Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **511,** 5120–5124 (2022) https://doi.org/10.1093/mnras/stac457
Advance Access publication 2022 February 18

Galaxy classification: a deep learning approach for classifying Sloan Digital Sky Survey images

Sarvesh Gharat ★ and Yogesh Dandawate
Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune 411048, India

Accepted 2022 February 15. Received 2022 February 15; in original form 2021 September 17

ABSTRACT

In recent decades, large-scale sky surveys such as Sloan Digital Sky Survey (SDSS) have resulted in generation of tremendous amount of data. The classification of this enormous amount of data by astronomers is time consuming. To simplify this process, in 2007 a volunteer-based citizen science project called 'Galaxy Zoo' was introduced, which has reduced the time for classification by a good extent. However, in this modern era of deep learning, automating this classification task is highly beneficial as it reduces the time for classification. For the last few years, many algorithms have been proposed which happen to do a phenomenal job in classifying galaxies into multiple classes. But all these algorithms tend to classify galaxies into less than six classes. However, after considering the minute information which we know about galaxies, it is necessary to classify galaxies into more than eight classes. In this study, a neural network model is proposed so as to classify SDSS data into 10 classes from an extended Hubble Tuning Fork. Great care is given to disc edge and disc face galaxies, distinguishing between a variety of substructures and minute features which are associated with each class. The proposed model consists of convolution layers to extract features making this method fully automatic. The achieved test accuracy is 84.73 per cent which happens to be promising after considering such minute details in classes. Along with convolution layers, the proposed model has three more layers responsible for classification, which makes the algorithm consume less time.

Key words: methods: miscellaneous – techniques: image processing – Galaxy: structure – galaxies: general.

1 INTRODUCTION

There are more than 2 trillion galaxies in our Universe (Conselice et al. 2016; Smith 2016). Those observed so far exhibit a variety of morphologies through cosmic time, showing spiral, elliptical, or irregular shapes and substructural features such as bars and clumps. Morphology of Galaxy relates to its internal properties such as radio emission, star-forming activity (Kennicutt Jr 1998; Bell et al. 2003) and can reveal some insights about its evolutionary history, including mergers (Mihos & Hernquist 1996) and interaction with environment (Sol Alonso et al. 2006).

The importance of morphological classification was first given by Hubble in 1926 (Hubble 1926; Hernández-Toledo et al. 2008; Oswalt & Gilmore 2013) as a descriptive system. In general, the classification was visually done by professional astronomers. However, this method of manual classification has different pros and cons e.g. professional astronomers can classify images with high accuracy, but can only analyse limited amount of data. However, nowadays lots of algorithms have been proposed which extract different features such as colour, shape, brightness profile, concentration index, etc. Compared to past methodologies, the usage of automatic feature-extraction algorithms can significantly reduce the time spent on analysing Galaxy images. These techniques include parametric and non-parametric fitting. In parametric fitting 2D analytical functions are fitted on Galaxy image. During the fitting routine the assumed mathematical model is convolved with the point spread function to account for atmospheric effects. Similarly in non-parametric methods, analysis of light distribution in Petrosian radius is done (Tarsitano et al. 2018). Recently, a decade ago, a citizen science project was launched which has been a great success in classifying large volumes of data (Simmons et al. 2017). However the speed of actual classification has been decreased by smaller extent. In today's era, there are large number of surveys which are expected to yield large volumes of data. This large volume of data would need a large number of volunteers for doing classification. Therefore, research nowadays is seeing the rise of automated algorithms for Galaxy classification.

Recent advancement in machine learning and neural networks has given significant results. With that, the amount of knowledge we have about galaxies has also increased by a larger extent. Hence, it is a must to classify galaxies in more classes than done by traditional algorithms. Tarsitano et al. (2022) gave a detailed overview on different methods used in classification. The authors have also proposed a novel work in their paper. In this study, we propose an automated system to classify galaxies into multiple classes. The proposed algorithm uses convolution neural networks. Convolution layer is responsible for extracting features from images. Each convolution kernel extracts multiple features from the whole input plane (Liu 2018). The extracted features are further treated as an input to classifier which classifies galaxies into 10 classes. More information on why it is important to classify galaxies considering such deep features is given in Section 2.

★ E-mail: sarveshgharat19@gmail.com

The processed data used in this study are collected from a python package named 'AstroNN' (Leung & Bovy 2019) which has Sloan Digital Sky Survey (SDSS) data in it. SDSS is one of the most resolved surveys which has been active since last two decades. After completing one decade of observation, SDSS had collected enough data which created a new 3D map of massive galaxies and distant black holes (Leung & Bovy 2019).

2 RELATED WORK

Different classification techniques use different methodologies. Some techniques include manual feature extraction, while some include automatic feature extraction from Galaxy images or photometric data from Galaxy images.

Walmsley et al. (2020) proposed Bayesian convolution neural network. The proposed technique is similar to VGG16 (Visual Geometry Group). VGG16 is a CNN-based architecture which consists of 16 convolution layers (Simonyan & Zisserman 2014) developed by Visual Geometry Group Lab of Oxford University and has 3 × 3 kernel of different sizes. The proposed technique by Walmsley et. al. uses active learning which reduces the training data by 35 per cent to 60 per cent. The authors classify galaxies based on the bar present in it into barred and unbarred class. To do this, the authors use two variables namely $N_{\rm bar}$ and $k_{\rm featured}$. For galaxies having bar, they consider $N_{\rm bar} \geq 10$. Similarly for unbarred, $N_{\rm bar} < 10$ and $k_{\rm featured} < 10$. The authors have almost used 304 122 images for this study.

Mittal et al. (2020) have proposed a deep convolutional neural network to classify SDSS data in three classes i.e. spiral, elliptical, and irregular galaxies. The data used in this study are less i.e. 4614 images, hence the authors have used data augmentation. The achieved accuracy which in 98 per cent is the best accuracy attained by any model till now, however it is restricted only for three classes.

Eassa et al. (2022) have focused on raw brightness level of galaxies. Based on brightness level and Euclidean distance from centre, the proposed technique classifies the data into spherical, elliptical, and irregular classes. Here, the authors have initially subtracted the background brightness due to stars and other objects so as to get raw brightness level from the Galaxy. This is done by deciding a particular threshold and subtracting all the values from that threshold. Then on using techniques like k means clustering, the galaxies are classified into three classes. The authors have used 1000 images to test and have achieved accuracy of 97.2 per cent.

Jiménez et al. (2020) have used auto-encoder for extracting features from images. The data used in this study are collected from GZ1(Galaxy Zoo 1) (Lintott et al. 2008) data set and consist of more than 650k data images. Out of 650k images, almost 41k of images are classified by professional astronomers, whereas remaining data are classified by amateur astronomers.

Bom et al. (2021) have proposed an RCNN model which allows pixel level segmentation. The study done in this work also focuses on detailed analysis of different learning techniques such as transfer learning, differential learning, and data augmentation. However, the number of classes in which the algorithm classifies the data is just two i.e. spiral and elliptical.

Cavanagh, Bekki & Groves (2021) have proposed four different models for classifying data into two (elliptical and spiral), three (elliptical, spiral, and lenticular), and four (elliptical, spiral, lenticular, and others) classes, respectively. Due to limitation of data, the authors have rotated and flipped the images to generate additional data images. The achieved accuracy after classifying Galaxy Zoo 2 (Willett et al. 2013) data is 92 per cent, 82 per cent, and 77 per cent for two, three, and four classes, respectively.

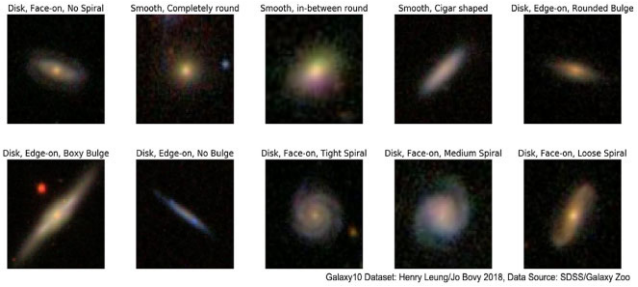

Figure 1. Images of each class in the data set.

All these studies have contributed to a good extent in the classification of SDSS images. In recent days the amount of theoretical knowledge we have is immense, hence it is a must to classify galaxies into more classes. However there's no algorithm to our knowledge which can classify these data into more than eight classes with accuracy more than 75 per cent. This study focuses on a deep learning approach which classifies the data into 10 classes which will solve this problem to a greater extent. The major reason behind classifying galaxies into eight classes or more is the varying properties of galaxies in each and every class. Having classified after considering every minute feature, it is easy to study one particular type of Galaxy in detail, e.g. a radio-loud Galaxy is generally elliptical in shape. At the same time the elliptical galaxies can be divided based on radio jets into fr1 and fr2. Similarly, there are many such unique properties which are found in galaxies with a particular morphology (De Vaucouleurs 1959; Ferguson & Binggeli 1994; De Paz, Madore & Pevunova 2003; Laurikainen, Salo & Buta 2005; Kormendy et al. 2009; Jiang et al. 2013; Graham 2019).

3 METHODOLOGY

This section focuses on the methodology proposed in the study.

3.1 Data collection and preprocessing

The SDSS catalogue considered in this study contains 21 785 visually classified samples from Galaxy Zoo data. All the images in this study are three band (r, g, and i band) coloured images which are classified into 10 classes. As Galaxy Zoo depends on volunteers for classification of data, there are chances that volunteers may make mistakes during classification, hence we only take the images for which more than 55 per cent of people have selected a particular class. The original image in this data set were of 424 × 424 pixels, which are further cropped to 207 × 207 pixels and then downscaled using bilinear interpolation (Gribbon & Bailey 2004; Rukundo & Maharaj 2014) to 69 × 69 pixels.

We have used a python package named astroNN (Leung & Bovy 2019) to import the data set for our study.

3.2 Data distribution

In this study, data have been divide into 10 classes.

Fig. 1 shows us examples of data belonging to each class and Table 1 gives us the number of images present in each class.

Table 1. Distribution of data.

Class	Name	Number of images
Class 0	Disc, Face-on, No Spiral	3461
Class 1	Smooth, Completely round	6997
Class 2	Smooth, in-between round	6992
Class 3	Smooth, Cigar shaped	394
Class 4	Disc, Edge-on, Rounded Bulge	1534
Class 5	Disc, Edge-on, Boxy Bulge	17
Class 6	Disc, Edge-on, No Bulge	589
Class 7	Disc, Face-on, Tight Spiral	1121
Class 8	Disc, Face-on, Medium Spiral	906
Class 9	Disc, Face-on, Loose Spiral	519
	Total	21 785

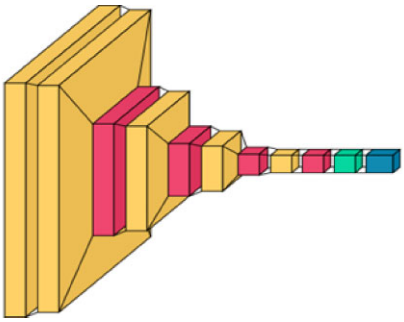

Figure 2. Feature extracting model.

3.3 Model architecture

The proposed Neural Network model is divided into two parts.

3.3.1 Feature extraction model

In this model, we have used convolution layers to extract features. Convolution layers are meant to extract high-level features from the image provided in input plane (Albawi, Mohammed & Al-Zawi 2017; Gu et al. 2018; Alzubaidi et al. 2021). Convolution layers apply the filter to input image to create a feature map. Each 2D convolution layer takes $n \times n$ array and applies k filters with the help of $m \times m$ kernel so as to extract k features. Here, $n \times n$ are dimensions of input image (individually r, g, b bands), $m \times m$ are the dimensions of convolution kernel, and k is the number of filters (LeCun et al. 1989; LeCun et al. 1998; Gu et al. 2018). With convolution layers, we use pooling layers to downsample the feature map. In this work, we have used MaxPooling which helps to preserve maximal elements and reduces the noise (Christlein et al. 2019; Gholamalinezhad & Khosravi 2020). Each convolution layer used in this model has 'Relu' as an activation function. Relu is an activation function which happens to be 0 for all negative inputs and same as input for all the positive values. This is done to increase the non-linearity of the last component in each convolution layer. After extracting features with the help of convolution and pooling layer, all the feature maps are flattened into a 1D array.

In Fig. 2, convolution layer is represented by yellow colour, pooling layer by red colour, dropout by green colour, and flattening layer by blue colour.

As it is evident from Fig. 2, the feature extracting model has multiple convolution blocks starting with convolution layer, followed by a block of alternating convolution layer and MaxPooling layer.

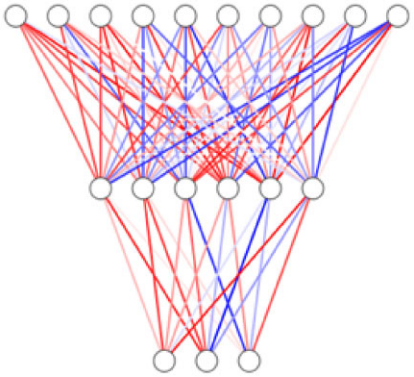

Figure 3. Classifier model.

3.3.2 Classifier model

The output of feature extracting model acts as an input to the classifier.

The classifier has two dense layers before the output layer as seen in Fig. 3. The first dense layer in the classifier has 64 neurons, whereas the second layer has 32 neurons. As the number of output classes are 10, the number of neurons in the output layer is 10. The activation function used in the first two dense layers is 'Relu'; in the output layer we use 'Softmax' as an activation function. This is done so that the output of the model will be the probability of image belonging to a particular class (Agarap 2018; Nwankpa et al. 2018; Feng & Lu 2019). To avoid overfitting, a regularization technique named dropout is used in both dense layers. As dropout of 0.2 is used, 20 per cent of neurons act as dead neurons (Srivastava et al. 2014; Srinivas & Babu 2016; Cai et al. 2019).

3.4 Training and validation

The proposed model is constructed using keras (Chollet et al. 2015). The training of model is conducted on Google Collaboratory, making use of NVIDIA K80 GPU (Carneiro et al. 2018). The data set in this study is divided in train, test, and validation data set in the ratio of 70:15:15 i.e. 70 per cent of data are used for training (15 249 images), 15 per cent of data are used for testing (3268 images), and remaining 15 per cent of data are used in validation data set (3268 images). Validation data set which is independent of testing data set is used for hyperparameter tuning so as to avoid any biasing in choice of hyperparameters. Thus, when the network is completely trained, evaluation is done on completely unseen test data set.

Training was conducted over maximum of 50 epochs. For efficient training, we use early stopping so that the training ends once there is no appreciable decrease in validation loss. Early stopping helps to avoid unnecessary computation once the validation accuracy has reached its peak and also avoids overfitting (Caruana, Lawrence & Giles 2001; Montavon, Orr & Müller 2012; Song et al. 2019; Ying 2019).

In Fig. 4, blue colour represents training loss and accuracy, whereas red colour represents validation loss and accuracy.

As seen in Fig. 4, during initial epochs the gradient of loss is high but at later stage it decreases. Here is when early stopping helps. If there was no early stopping the loss would have further increased, making the accuracy to decrease. The validation loss and accuracy as seen in Fig. 4 represents a metric corresponding to repeated evaluation of networks.

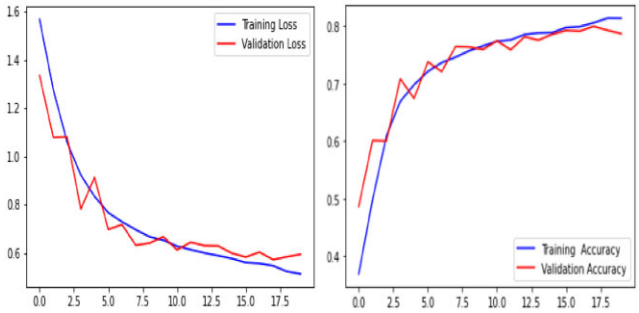

Figure 4. Graph of accuracy and loss against number of epochs.

Table 2. Confusion matrix for testing data set.

	0	1	2	3	4	5	6	7	8	9
0	342	52	60	3	6	0	1	29	11	10
1	15	987	11	0	0	0	0	4	0	0
2	30	60	850	2	1	0	0	0	0	0
3	2	0	4	53	4	0	7	0	0	2
4	16	0	7	3	206	1	15	0	0	1
5	0	0	0	0	1	3	0	0	0	0
6	0	0	0	3	20	0	67	0	0	1
7	22	5	2	0	2	0	0	101	10	0
8	20	1	2	0	0	0	0	11	99	9
9	10	0	3	1	4	0	2	3	16	61

4 RESULTS AND DISCUSSION

Almost all the previous studies done in classification of galaxies are done to classify galaxies into at most five classes (Jiménez et al. 2020; Mittal et al. 2020; Walmsley et al. 2020; Bom et al. 2021; Cavanagh et al. 2021; Eassa et al. 2022). But considering the development in theoretical aspects, we found that there was a need of classifying it into more classes. Reason for that is given in Section 2.

In this study, we define a convolutional neural network as mentioned in Section 3 which has been trained on SDSS data set to classify galaxies in 10 classes.

In Table 2, vertically we have the number of images predicted by algorithm belonging to a particular class. On the other hand, on horizontal axes we have a number of images that actually belong to a particular class. We know that the overall accuracy of the model is given by

$$\text{Accuracy} = \frac{\sum_{i=0}^{9} a_{ii}}{\sum_{i=0}^{9}\sum_{j=0}^{9} a_{ij}} \times 100$$

In our case it will be

$$\text{Accuracy} = \frac{342 + 987 + \cdots + 61}{342 + 52 + 60 + \cdots + 4 + 16 + 61} \times 100 = 84.73 \text{ per cent}$$

From confusion matrix we can see that 172 images (33.4 per cent) are misclassified as class 0; this is because of the similarity which class 0 i.e. Disc Face on and No spiral has with other classes. Also classes 7, 8, and 9 just vary in the structural wound they have. This is commented after visually looking images from all the classes.

Another mean of testing the model is to calculate overall F1 score.

$$\text{F1} = \frac{2PR}{P + R}$$

With reference to Table 3, Precision tells about the false positive rate i.e. precision of class 0 will tell about the images misclassified as class 0. Similarly, Recall tells about the false negative rate. The value of Precision and Recall ranges from 0 to 1. The F1 score for any model is defined as the harmonic mean of Precision and Recall. It is calculated so as to balance these two parameters.

Table 3. Precision, recall, and F1 score.

Class	Precision	Recall	F1 score
Class 0	0.748	0.665	0.704
Class 1	0.893	0.971	0.930
Class 2	0.905	0.901	0.902
Class 3	0.815	0.736	0.773
Class 4	0.844	0.827	0.835
Class 5	0.750	0.750	0.750
Class 6	0.728	0.736	0.731
Class 7	0.682	0.711	0.696
Class 8	0.728	0.697	0.712
Class 9	0.726	0.610	0.662
Average	0.782	0.760	0.769

For each class reported in the table, the higher the values of the three parameters, the better the model. The last three classes have very similar morphology and they show similar accuracies, hence the probability of misclassification is higher. Also, as discussed earlier, lots of images (172 images) are misclassified as class 0, so the lower F1 score of classes 0, 7, 8, and 9 contributes to decrease the average F1 score of the model.

5 CONCLUSIONS

In this study, we propose a convolutional neural network to classify galaxies into 10 classes. This is one of the initial works wherein the galaxies are classified into 10 classes by considering such minute details. This detailed classification happens to be of need after considering the theoretical knowledge we have. This was initially done by professional astronomers but due to the large amount of data we have, it was impossible for them to continue. Further, different citizen scientists were trained to do this task, but in modern era the data we have are huge in number compared to that of available volunteers. Hence, this algorithm happens to solve this problem. Also, the time taken by the algorithm to classify large volume of data set is less than 10 min, which is another advantage of using the automated algorithms over manual classification. The proposed algorithm gives accuracy of 84.5 per cent which is good after considering such minute details in classification.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

All the data used in this study are publicly available on astronn's official website. The code used for the proposed model will be made available after 2 yr of publishing on request.

REFERENCES

Agarap A. F., 2018, Deep learning using rectified linear units (relu), preprint (arXiv:1803.08375)

Albawi S., Mohammed T. A., Al-Zawi S., 2017, in International Conference on Engineering and Technology (ICET). Akdeniz University, Antalya, Turkey, p. 1

Alzubaidi L. et al., 2021, J. Big Data, 8, 53
Bell E. F., Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 2003, MNRAS, 343, 367
Bom C. et al., 2021, MNRAS, 507, 1937
Cai S., Shu Y., Chen G., Ooi B. C., Wang W., Zhang M., 2019, in Effective and efficient dropout for deep convolutional neural networks. preprint (arXiv:1904.03392)
Carneiro T., Da Nóbrega R. V. M., Nepomuceno T., Bian G.-B., De Albuquerque V. H. C., Reboucas Filho P. P., 2018, IEEE Access, 6, 61677
Caruana R., Lawrence S., Giles L., 2001, Advances in neural information processing systems. MIT, p. 402
Cavanagh M. K., Bekki K., Groves B. A., 2021, MNRAS, 506, 659
Chollet F. et al., 2015, Keras, https://github.com/fchollet/keras
Christlein V., Spranger L., Seuret M., Nicolaou A., Král P., Maier A., 2019, in International Conference on Document Analysis and Recognition ICDAR. IEEE, p. 1090
Conselice C. J., Wilkinson A., Duncan K., Mortlock A., 2016, ApJ, 830, 83
De Paz A. G., Madore B., Pevunova O., 2003, ApJS, 147, 29
De Vaucouleurs G., 1959, in Astrophysik iv: Sternsysteme/astrophysics iv: Stellar systems. Springer, Berlin, Heidelberg, p. 275
Eassa M., Selim I. M., Dabour W., Elkafrawy P., 2022, Alex. Eng. J., 61, 1145
Feng J., Lu S., 2019, J. Phys. Conf. Ser., 1237, 022030
Ferguson H. C., Binggeli B., 1994, A&AR, 6, 67
Gholamalinezhad H., Khosravi H., 2020, Pooling methods in deep neural networks, preprint (arXiv:2009.07485)
Graham A. W., 2019, MNRAS, 487, 4995
Gribbon K. T., Bailey D. G., 2004, in Proceedings. DELTA 2004. Second IEEE International Workshop on Electronic Design, Test and Applications. IEEE, Perth, WA, Australia, p. 126
Gu J. et al., 2018, Pattern Recognit., 77, 354
Hernández-Toledo H., Vázquez-Mata J., Martínez-Vázquez L., Reese V. A., Méndez-Hernández H., Ortega-Esbrí S., Núñez J. M., 2008, ApJ, 136, 2115
Hubble E. P., 1926, ApJ, 64
Jiang L. et al., 2013, ApJ, 773, 153
Jiménez M., Torres M. T., John R., Triguero I., 2020, IEEE Access, 8, 47232
Kennicutt R. C. Jr, 1998, ARA&A, 36, 189
Kormendy J., Fisher D. B., Cornell M. E., Bender R., 2009, ApJS, 182, 216
Laurikainen E., Salo H., Buta R., 2005, MNRAS, 362, 1319
LeCun Y., Boser B., Denker J., Henderson D., Howard R., Hubbard W., Jackel L., 1989, Advances in neural information processing systems, 2
LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
Leung H. W., Bovy J., 2019, MNRAS, 483, 3255
Leung H. W., Bovy J., III Press Releases, http://www.sdss3.org/press/
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Liu Y. H., 2018, J. Phys. Conf. Ser., 1087, 062032
Mihos C., Hernquist L., 1996, ApJ, 464, 641
Mittal A., Soorya A., Nagrath P., Hemanth D. J., 2020, Earth Sci. Inform., 13, 601
Montavon G., Orr G., Müller K.-R., 2012, Neural Networks: Tricks of the Trade. Vol. 7700, Springer, Berlin, Heidelberg
Nwankpa C., Ijomah W., Gachagan A., Marshall S., 2018, Activation functions: Comparison of trends in practice and research for deep learning, preprint (arXiv:1811.03378)
Oswalt T. D., Gilmore G., 2013, Planets, Stars and Stellar Systems: Volume 5: Galactic Structure and Stellar Populations. Springer, Berlin, Heidelberg
Rukundo O., Maharaj B. T., 2014, in 2014 International Conference on Computer Vision Theory and Applications (VISAPP). IEEE, Lisbon, Portugal, p. 641
Simmons B. D. et al., 2017, MNRAS, 464, 4420
Simonyan K., Zisserman A., 2014, Very deep convolutional networks for large-scale image recognition, preprint (arXiv:1409.1556)
Smith K. T., 2016, Science, 354, 844
Sol Alonso M., Lambas D. G., Tissera P., Coldwell G., 2006, MNRAS, 367, 1029
Song H., Kim M., Park D., Lee J.-G., 2019, Prestopping: How does early stopping help generalization against label noise?
Srinivas S., Babu R. V., 2016, Generalized dropout, preprint (arXiv:1611.06791)
Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929
Tarsitano F. et al., 2018, MNRAS, 481, 2018
Tarsitano F., Bruderer C., Schawinski K., Hartley W., 2022, MNRAS, 511, 3330
Walmsley M. et al., 2020, MNRAS, 491, 1554
Willett K. W. et al., 2013, MNRAS, 435, 2835
Ying X., 2019, J. Phys. Conf. Ser., 1168, 022022

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