The Predictive Power of Computational Astrophysics as a Discovery Tool
Proceedings IAU Symposium No. 362, 2023
D. Bisikalo, D. Wiebe & C. Boily, eds.
doi:10.1017/S1743921322001259

The CNN classification of galaxies by their image morphological peculiarities

D. Dobrycheva[1], **V. Khramtsov**[2], **M. Vasylenko**[1,3] **and I. Vavilova**[1]

[1]Main Astronomical Observatory of the NAS of Ukraine, 27 Akademik Zabolotny Str., Kyiv, 03143, Ukraine

[2]Institute of Astronomy, V.N. Karazin Kharkiv National University, 35 Sumska St., Kharkiv, 61022, Ukraine

[3]Institute of Physics of the National Academy of Sciences of Ukraine, 46 avenue Nauka, Kyiv, 03028, Ukraine

Abstract. Multidimensional mathematical analysis, like Machine Learning techniques, determines the different features of objects, which is difficult for the human mind. We create a machine learning model to predict galaxies' detailed morphology (∼300000 SDSS-galaxies with $z < 0.1$) and train it on a labeled dataset defined within the Galaxy Zoo 2 (GZ2). We use convolutional neural networks (CNNs) to classify the galaxies into five visual types (completely rounded, rounded in-between, smooth cigar-shaped, edge-on, and spiral) and 34 morphological classes attaining $> 94\%$ of accuracy for five-class morphology prediction except for the cigar-shaped (∼87%) galaxies.

Keywords. methods: data analysis – galaxies: general– surveys – methods: convolutional neural networks, etc.

1. Introduction

The morphological classifications of galaxies play a vital role in reflecting the evolutionary history of various types of galaxies and the large-scale structure of the Universe as a whole (Barrow and Saich (1993); Peng et al. (2010); Reid et al. (2012); Dobrycheva et al. (2018); Vavilova et al. (2021a, 2020a); Elyiv et al. (2020)).

Basically, the morphological classification of galaxies is manual and requires extensive use of human resources or from highly qualified professionals, or in some cases, amateur astronomers and volunteers (for example, the Galaxy Zoo project, GZ, Willett et al. (2013)). Current and near-term galaxy observational surveys are approaching the Exabyte scale multiwavelength databases of hundreds of millions of galaxies, which is impossible to classify manually (Vavilova et al. (2020b)). That magnifies the interest to use the alternatives in the form of machine learning (ML) techniques, including deep learning (DL), for classification of galaxies by their features.

We found that Support Vector Machine gives the highest accuracy for binary morphological classification with photometry-based approach, namely 96.1 % early E and 96.9 % late L types of galaxies (Vavilova et al. (2021b)). We exploited different galaxy classification techniques (human labeling, multi-photometry diagrams, and five supervised ML methods). The photometry-based approach was applied to the SDSS DR9 dataset, which contained $\sim 310\,000$ of galaxies with redshifts of $0.02 < z < 0.1$ and absolute stellar magnitudes of $-24^m < M_r < -19.4^m$ (see, more details about this sample, data cleaning, and

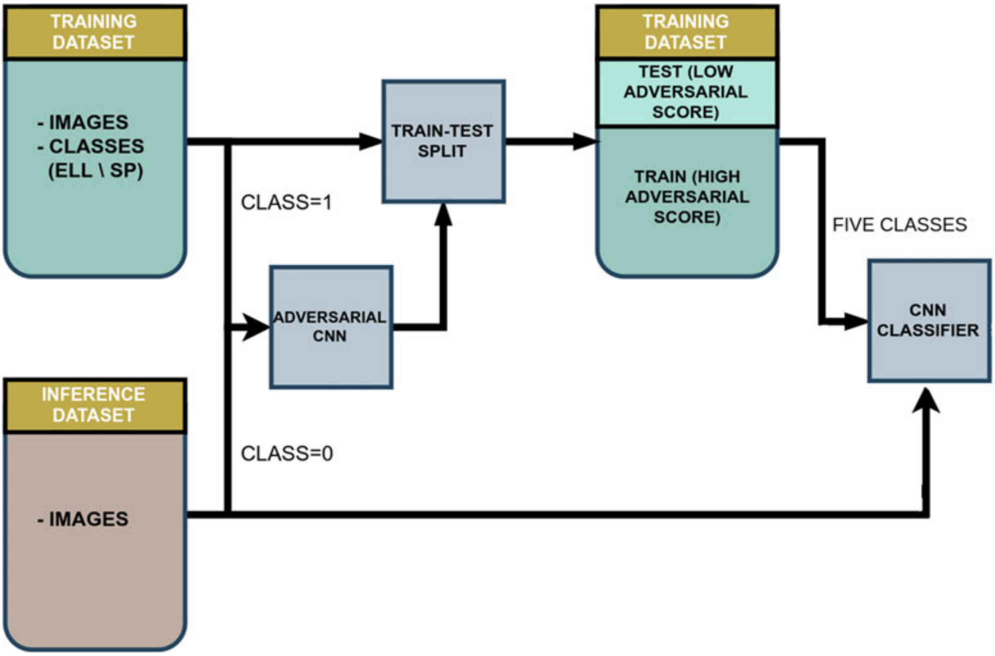

Figure 1. Scheme of our approach for morphological classification of galaxies with CNN.

correction procedures by Dobrycheva (2013); Dobrycheva et al. (2018); Vasylenko et al. (2019); Vasylenko (2020); Vavilova et al. (2021)). We tested such photometry parameters of galaxies as stellar magnitudes, color indices, inverse concentration indexes, which well correlate with the morphological type.

The aim of this work is to present the results of applying the machine learning model to the labeled images of galaxy dataset for prediction of the morphological type (morphological peculiarities) of galaxies.

2. Methods and Results

We continue to work with the above-mentioned sample, which included $\sim 310\,000$ galaxies (Dobrycheva (2013); Vavilova et al. (2021)). First of all, we have estimated how many galaxies from our sample belonged to the Galaxy Zoo 2 (GZ2) sample. We determined that more than a half of our sample do match with GZ2. So we divided our sample into two sub-samples (Fig. 1): **Training sample**: ∼170 000 galaxies (which do match GZ2 dataset); **Target sample (inference)**: ∼140 000 galaxies (which do not match GZ2 dataset).

Next, we used an adversarial neural networks to compare these two subsamples (Training and Inference). We trained the Convolutional Neural Network (CNN) on all the galaxy images of our sample, passing the class '0' for inference dataset and class '1' for the training one. The main idea of this procedure is to analyze potential differences between images from two datasets, because a classifier trained on one domain may behave incorrectly on another one. It turned out that the inference dataset can be easily separated from the GZ2 dataset with using `ResNet-50` CNN. The accuracy of such separation is $\approx 90\%$. This occurred mostly because the galaxies from the sample of $\sim 310\,000$ were pre-selected via $m_r < 17.7$ limitations by stellar magnitude in r-band. Analyzing properties of the galaxies from both datasets, we observed that the galaxies from the target (inference) sample are, on average, fainter and smaller (90% Petrosian flux) than

https://doi.org/10.1017/S1743921322001259 Published online by Cambridge University Press

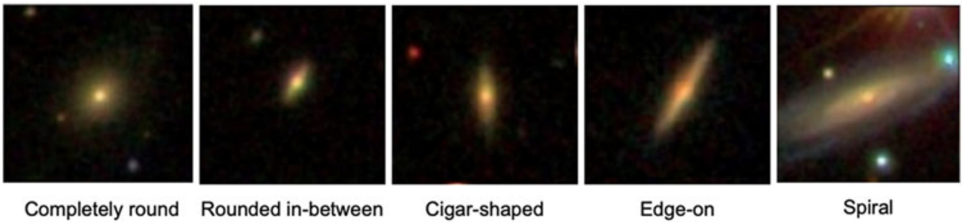

Figure 2. Examples of galaxy images from the GZ2 randomly selected from each of five morphological classes.

the galaxies from the training GZ2 sample. So, we did not use the same architectures for the adversarial CNN and final CNN classifiers. The main aim of adversarial validation is to investigate differences between training and inference galaxy samples.

So, according to the adversarial result, we can conclude that our training sample contains galaxies, properties of which are not common with inference one. This means that any validation of the morphological classifier has to be done with the galaxies from the training set, which have a low adversarial score. But we have found a way out of this obstacle.

We took into consideration only those galaxies from the training sample for which GZ2's volunteers gave the most votes for a more accurate result. It turned out to be $\sim 72\,000$ galaxies. To do so, we randomly choose $\sim 9\,000$ galaxies with adversarial score less than 0.7 from the training sample of $\sim 72\,000$ galaxies (comprising five different morphological classes, Fig. 2). Within this train-test split, the test part of training galaxies ($\sim 9\,000$) was used to validate the morphology by CNN classifier, and the rest part of galaxies ($\sim 63\,000$) to train the CNN. We have added the augmentation procedures to increase the validation sample for prediction of the classes of fainter and smaller galaxies.

Also, we did work on the selection of the best neural network for our task (described in more detail in the works by Khramtsov et al. (2019); Vasylenko (2020)). We used `DenseNet-201` and images of galaxies. Images were requested from the SDSS cutout server. We have retrieved RGB images composed of *gri* bands colour scaling, each of $100{\times}100{\times}3$ pixels (39.6×39.6 arcsec in each channel of the RGB image, respectively).

Confusion matrix for the classification CNN model for the pre-selected test sample of $\sim 9\,000$ galaxies is presented in Fig. 3. Each row represents the fraction of galaxies from a certain class classified as galaxies from other classes. One can see that the cigar-shaped sample has the lower accuracy. We see at least two reasons. First of all, the votes of volunteers for cigar-shaped galaxies were quite different, see, for example, their forum ZooUniverse. The opinions on classifying such a galaxy as the "flat but with an angle toward the tips and no structures like a bulge visible" are varied from the "cigar-shape elliptical" to the "spiral edge-on bulge". Secondly, this misclassification becomes more evident for small-sized galaxies at higher redshifts.

We have found that the inference catalog comprises 27 378 completely round, 59 194 round in-between, 18 862 cigar-shaped, 7 831 edge-on, and 23 119 spiral galaxies (see, examples, in Fig. 2).

3. Conclusion

Our approach and developed method, which includes CNN model and adversarial validation, allows one to classify galaxies with the SDSS images into five classes automatically (completely round, round in-between, cigar-shaped, edge-on, and spiral). It has the state-of-art performance giving $> 94\%$ of accuracy for all classes, except cigar-shaped galaxies ($\sim$87%). Each of the five classes has an uneven number of samples that could led to

https://doi.org/10.1017/S1743921322001259 Published online by Cambridge University Press

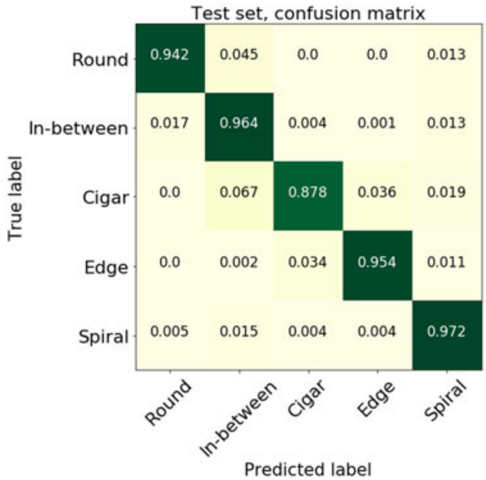

Figure 3. Confusion matrix for the classification CNN model for the pre-selected test sample of 9,000 galaxies. Each row represents the fraction of galaxies from a certain class (defined at the horizontal axis), classified as galaxies from other classes.

the categorical bias and measurable impact on overall accuracy. Due to the uneven class distributions, an overall accuracy metric may be not the best option. But we do not use it providing the accuracy scores per each class and noting that the accuracy is $> 94\%$ for all the classes except cigar-shaped galaxies. The preliminary visual inspection of classification shows the excellent agreement between the estimated classes and morphological parameters of the galaxies with their corresponding images.

4. Acknowledgements

This work was done in frame of the budgetary program "Support for the development of priority fields of scientific research" of the NAS of Ukraine ("e-Astronomy", CPCEL 6541230, 2020-2021). Vavilova I.B. thanks the Wolfgang Pauli Institute, Vienna, Austria, for the support in frame of "The Pauli Ukraine Project" (2022) under the "Models in plasma, Earth and space science" program.

References

Barrow, J.D. and Saich, P. 1993, *MNRAS*, 3, 717
Dobrycheva, D.V. 2013, *Odessa Astron. Publ.*, 26, 187
Dobrycheva, D.V., Vavilova, I.B., Melnyk, O.V. et al. 2018, *Kinemat. Phys. Cel. Bodies*, 34, 290
Elyiv, A.A., Melnyk, O.V., Vavilova, I.B. et al. 2020, *AA*, 635, A124
Khramtsov, V., Dobrycheva, D.V., Vasylenko, M.Y. et al. 2019, *Odessa Astron. Publ.*, 32, 21
Peng, Y., Lilly, S.J., Kovac, K. et al. 2010, *ApJ*, 721, 193
Reid, B.A., Samushia, L., White, M. et al. 2010, *MNRAS*, 426, 2719
Vasylenko, M. Y., Dobrycheva, D. V., Vavilova, I. B. et al. 2019, *Odessa Astron. Publ.*, 32, 46
Vasylenko, M., Dobrycheva, D., Khramtsov, V., et al. 2020, *Communications of BAO*, 67, 354
Vavilova, I., Dobrycheva, D., Vasylenko, M. et al. 2020, in: P. Skoda & F. Adam, *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (Elsevier), p. 307–323
Vavilova, I., Pakuliak, L., Babyk, I. et al. 2020, in: P. Skoda & F. Adam, *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (Elsevier), p. 57–102
Vavilova, I., Elyiv, A., Dobrycheva, D., Melnyk, O. 2021, in: I. Zelinka, M. Brescia, & D. Baron *Intelligent Astrophysics* (Springer), p. 57–79
Vavilova, I.B., Dobrycheva, D.V., Vasylenko, M.Y. et al. 2021, *AA*, 648, A122

Vavilova, I.B., Dobrycheva, D.V., Vasylenko, M.Yu. et al. 2018, *VizieR Online Data Catalog*, J/A+A/648/A122

Willett, K.W., Lintott, C.J., Bamford, S.P. et al. 2013, *MNRAS*, 435, 2835

Discussion

D. Bisikalo: I have question on accuracy of this method.

D. Dobrycheva: Accuracy classification of galaxies of five classes is 94 % for completely round, 96 % for round in-between, 87 % for cigar-shaped, 95 % for edge-on, and 97 % for spiral galaxies.

D. Bisikalo: Does it mean that 94 % of all objects are classified correctly?

D. Dobrycheva: Yes, correctly. For classification of features of galaxies is the same situation, likely for arms, bars, merging.

T. Hanawa: Could you show the confusion matrix again? The confusion comes from the data quiality or not? As you mentioned, some galaxies are less bright. And do you find any tendency that less bright galaxies are classified worse? I suppose that dim galaxies are hardly classified.

D. Dobrycheva: We had a big sample of 170000 galaxies from Galaxy Zoo. We have done data cleaning. Then we have 72000 galaxies, which were very well classified by volunteers. So, we didn't work with galaxies with low score. We selected randomly 9000 galaxies and used them as a test dataset for CNN.

T. Hanawa: I am interested in whether the confused galaxies are less bright or as bright as on average.

C. Boily: I have a similar question. I understand that eventually fainter galaxies should be more and more of those detected and, so, they will be more difficult to classify. I would imagine for all sorts of reasons, a lower signal noise and that kind of issue.

But my question was more about the SDSS dataset, which has something likely, I think, 4 or 5 wave bands. Are you able to remind what are the actual wave bands you used to define the reference datasets? And do you think that there are some biases (that are possible) if you choose one wave band, one color as opposed to another, for example, to look at the images of the galaxies because the morphology is depending on the color.

D. Dobrycheva: As about deep galaxies, we worked with a redshift less than 0.1 and in this range ...

T. Hanawa: I suppose that Chris means that irregular galaxies are bluer than round ones.

C. Boily: There would be a bias due to the color, for example, of the stellar population.