorithms, which do not require a pre-selected set of parameters to be fit into the model but are able to automatically extract high-level features at the pixel level. In particular, CNNs have been proven very successful in the last years for many different image recognition purposes: manuscript numbers, facial identification, etc. (e.g. Ciresan, Meier & Schmidhuber 2012; Krizhevsky, Sutskever & Hinton 2012; Russakovsky et al. 2015). CNNs have also been used for morphological classification of galaxies, with a high success rate. The use of these automated classification algorithms has been possible thanks to a series of advances in the last few years: the existence of large number of classified objects needed for the training (thanks to Galaxy Zoo project, in particular), the available computing power and a new set of techniques (e.g. rectified linear units – ReLUs – Nair & Hinton 2010 or dropout regularization, Hinton et al. 2012, Srivastava et al. 2014), as well as open source codes which facilitate the task. For example, Huertas-Company et al. (2015) applied CNNs to classify 50 000 CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) galaxies into five groups (spheroid, disc, irregular, point source, and unclassifiable). They obtained zero bias, ∼10 per cent scatter and less than 1 per cent of misclassification. The CNN model presented in Dieleman, Willett & Dambre (2015, hereafter D15), was able to reproduce the GZ2 classification with large accuracy for galaxies with certain classifications. However, one problem with this work is that all biases from GZ2 visual classifications are included; i.e. all the GZ2 catalogue is used for training the models, even galaxies with uncertain classifications.

We follow up that work and create an improved version of the GZ2 catalogue by training our models only with galaxies with very robust GZ2 classification. We also simplify the galaxy decision tree by giving only one probability value for each question (see Section 4). In addition, we complement the GZ2 classification scheme with a T-type, trained with the visually classified catalogue from Nair & Abraham (2010, N10 hereafter). The T-type is an extremely useful parameter for morphological classification because it gives information about the relative importance of the bulge and disc components by one single number. We also use the N10 catalogue to provide a model to separate pure E from S0's and an alternative bar classification to the GZ2-based one. We provide all these values for the sample of ∼670 722 galaxies in the Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7) Main Galaxy Sample (Abazajian et al. 2009) with r-band Petrosian magnitude limits $14 \leq m_r \leq 17.77$ mag published by Meert, Vikram & Bernardi (2015, 2016, see Section 2.3). This is a significant increase in the number of classified galaxies compared to similar available morphological catalogues (almost three times larger than the GZ2 and ∼50 times larger than the N10).

The paper is organized as follows. in Section 2 we introduce the data sets used for training and testing our models, as well as the sample for which the catalogue described in this paper is released. In Section 3, we describe the Deep Learning model and its network architecture. In Sections 4 and 5, we present the methodology and results of our models trained with the GZ2 and the N10 catalogues, respectively. Finally, Section 6 details the content of our morphological catalogue and Section 7 summarizes our main results.

2 DATA SETS

To carry out this work we have benefited from a series of morphological galaxy catalogues, which we use to train and test our Deep Learning models. In this section, we describe the data sets used for training and testing, as well as the final sample to which we apply our models and for which we release our catalogue.

2.1 Catalogues used for training the models

2.1.1 The Galaxy Zoo 2 catalogue

The GZ2 is a public catalogue for ∼240 000 galaxies ($m_r < 17$ mag, $z < 0.25$) of the SDSS DR7 Legacy Survey, with classifications from volunteer citizens. The volunteers have to answer a set of questions for each galaxy image. Depending on the answer, the user is directed to a different question following the GZ2 decision tree. The GZ2 decision tree has 11 classification tasks with 37 possible responses (the number of possible answers per question ranges from 2–7). We encourage the reader to refer to Willett et al. (2013, W13 hereafter) for a detailed description and, in particular, to Fig. 1 for a better understanding of the GZ2 classification scheme, which will be of significant importance throughout this work. The GZ2 catalogue includes number counts of votes and fractions for each answer (weighted and debiased, to correct from observational effects). We take advantage of the GZ2 catalogue for training our models on galaxy classifications similar to the GZ2 decision tree scheme. We base our analysis on weighted fraction values. The weighted fractions are calculated by correcting the vote fractions with a function which downweights classifiers in the tail of low consistency (see W13 for a detailed explanation). Classification bias corrections have been derived in W13 (and refined in a recent work by Hart et al. 2016). The debiased fractions account for changes in the observed morphology as a function of redshift, independent of any true evolution in galaxy properties. The debiased values contain additional information which is not actually included in the images. Therefore, we prefer to restrict our analysis to weighted fractions. Weighted factions are used exclusively hereafter and we will refer to them as $P_{\rm task}$, where *task* is the particular question being discussed.

2.1.2 Nair et al. 2010 catalogue

The Nair & Abraham (2010) is a catalogue based on visual classifications of monochrome g-band images by an expert astronomer for 14 034 galaxies in the SDSS-DR4 in the redshift range $0.01 < z < 0.1$ down to an apparent extinction-corrected limit of $m_g < 16$ mag. The data include RC3 T-types, as well as the existence of bars, rings, lenses, tails, warps, dust lanes, etc. The N10 catalogue

Table 1. Questions from the GZ2 scheme addressed in this work (note that question numbers do not correspond to the ones in table 2 from W13). Also shown the total number (and fraction) of galaxies with enough votes in GZ2 to be used in the training (>5, $N_{\rm votes}$), the number of *certain* galaxies ($N_{\rm certain}$) which fulfil our requirement for being used in the training ($a(p) > 0.3$, see text for a detailed explanation for each question) and the number of positive examples for each question ($N_{\rm pos}$, e.g. the number of galaxies with a bar signature in Q3). The percentages are derived from the parent sample of the previous column (i.e. the fraction of $N_{\rm certain}$ is the number of *certain* galaxies divided by the number of galaxies with enough votes).

Question	Meaning	$N_{\rm votes}$	$N_{\rm certain}$	$N_{\rm pos}$
Q1	Disc/Features	239 728 (99 per cent)	134 475 (56 per cent)	28 513 (21 per cent)
Q2	Edge-on disc	151 560 (63 per cent)	123 201 (81 per cent)	17 631 (14 per cent)
Q3	Bar sign	117 262 (48 per cent)	76 746 (65 per cent)	6 595 (8 per cent)
Q4	Bulge prominence	117 245 (49 per cent)	49 345 (42 per cent)	27 185 (55 per cent)
Q5	Cigar shape	180 223 (75 per cent)	124 610 (70 per cent)	28 230 (23 per cent)
Q6	Merger signature	239 669 (99 per cent)	110 079 (46 per cent)	1 399 (1 per cent)

provides a detailed bar classification, which distinguishes between strong, intermediate, and weak bars (plus additional features and combinations of them). We use the N10 catalogue to train our models for T-type classification, for a complementary bar classification, and to separate pure E form S0 galaxies.

2.2 Catalogues used for testing the models

To study the performance of our models, we combine tests on the catalogues used for training (described in Section 2.1) with tests on available catalogues which are not used in the training process.

2.2.1 Huertas-Company et al. 2011 catalogue

In order to test how our T-Type classification compares with previous automated classifications, we use Huertas-Company et al. (2011) catalogue. This data set contains an automated morphological classification in four types (E, S0, Sab, Scd) based on support vector machines of ∼670 000 galaxies from the Meert et al. (2015) SDSS DR7 sample. Each galaxy is assigned a probability of being in the four morphological classes instead of assigning a single class. We then transform these probabilities into T-types by using equation (7) from Meert et al. (2015).

2.2.2 Cheng et al. 2011

The Cheng et al. (2011) catalogue consists of 984 non-star-forming SDSS galaxies with apparent sizes >14 arcsec and is focused on making finer distinctions between ETGs. It includes a visual classification plus an automated method to closely reproduce the visual results. Galaxies are divided into three bulge classes by the shape of the light profile in the outer regions, roughly corresponding to Hubble types E, S0, and Sa. We use Cheng et al. (2011) catalogue to test the ability of our models to properly separate S0/Sa from pure E galaxies (see Section 5.2).

2.3 Parent sample of the morphological catalogue presented in this work

The catalogue released along with this paper is based on the sample described in Meert et al. (2015, 2016) in order to take advantage of the quality of processed data available for these galaxies. The Meert et al. catalogue contains 2D decompositions in the *g*, *r*, and *i* bands for each of the de Vaucouleur's, Sérsic, de Vaucouleur's + exponential disc and Sérsic + exponential disc models. As discussed in a series of papers (Bernardi et al. 2013; Meert et al. 2015; Bernardi et al. 2017b; Fischer, Bernardi & Meert 2017 and references therein), the SDSS pipeline photometry underestimates the brightness of the most luminous galaxies. This is mainly because (i) the SDSS overestimates the sky background and (ii) single- or two-component Sérsic-based models fit the surface brightness profile of galaxies better than the de Vaucouleur's model used by the SDSS pipeline, especially at high luminosities. In addition to having substantially improved photometry, stellar masses for the objects in this catalogue have recently been added (Bernardi et al. 2017a). Therefore, further augmenting this rich data set with morphological information represents a significant added-value. The reader can refer to Meert et al. (2015, M15 hereafter) for a more detailed description of the sample selection. Once trained and tested, we apply our morphological classification models to all galaxies in that data set. For each galaxy, we provide a probability for each of the questions listed in Table 1, based on GZ2 catalogue. We use the N10 catalogue to derive a T-type and also a probability value of being S0 versus E (to better separate galaxies with T-type ≤ 0), plus an additional bar classification. In Section 6, we summarize the catalogue content and give advise on how to properly use it.

3 DEEP LEARNING MORPHOLOGICAL CLASSIFICATION MODEL

In this work, we apply Deep Learning algorithms using CNNs to morphologically classify galaxy images. Deep Learning is a methodology which automatically learns and extracts the most relevant features (or parameters) from raw data for a given classification problem through a set of non-linear transformations. The main advantage of this methodology is that no pre-processing needs to be done: the input to the machine are the raw RGB cutouts for each galaxy. The main disadvantage is that, given the complexity of extracting and optimizing the features and weights in each layer, a large number of already classified images need to be provided to the machine. Fortunately, as explained in Section 2, there is a wealth of morphological catalogues in the literature overlapping our data set, which we can use for training and testing our model performance.

3.1 Network architecture

Given the high rate of success of previous works using CNNs for visual classification of galaxies (Dieleman et al. 2015; Huertas-Company et al. 2015), we adopt a similar (but not identical) CNN configuration. Testing the performance of different network architectures is beyond the scope of this paper, and we use the same input images and CNN configuration for each classification task.

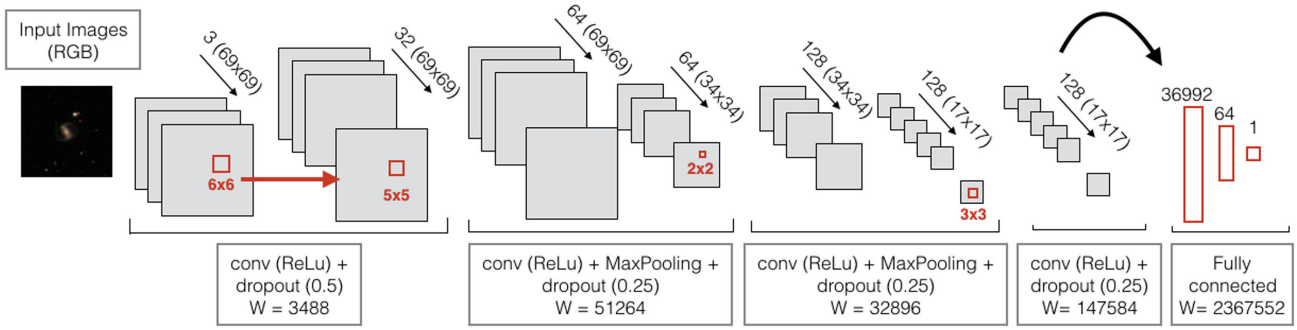

Figure 1. Network architecture used for training the models, consisting on four convolutional layers and a fully connected layer, as explained in the text. The number of weights at each level (*W*) are indicated.

We use the KERAS library,[2] a high-level neural networks application programming interface, written in PYTHON.

The input to our machine are the RGB cutouts downloaded from the SDSS DR7 server[3] in jpeg format, with 424×424 pixels of $0.02\times R_{90}$ arcsec in size (per pixel, where R_{90} is the Petrosian radius for each galaxy). The algorithm reads the images which are downsampled into (69, 69, 3) matrices, with each number representing the flux in a given pixel at a given filter. Downsampling the input matrix is necessary to reduce the computing time and to avoid overfitting in the models. The flux values are normalized to the maximum value in each filter for each galaxy. The network architecture, represented in Fig. 1, is composed of four convolutional layers with squared filters of different sizes (6, 5, 2, and 3, respectively) and a fully connected layer. Dropout is performed after each convolutional layer to avoid overfitting, and a 2×2 max pooling follows the second and third convolutional layers. The number of weights in each layer – before dropout – are also indicated. The output of the fully connected layer is a single value, which has different meanings for each model (see Sections 4 and 5).

We train the models in binary classification mode for GZ2-based questions and in regression mode for the T-type values. The output of the models trained in binary classification ranges from 0 to 1, and it can be interpreted as the probability of being a positive example (example labelled as $Y = 1$ in our input matrix). The output of the T–type model trained in regression mode ranges from -3 to 10, and the returned value is directly the T-type. We use 50 training epochs, with a batch size of 30 and (usually) a *learning rate* of 0.001. We tested the effect of using different *learning rate* values for questions which were more difficult to train (e.g. bars, bulge prominence, and roundness). In the training process, we perform *data augmentation*, allowing the images to be zoomed in and out (0.75–1.3 times the original size), rotated (within 45 degrees), flipped, and shifted both vertically and horizontally (by 5 per cent). This ensures our model does not suffer from overfitting since the input is not the same in every training epoch.

4 GALAXY ZOO 2-BASED MODELS

In this Section, we explain in detail the training methodology and the results obtained for the GZ2-based models listed in Table 1.

4.1 Training methodology

In this work, we use the W13 catalogue for training our GZ2-based models. In D15, a CNN able to reliably predict various aspects of GZ2 galaxy morphology directly from raw pixel data was presented. While their objective was to reproduce the whole GZ2 catalogue, we aim to provide an improved version of the GZ2 classification. In D15, the goal was to predict probabilities for each answer simultaneously solving a regression problem, while we train each question independently using a binary mode classification algorithm. Our main difference with respect to D15 approach is that we only use for the training of each question galaxies with low uncertainties in the GZ2 classification. This allows the model to better identify the important features for each task and to obtain a more evident classification for galaxies for which the GZ2 classification was uncertain (see Section 6).

We do not try to reproduce the whole GZ2 decision tree, but we restrict our analysis to the questions belonging to the third tier. Questions in the lower levels of the classification tree are usually classified by a smaller number of volunteers, reducing the statistics of robust samples, which is fundamental for training our models. Even though in the third tier, we do not address the spiral arm signature nor the bulge shape questions (Q4 and Q9 in W13, respectively), since we believe these tasks are too detailed for the resolution of our binned input images. The tasks included in this work are listed in Table 1.

The fact that GZ2 classifications are based on the answers of citizens, who may not have any background on galaxy images, has some inconveniences. One of the most troublesome tasks is the identification of bar signatures: only the most prominent bars have a high probability of being identified as such, while the weaker features are hardly recovered. For example, only 50 per cent of the weak bars identified by N10 have $P_{\rm bar} > 0.5$ in the GZ2 catalogue ($P_{\rm bar} > 0.5$ is the threshold used in Masters et al. 2011 to select GZ2 barred galaxies). Mergers are also difficult to identify simply by eye, and the sample of galaxies with large $P_{\rm merger}$ in GZ2 is heavily contaminated by projected pairs (see Darg et al. 2010; Casteels et al. 2013). On the other hand, the advantage of the GZ2 classification is that there are sufficient statistics to investigate and quantify these issues.

When the answer for a particular question is not obvious for the volunteers, the vote fractions take intermediate values, meaning that the GZ2 classification for those cases are rather uncertain (see Table 1). Following D15, we quantify the agreement between classifiers, $a(p)$:

$$a(p) = 1 - \frac{H(p)}{\log(n)}, \quad (1)$$

[2] https://keras.io/
[3] http://casjobs.sdss.org/ImgCutoutDR7

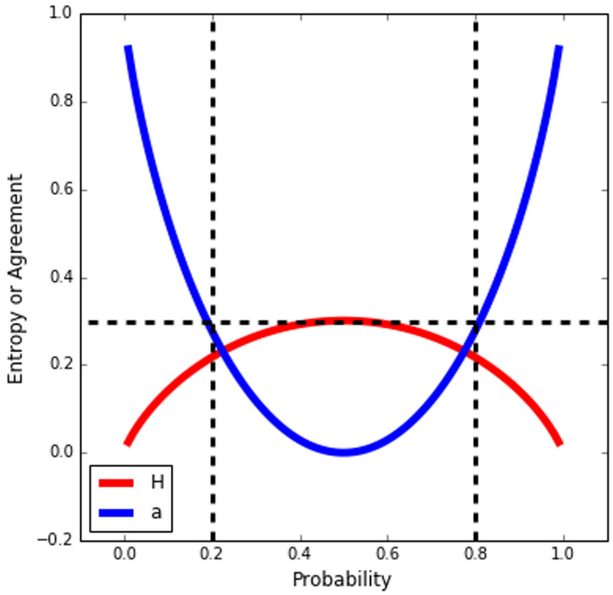

Figure 2. Entropy ($H(p)$, red line) and agreement (a(p), blue line) versus probability for binary questions, where P_1+P_2=1. The dashed line marks the limit used throughout the paper to consider a galaxy in GZ2 as *robust* classification: $P_i < 0.2$ or $P_i > 0.8$, roughly corresponding to $a(p) \geq 0.3$.

where $H(p)$ is the entropy of a question with n possible answers and probability $p(x_i)$ for answer i:

$$H(p) = -\sum_{i}^{n} p(x_i)\log p(x_i). \quad (2)$$

The meaning of $a(p)$ is a measurement of how consistent a classification is, for all the participants that answered that question. In Fig. 2, we show the behaviour of the two functions, $H(p)$ and $a(p)$, for a binary classification. Around 44 per cent of the galaxies in the GZ2 catalogue have an agreement lower than 0.3 for Q1, corresponding approximately to a probability between 0.2 and 0.8 for a binary question (see Table 1). This complicates the usage of the GZ2 catalogue in scientific studies. Another problem is the number of classifiers that have answered a particular question, i.e. the minimum number of votes needed to consider a classification as reliable.

Our methodology consists in only using galaxies with a very robust classification in GZ2 for training each question: we require $P > 0.8$ in one of the two possible answers (these limits are relaxed to 0.7 for questions where the statistic is limited) and a minimum of five vote counts (at least five people have answered that question) in order to use a galaxy in our training sample. This removes noisy galaxies, which are difficult to classify by humans, and allows the model to more rapidly converge. The price to pay is that we have fewer galaxies to train in every question, as can be seen in Table 1.

In addition, instead of allowing more than two answers for some questions, as in the original GZ2 scheme (e.g. the bulge prominence question has four possible outputs: *no bulge, just noticeable, obvious, dominant*), we train our models in binary classification mode, i.e. only positive or negative examples are provided. The loss function used throughout this work for binary classification tasks is `binary-crossentropy` with `adam` optimizer and `sigmoid` activation. Since the output of our model is a probability distribution, that number can be interpreted as the degree of, e.g. bulge importance or roundness.

To summarize, there are three main differences in our methodology compared to D15:

(i) We train each question individually, i.e. we use one model for obtaining each of the parameters contained in the catalogue.
(ii) We use ONLY robust classifications for training our models (more than five votes and $a(p) \geq 0.3$).
(iii) We train the models in binary mode, not in regression mode.

Fig. 3 shows the classification scheme for the GZ2 type questions. Here, we describe in detail some particularities on the training for each question in Table 1:

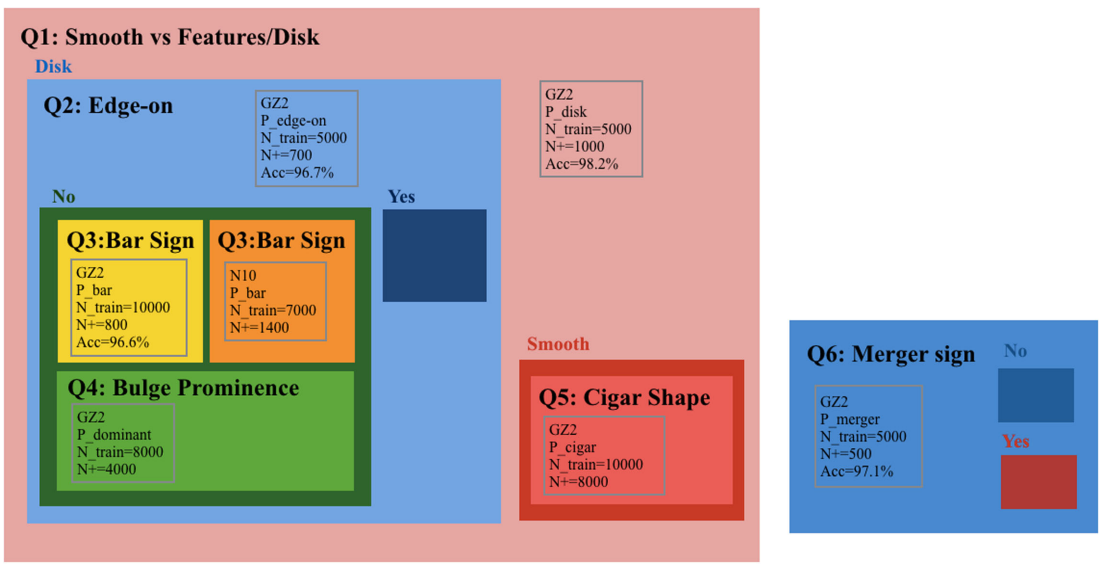

Figure 3. Scheme for our classification of GZ2 type questions. Each box represents a model, with some characteristics framed in grey (from top to bottom: the catalogue used for training, the output of the model, the number of galaxies used in the training, the number of positive examples in the training and the average accuracy – when its computation is feasible). Each box contains additional boxes representing the two possible answers of the model, which may, at the same time, contain additional boxes representing questions trained for that particular subset of galaxies (e.g. the bar classification is only trained with non edge-on disc galaxies).

(i) *Q1 – disc/features*: This question classifies smooth galaxies versus galaxies with the presence of disc or features. It is the first question in the GZ2 classification scheme and has, therefore, been answered by all the participants. Only 20 galaxies in the whole catalogue have less than five votes (adding the smooth and the disc/feature votes), meaning that statistics is not an issue when training this question. However, only 56 per cent of them have a *certain* classification, i.e. satisfy the requirement of having $P_{\rm smooth} > 0.8$ or $P_{\rm disc} > 0.8$, of which $\sim$ 21 per cent are classified as *disc/features*. We use 5000 galaxies in the training ($N_{\rm train}$). For this particular task, this number of galaxies is enough for the models to converge (i.e. setting $N_{\rm train} = 10\,000$ does not improve the model performance). We consider as positive examples galaxies with $P_{\rm disc} > 0.8$. The output of the model is the probability of galaxies having disc or features, $P_{\rm disc}$.

(ii) *Q2 – edge-on galaxies*: This question belongs to the second level of the GZ2 classification scheme (only participants who choose the *disc/features* path were asked this question) and $\sim$63 per cent of the galaxies have >5 votes. However, this is a pretty evident question and $\sim$81 per cent of the galaxies have a *certain* GZ2 classification ($P > 0.8$ in one of the two answers), of which only 14 per cent are edge-on (positive examples). To overcome the small number of positive examples, we use balanced weights (i.e. each instance of the smaller class – edge-on galaxies – contribute more to the final loss, whereas the larger class – non edge-on galaxies – contribute less). The output of the model, trained with $N_{\rm train} = 5000$ galaxies, is $P_{\rm edge\text{-}on}$.

(iii) *Q3 – bars*: This question belongs to the third level of the GZ2 classification scheme (only participants who choose the *disc/features* and *no edge-on* path were asked this question), reducing the sample of galaxies which have at least five votes to $\sim$48 per cent. The fraction of them having $P > 0.8$ in one of the two answers is $\sim$65 per cent, of which only 8 per cent are barred galaxies (positive examples). The small number of barred galaxies complicates the training, which we overcome by increasing the training sample ($N_{\rm train} = 10\,000$) and using balanced weights. The output of the model is the probability of having bar sign, $P_{\rm bar}$.

(iv) *Q4 – bulge prominence*: This question also belongs to the third level of the GZ2 classification scheme (only participants who choose the *disc/features* and *no edge-on* path were asked this question), reducing the sample of galaxies which have five votes to $\sim$49 per cent. In the GZ2 classification, this questions has four possible answers (*no bulge, just noticeable, obvious,* or *dominant*). The fraction of them having $P > 0.7$ in one of the answers is <30 per cent, of which only 132 are bulge dominated. Requiring $P_{\rm dom}$+$P_{\rm obvious} > 0.7$, the fraction increases to 42 per cent. Due to the scarce statistic and for simplicity reasons, we train the model related to this question in a binary classification mode: we consider as positive examples galaxies with obvious or dominant bulge ($P_{\rm dom}$+$P_{\rm obvious} > 0.7$, $\sim$55 per cent of the *certain* sample) against galaxies with no bulge ($P_{\rm no\text{-}bulge} > 0.7$). To obtain better results the *learning rate* value used for training this question was set to 0.000 1. The output of the model, trained with $N_{\rm train} = 8000$, is $P_{\rm bulge}$, i.e. the probability of having an obvious/dominant bulge. We tested that this is the configuration which returns the best results.

(v) *Q5 – roundness*: This question belongs to the second level of the GZ2 classification scheme (only participants who choose the *smooth* option in Q1 were asked this question) and $\sim$75 per cent of GZ2 galaxies have five or more votes. In the GZ2 classification, there are three possible answers to this question (*completely round, in between* and *cigar shaped*) and the fraction having $P > 0.7$ in one of the two answers is $\sim$70 per cent, of which more than a half (63 per cent) are in the *in between* category. We proceed as in Q4 and train the model related to this question in a binary classification mode: we consider as positive examples cigar shape galaxies ($P_{\rm cigar} > 0.7$) against completely round galaxies ($P_{\rm round} > 0.7$). To obtain better results the *learning rate* value used for training this question was set to 0.000 1. The output of the model, trained with $N_{\rm train} = 10\,000$, is $P_{\rm cigar}$, i.e. the probability of having a cigar shape instead of a round shape.

(vi) *Q6 – mergers*: This question belongs to the second level of the original GZ2 classification scheme. Although it is independent of the first answer to Q1, only users who answered *yes* to the question *Is there anything odd?* are then directed to the next question (*what is the odd feature?*), which has seven possible answers: merger, ring, arc/lens, distorted, irregular, dust lane or other. Only $\sim$7 per cent of the GZ2 galaxies have more than five counts in the merger answer, which limits the training sample. We choose a different approach to the GZ2 scheme: we train a model in binary classification mode, as we did with the previous questions. We consider as positive examples galaxies with high probability of being merger combined with a low probability of no presenting anything odd ($P_{\rm merger} > 0.7$ and $P_{\rm no\text{-}odd} < 0.45$), against galaxies which are clearly non merger ($P_{\rm no\text{-}odd} > 0.9$ and $P_{\rm merger} < 0.4$ and at least 10 votes in the *no-odd* answer). Since there are only $\sim$1400 clear merger examples, we use balanced weights. The output of the model, trained with $N_{\rm train} = 5000$, is $P_{\rm merger}$, i.e. the probability of presenting a merger signature. Given the scarce number of merger examples, this was the most challenging question to train in our models. We leave for a forthcoming paper the use of simulated mergers for training a more curated model for merger identification.

4.2 Testing the models

In this section, we detail the performance of our GZ2-based models when tested against a sample of robustly classified galaxies ($a(p) \geq 0.3$), comparable to the one used for training the models.

4.2.1 Questions with two possible answers

In order to quantify the performance of our models for the questions with only two possible answers in GZ2 (Q1, Q2, Q3, Q6), we use two standard methods from the literature: ROC curves and precision-recall versus probability threshold.

A very common way to measure the accuracy of the models is the ROC curve of the classifier (Powers & Ailab 2011). This curve represents the false positive rate (FPR= FP/N, i.e. the ratio between false positive and total negative cases) versus true positive rate (TPR=TP/P, the ratio between true positive and total positive cases) for different probability thresholds ($P_{\rm thr}$). The better the classifier, the closer to the left *y*-axis and upper *x*-axis, i.e. it should maximize TP, and minimize FP values. A complementary way to test the model performance is the precision (Prec) and recall (R) scores (e.g. Dieleman et al. 2015; Barchi et al. 2017), which can be defined as follows:

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}; \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \mathrm{TPR}$$

R, equivalent to the TPR, is a proxy of completeness, while Prec is a purity (contamination) indicator. By choosing different $P_{\rm thr}$ values to consider a galaxy as a positive example, the Prec and R also vary. In Fig. 4, we show these two tests when applying our models to a control sample with similar characteristics to the training sample (i.e. $a(p) \geq 0.3$ and at least five votes) but not used for the training.

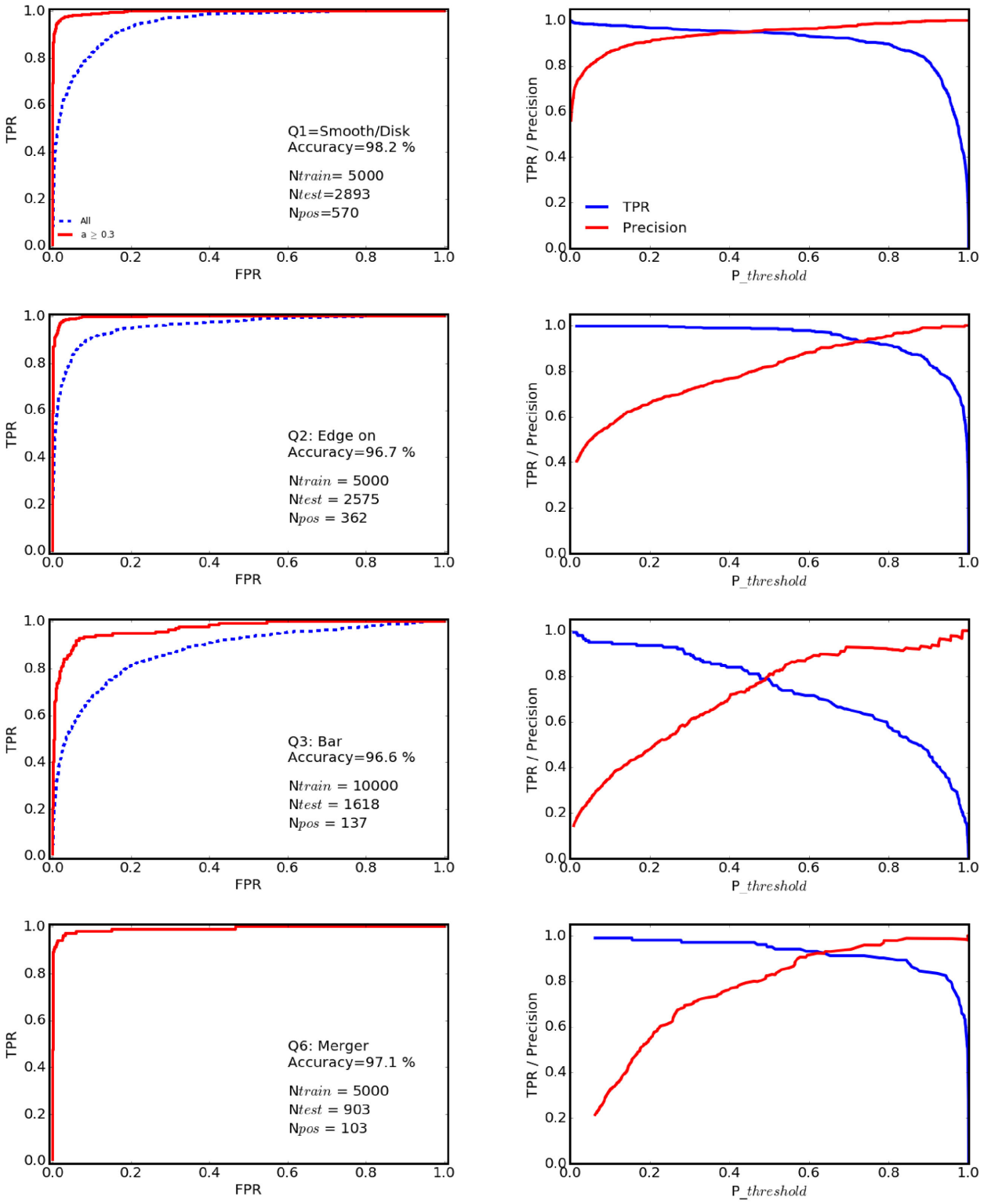

Figure 4. ROC curves (left panels) and TPR, Precision values (blue and red lines, respectively) as a function of P_{thr} (right panels) for the four questions with only two possible answers in GZ2 (disc/features, edge on, bar and merger, from top to bottom). The red lines in the left panels show the results when applying the model to a test sample with the same characteristics as the one used for training ($a(p) \geq 0.3$ and at least five votes). The dashed blue line shows the ROC curve when applied to a test sample without any cut in $a(p)$. Also shown is the number of galaxies used in the training, the number of test galaxies, the number of positive test examples and the average accuracy.

Table 2. Precision and recall (TPR) values for different $P_{\rm thr}$ and average accuracy for the questions which have two possible answers in GZ2 classification scheme.

Question	Meaning	$P_{\rm thr}$	TPR	Prec.	Acc.
		0.2	0.97	0.91	
Q1	Disc/features	0.5	0.95	0.96	0.98
		0.8	0.90	0.99	
		0.2	1.00	0.67	
Q2	Edge-on	0.5	0.99	0.83	0.97
		0.8	0.92	0.95	
		0.2	0.93	0.48	
Q3	Bar sign	0.5	0.79	0.80	0.97
		0.8	0.58	0.92	
		0.2	0.98	0.54	
Q6	Merger signature	0.5	0.96	0.82	0.97
		0.8	0.90	0.97	

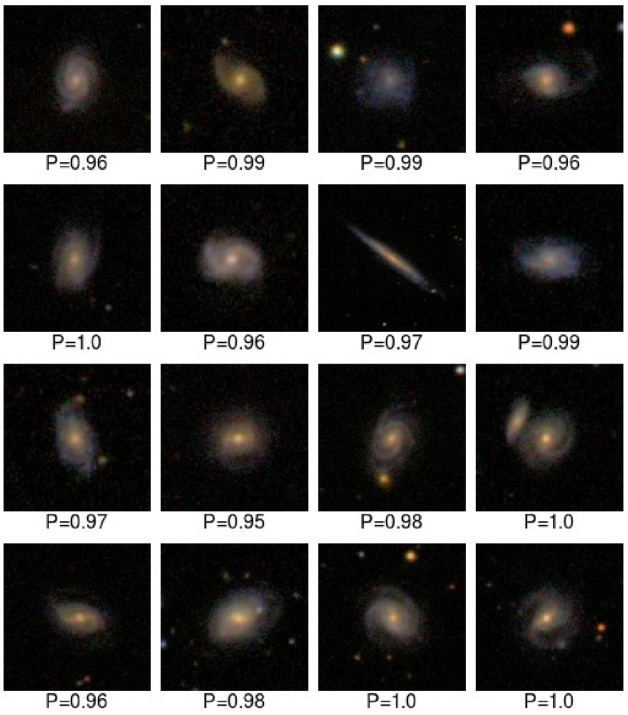

Figure 5. Random examples of galaxies with a high probability of having disc/features according to our model, shown in each cutout. We note that the cutouts have been zoomed-in to the central third of the input images used by the CNN, to better appreciate the detailed morphology. This applies to all the cutouts shown throughout this work.

The crossing point of the red and blue lines in the right panels is the $P_{\rm thr}$ value that optimizes both the Prec and R, but depending on the user purpose, one can vary the $P_{\rm thr}$ to obtain a more complete or less contaminated sample. We tabulate precision and recall values for $P_{\rm thr} = 0.2$, 0.5, and 0.8 for the four questions in Table 2.

The models have a high success rate for all the questions, with total accuracy values defined as:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{(P + N)}$$

higher than 96 per cent and reaching 98 per cent for Q1 (see Table 2, Fig. 4). Also, for all the questions there is a $P_{\rm thr}$ value for which

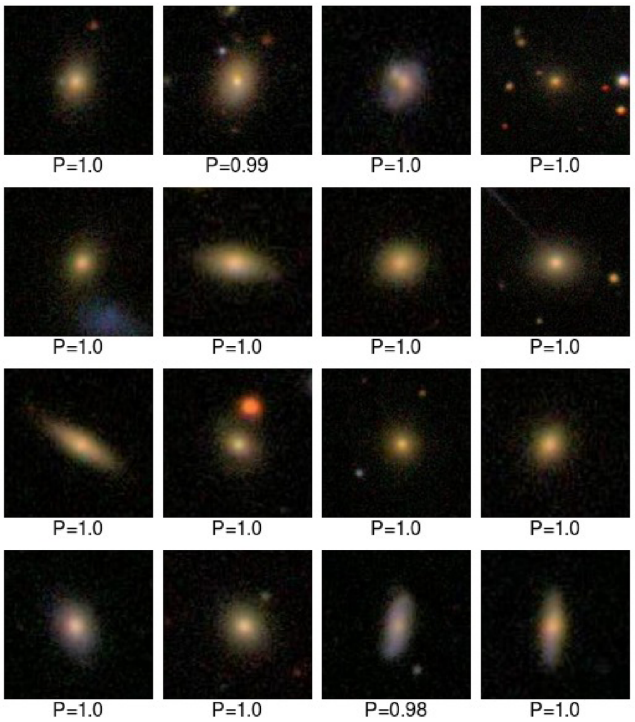

Figure 6. Random examples of galaxies with a high probability of being smooth according to our model.

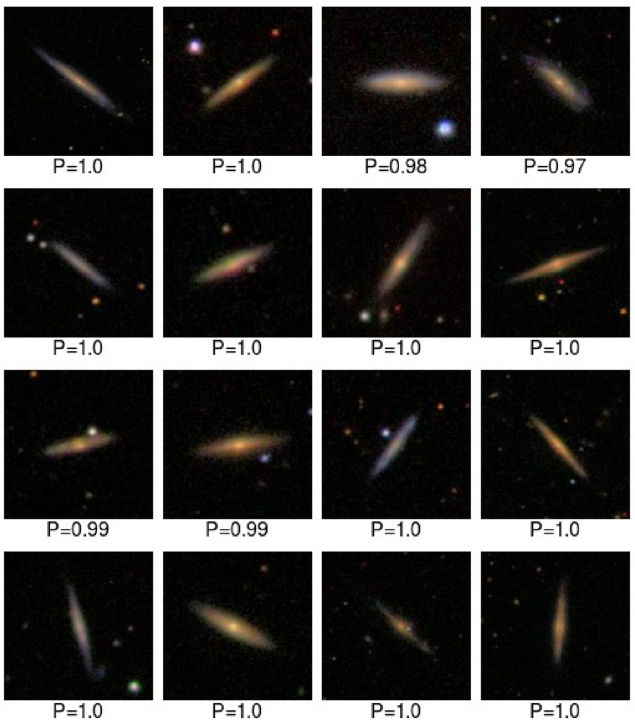

Figure 7. Random examples of galaxies with a high probability of being edge-on according to our model.

both Prec and $R > 0.9$, except for Q3 (barred galaxies, which will be further discussed in section 5.3), for which the maximum is ∼0.8. This is related to the fact that bars are not easily identified by amateur astronomers, but also to the few positive barred examples

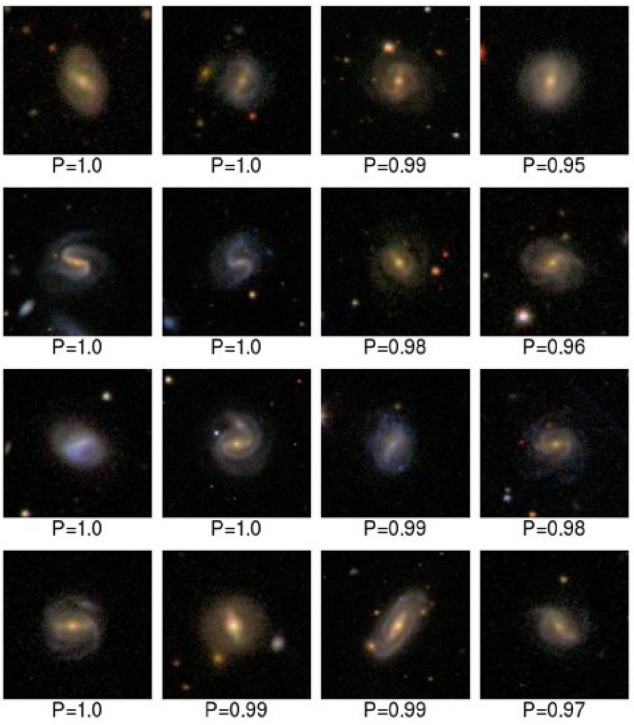

Figure 8. Random examples of galaxies with a high probability of having bar signature according to our GZ2 model. Smooth and edge-on discs galaxies have been removed from the selection.

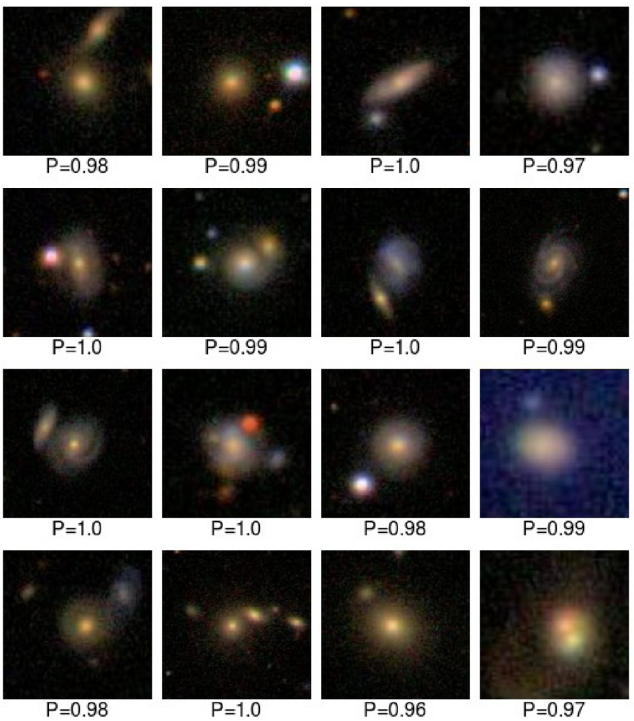

Figure 9. Random examples of galaxies with high probability of showing merger signatures, according to our model.

in our training and testing samples (<10 per cent), which causes the precision value to quickly decrease when few FP cases occur. If we consider the global accuracy of this question (fraction of the correctly classified galaxies), it reaches 96.6 per cent.

Finally, to visually inspect our models, we show some random examples of different galaxy types according to our classification: disc/features, smooth, edge-on, barred, and mergers (Figs 5–9).

4.2.2 *Questions with more than two answers*

The GZ2 scheme includes questions where more than two answers are possible, e.g. the number of spiral arms (five possible answers) or prominence of the bulge (four possible answers). As already mentioned, we do not aim to reproduce the GZ2 classification scheme. In the case of questions with more than two possible answers, we have focused on the prominence of the bulge and the roundness of the galaxy. As explained in Section 4.1, we train these questions on binary mode, discarding intermediate examples to avoid introducing noise in the training.

Testing the behaviour of the models trained in this way by comparison with the GZ2 catalogue is not straightforward since we can not really define TP, TN, FP, FN values as we did for the binary mode questions. We can test how well our derived probability distributions compare to the GZ2 classification for a sample with similar characteristics to our training set (see Section 4.1). This is shown in Fig. 10, for the probability of having a prominent bulge ($P_{\rm bulge}$) and the probability of having cigar shape ($P_{\rm cigar}$).

The extreme cases for each question are clearly separated in the two models. For Q4, there is only a 2 per cent of FN (galaxies classified as bulge dominated in GZ2 which have $P_{\rm bulge} < 0.4$) and less than 0.1 per cent of FP (only three galaxies classified as having no bulge in GZ2 have $P_{\rm bulge} > 0.5$). For galaxies classified as *just noticeable bulge* in GZ2 the distribution is much wider, spanning all possible $P_{\rm bulge}$ values, as expected for intermediate size bulges. There is a 6 per cent of those galaxies for which our model assigned a $P_{\rm bulge} > 0.9$ and 17 per cent with $P_{\rm bulge} < 0.1$.

For question Q5, cigar shape versus round shape, the agreement between the GZ2 classifications and the model distributions is excellent, with less than 0.1 per cent of FP or FN (i.e. galaxies classified as round in the GZ2 with a high $P_{\rm cigar}$ in our model and vice versa). The largest uncertainties are obtained for galaxies classified as *in between* in GZ2, for which we find a 27 per cent with $P_{\rm cigar} < 0.1$. This is probably due to the fact that most GZ2 volunteers, when having an intermediate option, only choose the extreme cases (round or cigar) for the most evident examples.

5 N10 BASED MODELS

This work aims to provide the most complete and accurate morphological classification up to date using Deep Learning models. For this reason, we complement the GZ2 classification with a T-type model trained with the N10 catalogue, as well as an alternative bar classification.

5.1 T-type model

As stated in Section 2, the N10 is a very detailed visual morphological catalogue which assigns an integer number to each galaxy (from −5 to 10) following a structural sequence. The detailed class for each number can be found in Table 1 of N10, but in short, T-type < 0 correspond to ETGs, T-type > 0 are spiral galaxies (from Sa to Sm), T-type = 0 are S0, while T-type = 10 are irregular galaxies.

We use 10 000 galaxies with flag = 0 (i.e. certain classification) for training our T-type models. We apply, though, a minor modification: in N10 the T-type minimum value is −5, but there are no

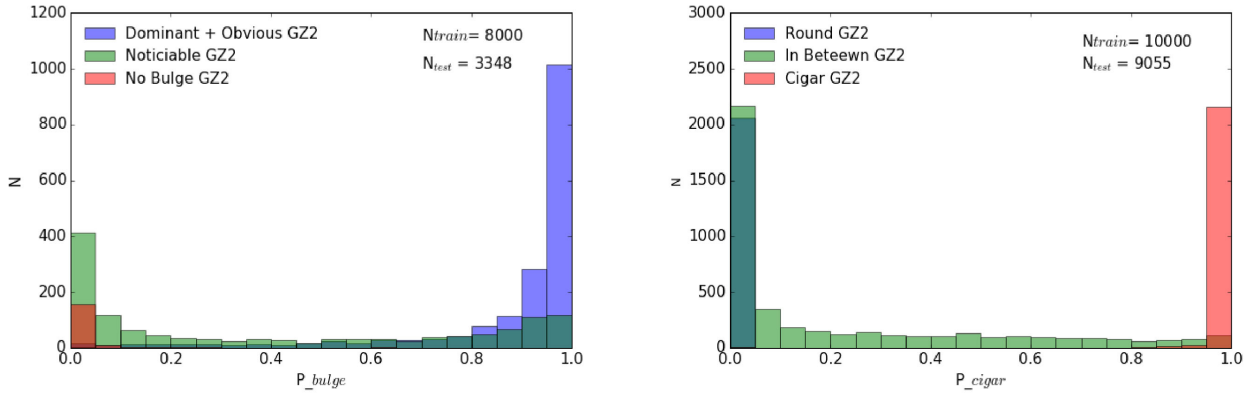

Figure 10. Probability distribution obtained by applying our models to a sample of well classified galaxies. The left-hand panel shows the probability of having a prominent bulge, while the right-hand panel shows the probability of being cigar shaped. Coloured bins represent galaxies with different GZ2 classifications, as stated in the legend. Also shown the number of galaxies used in the training and the number of test galaxies. Our classification is very efficient in separating the extreme cases in both questions.

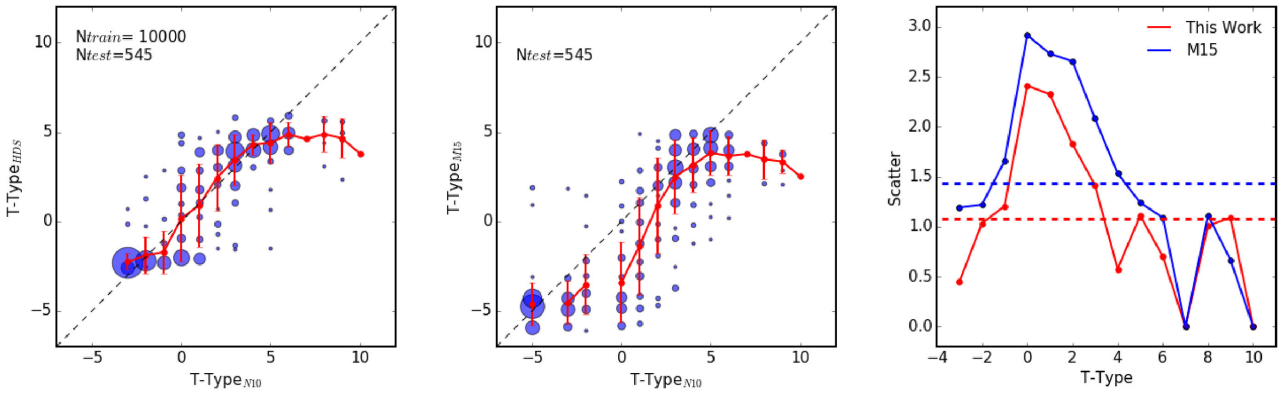

Figure 11. Comparison of our T-type classification with the N10 (left-hand panel) and M15 (central panel). To better visualize it, we plot average binned values, where the size is proportional to the number of objects in each bin. The red dots show the mean value at each T-type, while the error bars show the scatter. The right-hand panel shows the scatter as a function of T-type for our and M15 classifications. Our classification scatter is always smaller than the M15 classification one at all T-types (except for T-type = 10) and on average $\sigma = 1.1$ (red dashed line), comparable or even smaller than visual classification uncertainties ($\gtrsim$1.3; Naim et al. 1995).

galaxies defined as −4 or −1. To facilitate the model to fit a linear regression, we fill those gaps, so our T-types range from −3 to 10. The 0 still corresponds to S0/a, meaning that negative T-types correspond to early-type galaxies (E, S0-), positive T-types correspond to spiral galaxies (from Sa to Sm) and 10 to irregulars. In this case, we use `mean squared error (mse)` as the loss function, which is widely used for linear regression algorithms.

In Fig. 11, we show the comparison between the classification obtained with our models and the N10 classification for a test sample of ∼500 galaxies not used for training the model. The two classifications show an excellent agreement, with a median offset of $b = 0.03$ up to T-type ≤ 6. At higher values, the statistic is very scarce ($\leq$1 per cent) and the model fails to converge. As a comparison, we show, for the same test galaxies, the T-type obtained following equation (7) from M15, which transforms the probability values of being E, Sa, Sb, or Sc, derived by Huertas-Company et al. (2011) using support vector machine models, into a continuous T-type sequence. In this case, there is a median offset larger than one T-type ($b = 1.7$). In this plot, we use the original T-type value from N10 because equation (7) in M15 was optimized for the original catalogue. The scatter for our classification is on average $\sigma = 1.1$, comparable to or even smaller than expert classifier intercomparisons (Naim et al. 1995). The scatter values are always smaller than the scatter for M15 classification for T-type ≤ 6 ($\sigma = 1.4$ on average, right-hand panel Fig. 11). Therefore, we consider that this is an improved T-type catalogue compared to similar available catalogues, both in terms of accuracy and number of classified galaxies (∼50 times larger than the N10). In Fig. 12, we show random examples of galaxies sorted by the T-type derived with our models. The galaxies follow a smooth transition from E to spiral morphologies, as expected.

5.2 Ell versus S0 models

The performance of our model is excellent for the intermediate T-types. However, it shows some flattening at the edges (see Fig. 11). For T-type > 6 that is obviously due to insufficient statistics. On the other hand, the model trained to distinguish between such different morphological types as spirals or ellipticals, is not able to clearly separate between pure E and S0 galaxies, which share many characteristics. In fact, 70 per cent of galaxies classified as ETG (T-type ≤ 0) are assigned a T-type < -2 and the largest scatter is precisely found for T-type = 0. Given that we have enough ETGs to provide a more accurate classification, we train an additional model to separate E from S0 galaxies. We select galaxies with input T-type ≤ 0 (and flag = 0) and label as positive examples those with

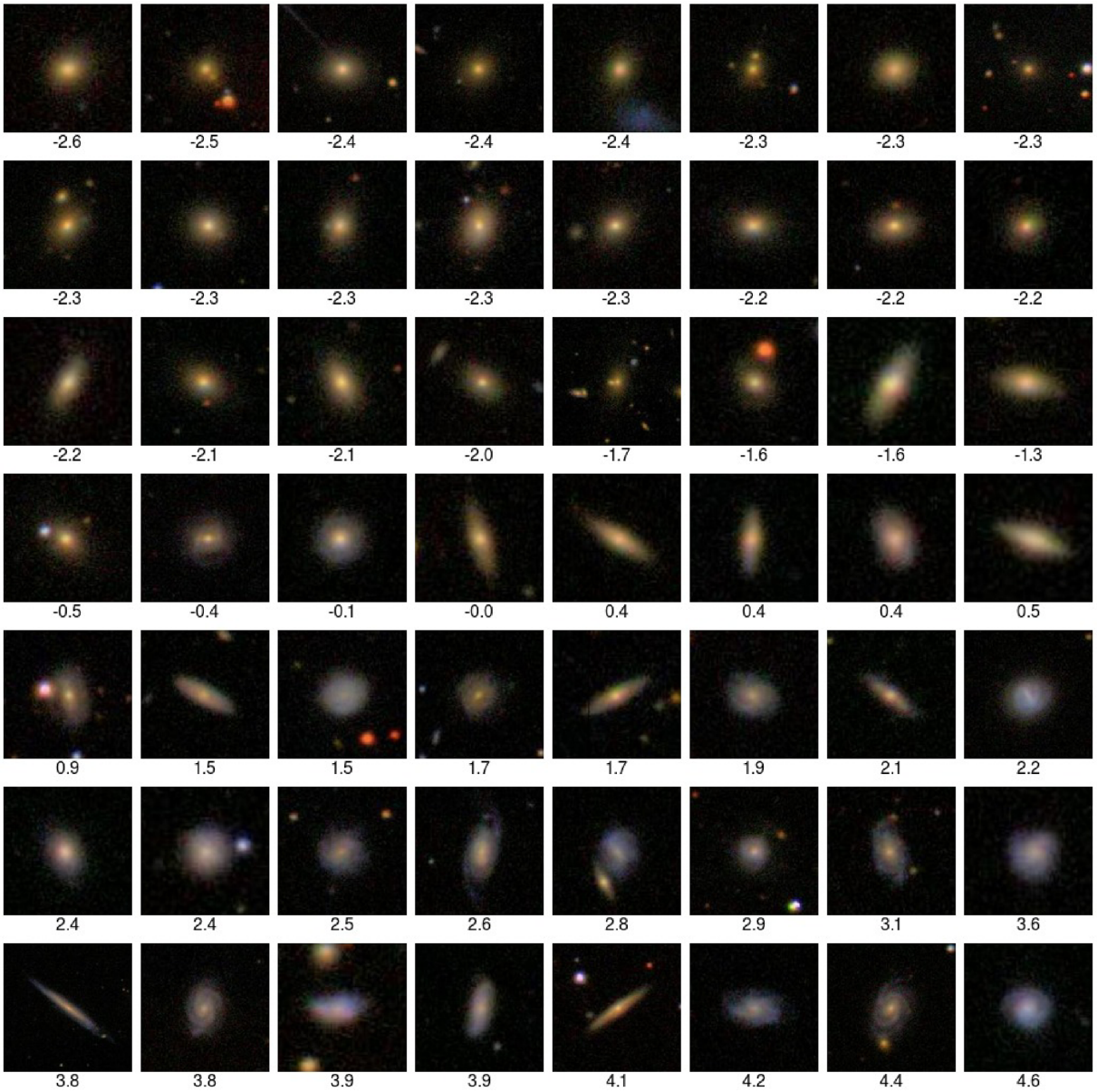

Figure 12. Examples of galaxies sorted by the T-type given by our model (shown in each cutout).

$-3 \leq$ T-type ≤ 0 (S0−, S0, S0+, and S0/a, as defined in table 1 form N10) and as negative those with T-type $= -5$ (c0, E0, E+). We train the model with 4000 galaxies loading the weights of the T-type model, i.e. the weights are initialized to the value learned by the CNN trained for the T-type classification described in Section 5.1. The model output is P_{S0}, i.e. the probability of being S0 rather than E. A schematic classification for the models presented in this section is shown in Fig. 13.

To test this model, we apply it to a sample of 681 galaxies not used in the training with T-type ≤ 0 and study the P_{S0} distribution for each ETGs sub-sample (Fig. 14). The model is very efficient at identifying pure ellipticals: only 6 per cent of the test sample with T-type $= -5$ in N10 is assigned $P_{S0} > 0.5$. Most of the S0/a are also correctly assigned a high P_{S0}, although there is a 10 per cent of them for which $P_{S0} < 0.5$. For the intermediate types (S0−, S0, and S0+), the P_{S0} spans over all P_{S0} values, as expected. We do a complementary check by comparing our P_{S0} values with the bulge classes (BC) values of the Cheng et al. (2011) catalogue (described in Section 2.2) for a sample of ∼600 galaxies in common. We find that 95 per cent of galaxies with BC = 3 (corresponding to Sa) have $P_{S0} > 0.5$, while only 11 per cent of galaxies with BC = 1 (corresponding to E) have $P_{S0} > 0.5$. The fraction of BC = 2 galaxies with $P_{S0} > 0.5$ is 62 per cent, as expected for an intermediate class. Our classification presents larger purity and completeness values when compared to the visual classifications from Cheng et al. (2011) than their automated classification method (75 per cent completeness and 73 per cent purity for the bulge identification, 83 per cent completeness and 70 per cent purity for the discs). When compared to the automated classification provided in Cheng et al. (2011), our classification is not so accurate: 25, 56, and 84 per cent of galaxies with BC = 1, 2, and 3, respectively, have $P_{S0} > 0.5$. This is an indication of our model being more efficient in distinguishing between E and Sa than the automated classification presented in Cheng et al. (2011).

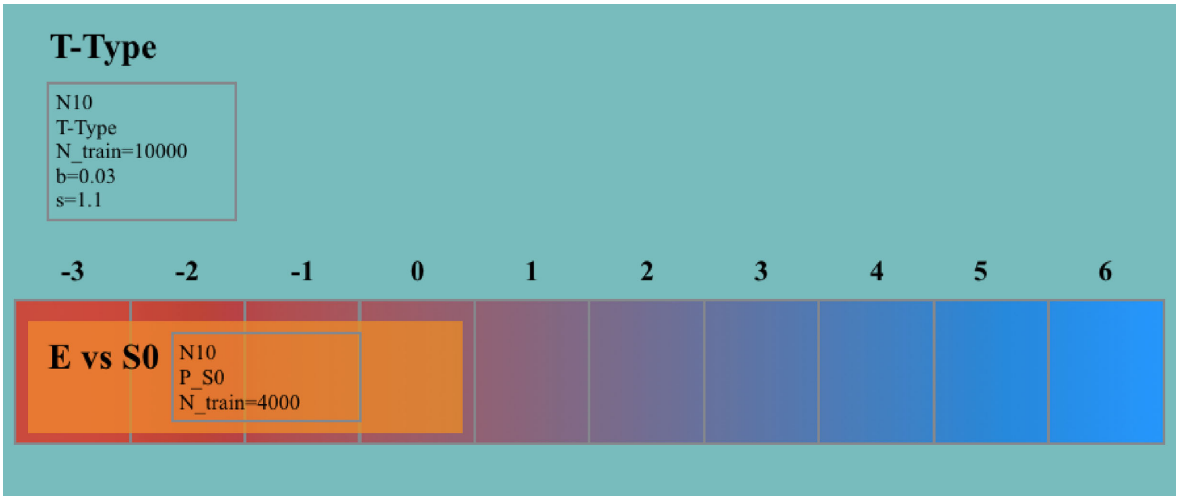

Figure 13. Scheme for our classification of T-type questions. The main turquoise box represents the model for obtaining a T-type value with some characteristics framed in grey (from top to bottom: the catalogue used for training, the output of the model, the number of galaxies used in the training, the average bias and scatter). The coloured smaller boxes (red to blue) show the reliable T-type outputs (at higher T-types the statistic is scarce and our models deviate from the expected values). An additional model to enable a distinction between pure E and S0 galaxies is represented as an orange box, and is only meaningful for galaxies with T-type ≤ 0.

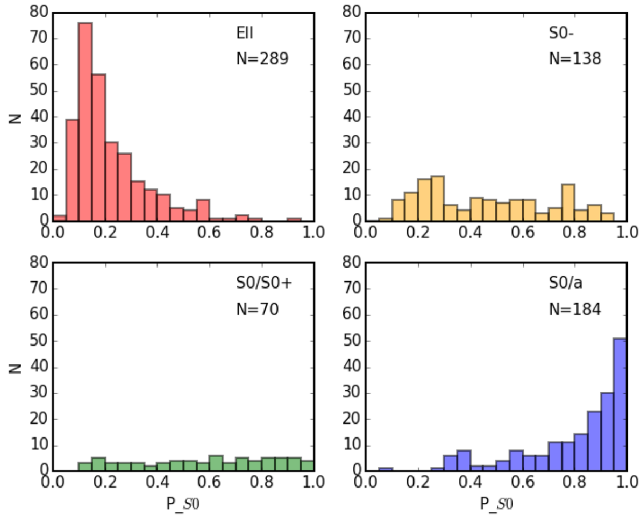

Figure 14. Distribution of the probability of being S0 rather than E (P_{S0}) obtained with our model for a test sample of ETGs divided in four classes (according to N10): E, S0−, S0/S0+, and S0/a. The number of galaxies of each class is shown in each panel. The distribution is clearly skewed towards low P_{S0} for the pure E galaxies and towards higher values for the S0/a. For the intermediate classes, the distribution spans over almost the whole probability range, as expected.

We conclude that this model efficiently allows distinguishing between pure E and S0 galaxies, which is a subtle classification task even for astronomers. We caution the reader that, although we provide a P_{S0} value for each galaxy in our catalogue, it should only be used for galaxies with T-type ≤ 0, for which the model was trained.

5.3 Barred galaxies

The N10 catalogue includes, in addition to the T-type, a detailed visual classification of bars, divided into different classes − strong, intermediate, weak, etc. We take advantage of their bar classification to train an alternative model to the GZ2 based for barred galaxies. We focus on this particular characteristic because our GZ2 bar model is the one with the worst results (see Section 4.2.1). In addition, the GZ2 bar classification is only efficient identifying the strongest bars, while half of the galaxies with weak bar signatures are missed, as already mentioned in Section 4.

We select a sample of barred galaxies (Bar flag > 0 from N10) and non-barred (Bar flag $= 0$). We train our model using 7000 galaxies, of which approximately 20 per cent are barred. We load the weights from the GZ2 bar model, i.e. the weights are initialized to the value learned by the CNN trained for the GZ2 bar classification described in Section 4.1. The results of applying the model to a test sample not used in the training, including 1595 unbarred galaxies and 341 barred galaxies, is shown in Fig. 15. We plot the P_{bar} distribution of our model for galaxies belonging to those four different classes. We correctly classify 90 per cent of unbarred galaxies ($P_{bar} < 0.5$, TN) and 80 per cent of strong bar galaxies ($P_{bar} > 0.5$, TP). However, the scarce number of strong bars (24) makes the statistics very noisy and there are actually five FN, of which only two have $P_{bar} < 0.4$. We visually checked those galaxies, finding that the two extreme cases ($P_{bar} < 0.3$) were affected by close neighbours. Setting $P_{thr} > 0.4$, we obtain 88 per cent and 80 per cent of TP for the intermediate and weak bar samples, respectively. The P_{bar} distribution for the weak sample takes smaller values than the stronger bar sample, indicating that our P_{bar} could also be used as a proxy of bar strength. We visually inspected the FP cases, i.e. 39 galaxies classified as unbarred in N10 but for which our model predicts $P_{bar} > 0.7$ (shown in Fig. 16). In most of the cases, there is a significant central structure, which could be considered as a bar or strong central bulge plus spiral arms. We conclude that no catastrophic failures are found.

As a complementary exercise, we study how well the model trained with the GZ2 catalogue performs with respect to the N10 classification. The GZ2 bar model recovers 96, 80, and 45 per cent of the strong, intermediate, and weak N10 bars. Note that the model trained with GZ2 bar classification is even more efficient in identifying the galaxies with the strong bar signatures, but it fails to recover the weak ones. This demonstrates how the Deep Learning models are affected by the training sample.

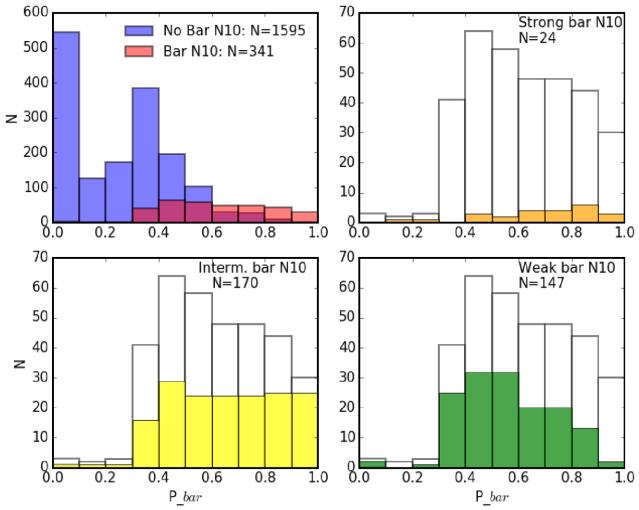

Figure 15. Distribution of the probability of having bar signature, $P_{\rm bar}$, given by our N10 based model for a test sample divided in four classes (according to N10): no bar, strong bar, intermediate bar, and weak bar (filled histograms in blue, orange, yellow, and green, respectively). In the upper left panel, we show the $P_{\rm bar}$ distribution for the barred (red) and unbarred (blue) galaxies. In the other panels, we show the $P_{\rm bar}$ distribution for the barred galaxies (white empty histogram) and for the different classes of barred test galaxies (filled coloured histograms). The number of galaxies of each class is shown in each panel. The values are clearly skewed towards low values for the unbarred sample and towards higher values for the strong and intermediate bars. For the weak bars, the distribution peaks around $P_{\rm bar} \sim 0.5$.

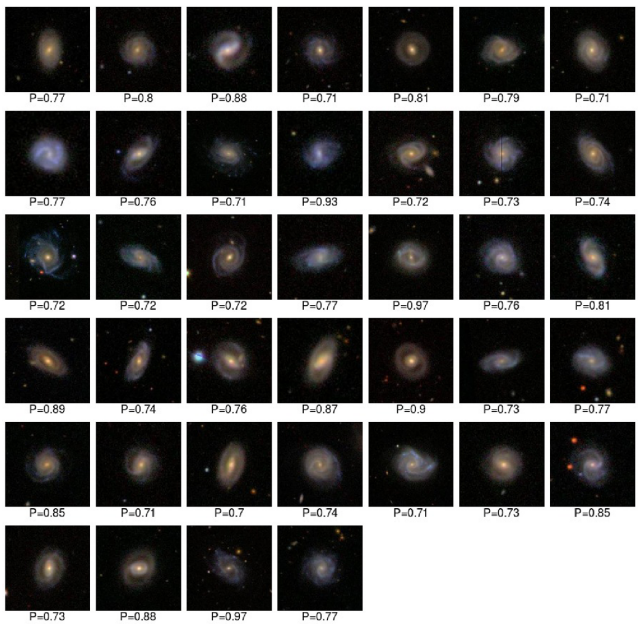

Figure 16. The 39 FP cases in our N10-based bar classification. These are galaxies classified as unbarred by N10 but for which our model predicts $P_{\rm bar} > 0.7$. In most of the cases, there is a significant central structure, which could be considered as a bar or strong central bulge plus spiral arms. No catastrophic failures are found.

6 COMPARING THIS CATALOGUE WITH THE GALAXY ZOO 2

In this section, we summarize the content of the catalogue released with this paper and compare it with the GZ2 catalogue.

Table 3. Content of the catalogue released with this paper. The catalogue contains 670 722 rows, each corresponding to a galaxy from the M15 sample. The last column of this table indicates which catalogue has been used for training each model.

Col.	Name	Meaning	Train sample
1	dr7objid	SDSS ID	
2	galcount	Meert15 ID	
3	$P_{\rm disc}$	Prob. features/disc	GZ2
4	$P_{\rm edge\text{-}on}$	Prob. edge on	GZ2
5	$P_{\rm bar\text{-}GZ2}$	Prob. bar signature	GZ2
6	$P_{\rm bar\text{-}N10}$	Prob. bar signature	N10
7	$P_{\rm merg}$	Prob. merger	GZ2
8	$P_{\rm bulge}$	Prob. bulge prominence	GZ2
9	$P_{\rm cigar}$	Prob. cigar shaped	GZ2
10	T-type	T-Type	N10
11	$P_{\rm S0}$	Prob. S0 versus E	N10

Table 3 summarizes the content of the catalogue presented in this work. A detailed explanation on the training procedure and the performance of all the models has been presented throughout the paper. The catalogue includes classification values for all 670 722 galaxies from the Meert et al. (2015) sample, as explained in Section 2.3. We provide a probability value for each question and galaxy. Depending on the user purpose, a $P_{\rm thr}$ value should be chosen to select *positive* examples. Values of precision (∼purity) and TPR (∼completeness) for three $P_{\rm thr}$ values are tabulated in Table 2 with this objective. For example, if one aims to select a very pure *edge-on* sample, $P_{\rm thr} \sim 0.85$ would be a good choice, while for *disc/feature* galaxies $P_{\rm thr} \sim 0.50$ would be enough.

This is the largest and more accurate morphological catalogue available for the SDSS data up to date. Once trained, applying the models to images of galaxies without any previous classification is straightforward and no time-consuming. Therefore, our catalogue contains a homogeneous GZ2-type classification for a sample of galaxies twice as large as the W13 catalogue. It also provides a T-type value for a sample of galaxies 50 times larger than the previous available T-type catalogue (N10) and a finer separation between E and S0 galaxies. This is the first time, to the best of our knowledge, that a T-type and an E/S0 classification are obtained with Deep Learning algorithms.

The probability distributions of our models are compared with the GZ2 ones in Fig. 17. We recall that we use a `sigmoid` activation function for our binary classification models. This function tends to bring the output values to either end of the probability distribution (0 or 1). In addition, by training the models with robust examples, the machine learns how to recognize the features and the output probabilities. This causes our probability distributions to be generally more bimodal for most of the tasks. Our probabilities should be more objective in the sense that they measure similarity to robustly classified objects. Having a bimodal probability distribution is helpful because it removes galaxies with intermediate probabilities, low $a(p)$, which are difficult to interpret for scientific purposes. This is very evident for Q1, where the fraction of galaxies with $a(p) > 0.3$ increases from 56 per cent for the GZ2 to 86 per cent for our catalogue. The comparison of the fraction of galaxies with a *certain* classification for the questions belonging to the second or third tier of the GZ2 tree, such as $P_{\rm bar}$, $P_{\rm bulge}$, or $P_{\rm merger}$ is more complicated due to the thresholds for determining well-sampled galaxies in GZ2 (according to table 3 in W13). On the other hand, there are tasks, such as $P_{\rm edge}$ or $P_{\rm cigar}$, which show similar distributions for both the GZ2 and our model.

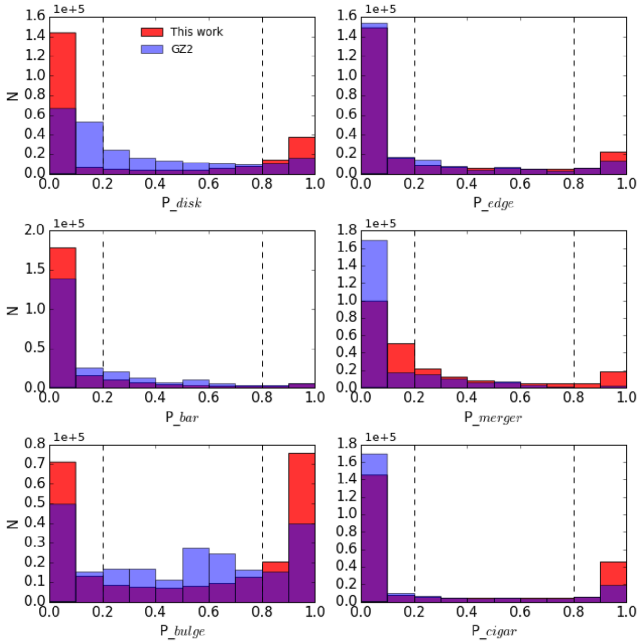

Figure 17. Probability distributions of our models (red) compared to the original GZ2 probabilities (blue) for each of the GZ2-type classification tasks presented in our catalogue, for the sample of galaxies in common (233 472). The dashed vertical lines mark the probability threshold which translates into $a(p) > 0.3$ (for binary classifications). The GZ2 $P_{\rm bulge}$ value is the sum $P_{\rm dominant} + P_{\rm obvious}$, the same we use for training our models (see Section 4.2.2). Note that the $P_{\rm merg}$ comparison is not straightforward due to the different approach used in our models with respect to the GZ2 decision tree (see Section 4.2.1).

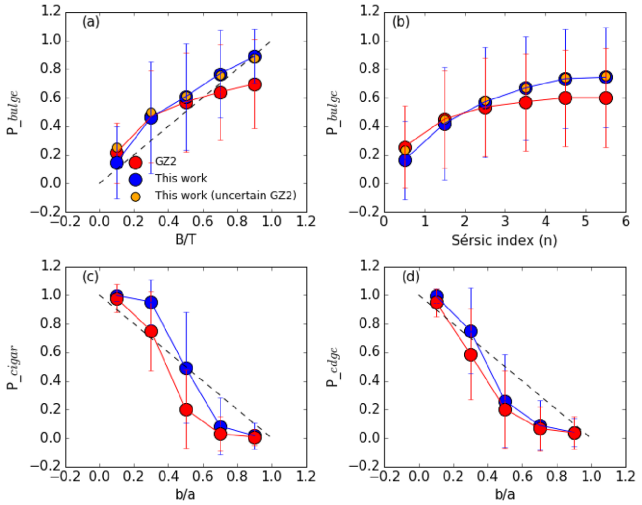

Figure 18. Mean probability values for our models (blue) and the GZ2 catalogue (red) in bins of morphological parameters (extracted from the Sérsic-exponential photometric catalogue presented in M15) for the sample of 233 472 galaxies in common with the GZ2 catalogue. The error bars represent the standard deviation in each bin. (a) $P_{\rm bulge}$ versus bulge-to-total ratio; (b) $P_{\rm bulge}$ versus Sérsic index; (c) $P_{\rm cigar}$ versus ellipticity; (c) $P_{\rm edge}$ versus ellipticity. The orange dots in panels (a) and (b) show the mean $P_{\rm bulge}$ distributions according to our catalogue for galaxies with $0.2 > P_{\rm bulge\text{-}GZ2} > 0.8$. The existence of a similar correlation for this subset of galaxies for which the GZ2 classification is *uncertain* is an indication of the quality of our classification.

A test on the reliability of the output probabilities of our models is their correlation with other morphological parameters. As a reference, we use the morphological parameters provided in the Sérsic-exponential photometric catalogue presented in M15. In Fig. 18, we show mean probability values in bins of bulge-to-total ratio (B/T), bulge Sérsic index (n), or ellipticity (a/b). There is a clear correlation between $P_{\rm bulge}$ and B/T, as well as n. This correlation is stronger for the probabilities provided by our catalogue than for the GZ2 values, demonstrating the physical meaning of the output probabilities of our models. The correlation between a/b and $P_{\rm cigar}$ or $P_{\rm edge}$ is also very evident, for both our probabilities and the GZ2 ones. This is expected, given the similar probability distributions for these two tasks for the GZ2 and our models (see Fig. 17). We also show the mean $P_{\rm bulge}$ according to our model in bins of B/T and n for the sub-sample of galaxies with low $a(p)$ in the GZ2 catalogue (i.e. $0.2 > P_{\rm bulge\text{-}GZ2} > 0.8$). The correlation is also evident for this sub-sample, demonstrating that our probabilities have a physical meaning even for the galaxies with uncertain GZ2 classifications.

Unfortunately, there is no quantitative way to demonstrate that our classification works better than the GZ2 one for galaxies with low $a(p_{\rm GZ2})$, since there is no 'true reference' catalogue. We can only test our models by visual inspection. In Figs 19–21, we show arbitrary examples of galaxies with high output probabilities from our models ($P > 0.9$) and low $a(p_{\rm GZ2})$. In most of the cases, the classification given by our model is robust and correct, while the GZ2 probabilities are much lower (and thus, uncertain).

We also checked the number of *catastrophic errors*, defined as galaxies for which GZ2 and our classification are very different. The fraction of galaxies for which our model predicts $P > 0.8$ and GZ2

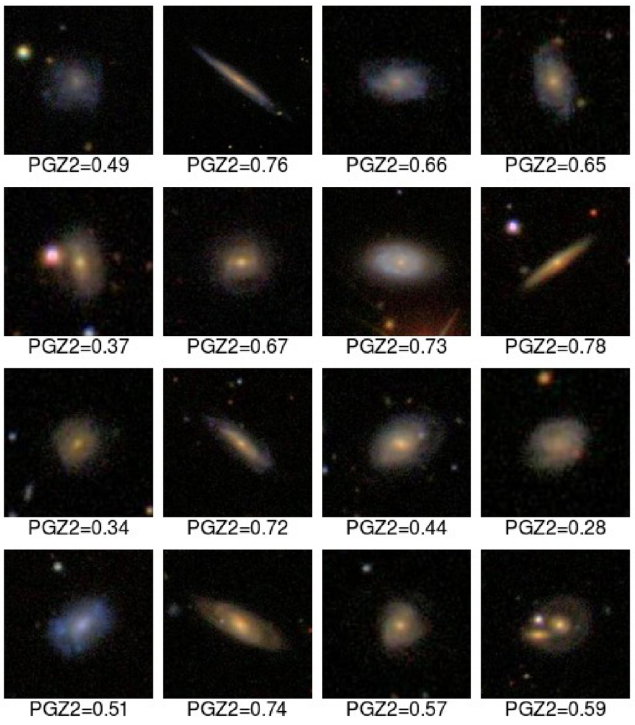

Figure 19. Examples of galaxies with high probability of having disc/features by our model (Q1, $P_{\rm disc} > 0.9$) but with uncertain GZ2 classification ($a(p_{\rm GZ2}) < 0.25$). The number shown in the cutouts is the probability given by the GZ2 catalogue.

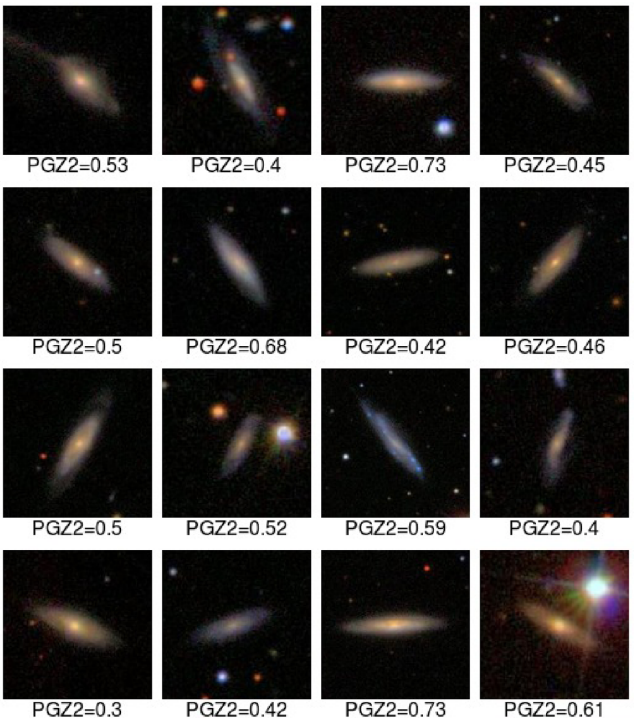

Figure 20. Examples of galaxies with high probability of being edge on (Q2, $P_{\text{edge-on}} > 0.9$) by our model and an uncertain GZ2 classification ($a(p_{\text{GZ2}}) < 0.25$). The number shown in the cutouts is the probability given in GZ2 catalogue.

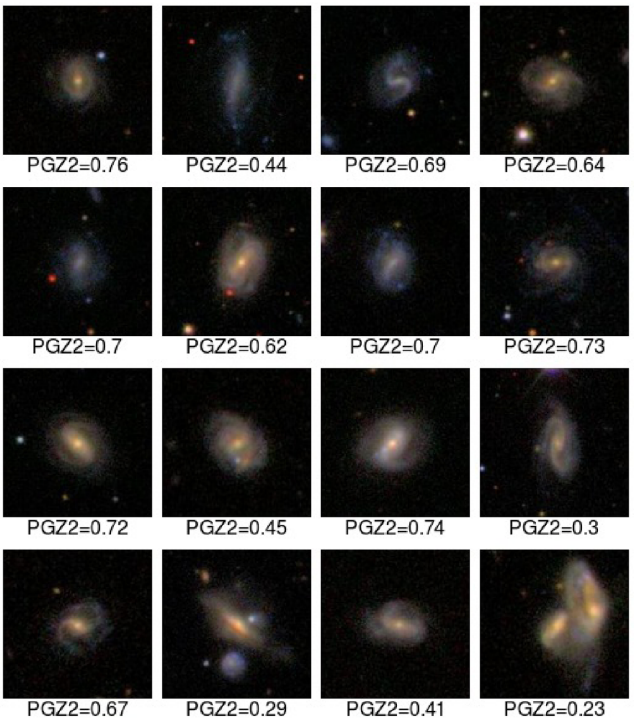

Figure 21. Examples of galaxies with high probability of having bar signatures (Q3, $P_{\text{bar}} > 0.9$) by our model and an uncertain GZ2 classification ($a(p_{\text{GZ2}}) < 0.25$). The number shown in the cutouts is the probability given in GZ2 catalogue.

has $P < 0.4$ (or vice versa) is 2.5, 1.7, and 1.9 per cent for Q1, Q2, Q3, i.e. less than 3 per cent for all questions with two possible answers in GZ2 decision tree. For Q6, mergers, the discrepancy is a bit larger (7.2 per cent), but we want to stress the difficult comparison between our model and GZ2 for this particular question.

An advantage of our catalogue with respect to the GZ2 is that our probabilities are not affected by the number of votes (i.e. the number of volunteers who answered a particular question). Note, however, that the minimum five vote requirement for the training sample of each GZ2 task means that the models somehow contain selection effects of previous questions of the classification scheme (Fig. 3). For example, the probabilities of being bar or edge-on have been trained with galaxies which at least five volunteers have classified as *disc/features*. Therefore, the probability of a question contained in an upper-level box of Fig. 3 should only be completely trusted for positive examples of that particular answer.

Care should also be taken with the P_{S0} value, whose meaning is to differentiate E from S0, i.e. should only be applied to galaxies with T-type ≤ 0. We also advise that, due to the limited merger examples, P_{merger} has difficulty in selecting real on-going mergers. After visual inspection, the P_{merg} value looks like a better proxy to clustered galaxies or projected pairs than to actually interacting galaxies. For simplicity in the catalogue construction and release, we provide one value for each question and galaxy, but we caution the user to properly understand the meaning of each probability when using it.

7 CONCLUSIONS

In this work, we present a morphological catalogue for a sample of ∼670 000 galaxies from the SDSS DR7 corresponding to the sample analysed by Meert et al. (2015, 2016). The morphological classifications are obtained with Deep Learning algorithms using CNNs, and the models are trained with the best available visual classification catalogues (Nair & Abraham 2010; Willett et al. 2013).

We use the GZ2 catalogue presented in W13 to train GZ2 classification models: presence of a disc or features, edge-on discs, bar signature, roundness, bulge predominance, and merger signature. The main novelties of our training approach with respect to previous works (e.g. Dieleman et al. 2015) are:

(i) we independently train each question from the GZ2 scheme listed in Table 1;

(ii) we use in the training only galaxies with small GZ2 classification uncertainties (large agreement, $a(p) \geq 0.3$, between classifiers). This allows the models to easily extract the fundamental features for each question;

(iii) we train the questions in binary classification mode, i.e. only two answers (yes or no) are allowed for each question. The output of each model is the probability of being a positive example, as shown in Fig. 3, and takes values between 0 and 1.

Our models show large accuracy values (>97 per cent) when tested against a sample with the same characteristics as the one used in the training (i.e. *robust* GZ2 classifications). There is a P_{thr} value for each question for which both TPR (∼completeness) and precision (∼purity) are >90 per cent (except for the bar sign, for which TPR and precision only reach ∼80 per cent, see discussion in Section 4.2.1). These values are listed in Table 2. Our morphological catalogue includes a homogeneous classification for 670 722 galaxies, increasing by a factor ∼3 the statistics with respect to GZ2. In addition, we obtain a more unambiguous classification for some of the GZ2-type tasks (see Fig. 17). This result is particularly important

regarding the question about the presence of *disc/features*, where the number of galaxies with $a(p) > 0.3$ increases from 56 per cent in the GZ2 to 86 per cent in our catalogue.

We complement the GZ2 type classification with a T-type value. To this purpose, we train the models with the visual classification catalogue presented in N10. The catalogue presented in this paper is the first T-type classification obtained with CNNs – to the best of our knowledge – and represents a significant increase in terms of statistics compared to previous available T-type catalogues (∼50 times larger than the N10 catalogue). In this case, we train the model using a regression mode, so the output ranges from −3 (E) to 10 (irregular). As shown in Fig. 11, when compared to the T-type from N10, our classification shows no offset and a scatter comparable to or even smaller than typical expert visual classifications ($b = 0.03$, $\sigma = 1.1$). These values are smaller than the ones obtained when comparing N10 T-type with the classification proposed by M15 ($b = 1.7$ and $\sigma = 1.4$). We provide an additional model enabling a separation between E and S0 galaxies. This classification is tested against the N10 and Cheng et al. (2011) catalogues with a great success rate (94 per cent of TP pure E galaxies when compared to N10, see Fig. 14). We also use the N10 bar classification to obtain an alternative model to the GZ2 based for the bar signature, in order to have a complementary bar indicator to the GZ2 based, with a high success rate (>80 per cent TP and TN, see Fig. 15).

We remind that applying the models to images of SDSS galaxies without any previous classification is straightforward and no time-consuming. Therefore, in a forthcoming work (Domínguez Sánchez et al., in preparation), we plan to complement the morphological classification catalogue by applying the models to other SDSS samples, such as the MaNGA data set (Bundy et al. 2015).

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referee for comments that improved the paper. This work was funded by the French National Research Agency (ANR) project ASTROBRAIN (P.I. MHC) and by UPenn research funds (MB). The authors are also thankful to Google for the gift given to UCSC ('Deep-Learning for Galaxies') which has also greatly contributed to making this work possible. Finally, HDS and MHC would like to thank our UCSC colleagues Sandy Faber, David Koo, Joel Primack for very productive discussions which helped to improve the content of the paper.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Abraham R. G., van den Bergh S., Glazebrook K., Ellis R. S., Santiago B. X., Surma P., Griffiths R. E., 1996, ApJS, 107, 1
Barchi P. H., da Costa F. G., Sautter R., Moura T. C., Stalder D. H., Rosa R. R., de Carvalho R. R., 2017, preprint (arXiv:1705.06818)
Barro G. et al., 2017, ApJ, 840, 47
Bernardi M., Meert A., Sheth R. K., Vikram V., Huertas-Company M., Mei S., Shankar F., 2013, MNRAS, 436, 697
Bernardi M., Meert A., Sheth R. K., Fischer J.-L., Huertas-Company M., Maraston C., Shankar F., Vikram V., 2017a, MNRAS, 467, 2217
Bernardi M., Fischer J.-L., Sheth R. K., Meert A., Huertas-Company M., Shankar F., Vikram V., 2017b, MNRAS, 468, 2569
Blanton M. R., Moustakas J., 2009, ARA&A, 47, 159
Bundy K. et al., 2015, ApJ, 798, 7
Casteels K. R. V. et al., 2013, MNRAS, 429, 1051
Cheng J. Y., Faber S. M., Simard L., Graves G. J., Lopez E. D., Yan R., Cooper M. C., 2011, MNRAS, 412, 727
Ciresan D., Meier U., Schmidhuber J., 2012, Proc. CVPR '12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society Washington, DC, USA, p. 3642
Conselice C. J., Bershady M. A., Jangren A., 2000, ApJ, 529, 886
Darg D. W. et al., 2010, MNRAS, 401, 1043
DES Collaboration, 2016, MNRAS, 460, 1270
de Vaucouleurs G., 1963, ApJS, 8, 31
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441 (D15)
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fischer J.-L., Bernardi M., Meert A., 2017, MNRAS, 467, 490
Grogin N. A. et al., 2011, ApJS, 197, 35
Hart R. E. et al., 2016, MNRAS, 461, 3663
Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R., 2012, preprint (arXiv:1207.0580)
Hubble E. P., 1926, ApJ, 64, 321
Hubble E. P., 1936, Realm of the Nebulae. Yale University Press, New Haven
Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, A&A, 525, A157
Huertas-Company M. et al., 2013, MNRAS, 428, 1715
Huertas-Company M. et al., 2015, ApJS, 221, 8
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Bartlett P. L., Pereira F. C. N., Burges C. J. C., Bottou L., Weinberger K. Q., eds, Image Net Classification with Deep Convolutional Neural Networks. Curran Associates, Inc. Red Hook, NY, p. 1097
Linott C. J. et al., 2008, MNRAS, 389, 1179
Lotz J. M. et al., 2008, ApJ, 672, 177
Masters K. L. et al., 2011, MNRAS, 411, 2026
Meert A., Vikram V., Bernardi M., 2015, MNRAS, 446, 3943 (M15)
Meert A., Vikram V., Bernardi M., 2016, MNRAS, 455, 2440
Naim A. et al., 1995, MNRAS, 274, 1107
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427 (N10)
Nair V., Hinton G. E., 2010, in Fürnkranz J., Joachims T., eds, Proc. 27th Int. Conf. Machine Learning (ICML-10). Omnipress, Haifa, Israel, p. 807
Powers D. M. W., 2011, J. Mach. Learn. Technol., 2, 37
Pozzetti L. et al., 2010, A&A, 523, A13
Racca G. D. et al., 2016, in MacEwen H. A., Fazio G. G., Lystrup M., Batalha N., Siegler N., Tong E. C., eds, Proc. SPIE Conf. Ser. Vol. 9904, Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave. SPIE, Bellingham, p. 99040O
Roberts M. S., Haynes M. P., 1994, ARA&A, 32, 115
Russakovsky O. et al., 2015, Int. J. Comput. Vis., 115, 211
Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929
Willett K. W. et al., 2013, MNRAS, 435, 2835 (W13)
Wuyts S. et al., 2011, ApJ, 742, 96

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Morphology_catalogue_DS18.zip

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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