Morphological Classification of Galaxies From the SDSS Using Machine Learning Techniques

Clara Valle Gómez, Manuel Francisco González Penedo, and Minia Manteiga Outeiro

Laboratorio Interdisciplinar de Aplicaciones de la Inteligencia Artificial
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: clara.valle@udc.es

DOI: *https://doi.org/10.17979/spu.23.c62*

Abstract: Given the vast amount of labeled data from surveys like SDSS and Galaxy Zoo 2, AI has become essential for galaxy morphology classification. Challenges such as redshift, low resolution, and similar shapes make even expert labeling difficult. We developed a hierarchical two-stage CNN model to classify galaxy images. The first model categorized them into four classes, merging the underrepresented and visually similar cigar-shaped and edge-on types. A second submodel was then trained to distinguish between these two classes, using data augmentation to address class imbalance. Using these approaches, we achieved 95% accuracy, outperforming other research models using the same classes and demonstrating better classification accuracy and generalization on imbalanced datasets.

1 Introduction

Galaxies are vast systems composed of stars, gas, dust, and dark matter that evolve over billions of years. Their shapes and structures are not static, but change significantly over time due to both internal processes and external interactions. Traditionally, the classification of galaxy images was performed manually. However, given the vast number of images produced by modern telescopes, this manual task is no longer feasible and requires automatic classification methods.

Deep Learning (DL), a branch of Machine Learning (ML), has become the most common approach for creating automatic image classifiers. In particular, Convolutional Neural Networks (CNNs) have demonstrated strong potential for automatically learning complex features in images and exhibit good generalization capabilities.

Several studies have used CNNs to solve this task. For instance, Murrugarra LL. and Hirata (2017) evaluated a CNN to classify galaxies from GZ1 into two classes (ellipticals and spirals), achieving an accuracy of 90%. Vavilova et al. (2022), which serves as the baseline for our work, trained a CNN with GZ2 data and reported over 93% accuracy across five morphological classes. However, they observed lower performance in specific categories, such as cigar-shaped galaxies (75%) and edge-on galaxies (83%), likely due to class imbalance or visual ambiguity. Finally, Ma et al. (2022) introduced a hierarchical approach, training a 7-stage CNN on the GZ2 dataset to classify galaxies into five classes using One-Hot Encoding. Their model achieved an accuracy of 96% after more than 100 training epochs.

For this project, we aim to address the issues reported in the literature, caused by similarities in underrepresented classes, by employing a hierarchical approach, training specialized models, and applying machine learning techniques such as data augmentation.

2 Data preparation

The dataset used in this project is derived from the SDSS DR7 (Sloan Digital Sky Survey, Data Release 7) (Abazajian et al., 2009). The SDSS telescope captures large images of the sky (see Figure 1), from which celestial objects are then cropped.

The labeled data is obtained from Galaxy Zoo 2 (GZ2), a public survey that classifies images using a decision tree (Willett et al., 2013). By applying specific thresholds and a debiasing method (Hart et al., 2016), the most reliable responses were selected to create class labels for more than 304122 galaxies (see Figure 2).

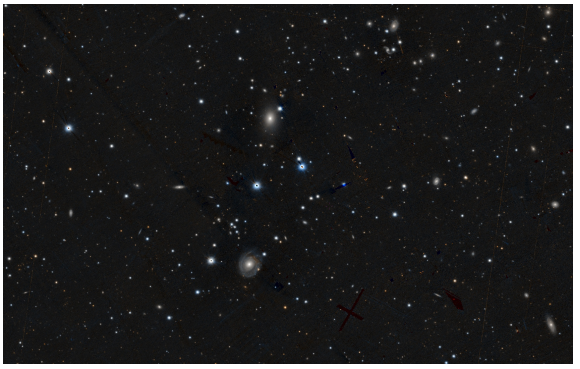

Figure 1: SDSS example image

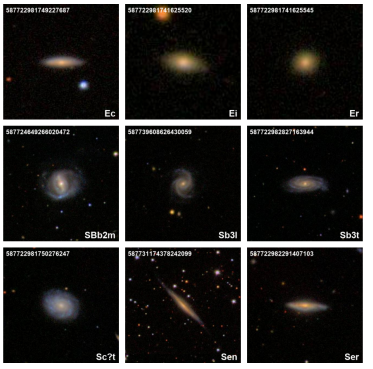

Figure 2: GZ2 classes example

Galaxies are cropped from the SDSS image and scaled to always represent around the same proportion in the center of the cutout. This samples are required to meet certain criteria to ensure data quality and reduce possible biases:

- **Limited aparent magnitude:** $m_r \leqslant 17.0$ mag to ensure bright enough galaxies
- **Limited angular size:** Petrosian radius $R_{90r} > 3''$ to ensure enough detail
- **Limited redshift:** $0.0005 < z < 0.25$, though most galaxies lie at $z \lesssim 0.1$

The classification data is downloaded from the Galaxy Zoo portal (GZ2 Table1)[1]. Cropped images and mapping file (which relates image name with SDSS galaxy ID) are downloadable from the Zenodo page[2].

The number of downloaded images is approximately 0.08% fewer than the total entries in the GZ2 table. The project documentation indicates these missing images are not systematically chosen, suggesting the dataset remains suitable for unbiased analysis, but it will require extra preprocessing.

2.1 Dataset generation

The dataset is preprocessed and cleaned by removing the GZ2 table entries that do not correspond to any image, in order to avoid issues during image handling.

After the cleaning step, a five-class morphological scheme from Vavilova et al. (2022) is applied to generate the dataset. Each classification requires both debiased vote fractions exceeding specified thresholds and minimum vote counts to ensure statistical reliability, as shown in Table 1.

[1] https://data.galaxyzoo.org/
[2] https://zenodo.org/records/3565489

Class	Criteria	N_samples
0 - Completely Round	Smooth debiased > 0.469 Completely round debiased > 0.469	31294
1 - Rounded In Between	Smooth debiased > 0.469 In-between roundness debiased > 0.5	36586
2 - Cigar-Shaped	Smooth debiased > 0.469 Cigar-shaped debiased > 0.5	4909
3 - Edge-On Disk	Disk features debiased > 0.43 Edge-on detection debiased > 0.602	4342
4 - Spiral Disk	Disk features debiased > 0.43 Not edge-on debiased > 0.715 Spiral structure debiased > 0.619	24003

Table 1: Classification criteria from Vavilova et al. (2022), based on GZ2 (Willett et al., 2013) selection thresholds for each class (all > 25 votes), along with the resulting number of samples per class.

The resulting dataset consists of **101,134 galaxies**. Table 1 also shows the class distribution. A clear imbalance can be observed, with classes 2 and 3 being underrepresented relative to the others.

As previously discussed in the literature by Vavilova et al. (2022), these underrepresented classes present a problem, not only because of the limited number of samples, but also due to their morphological similarity (see Figure 3). To mitigate this issue, we merged these two classes into a single class "cigar+edgeon," which a specialized model will later attempt to separate.

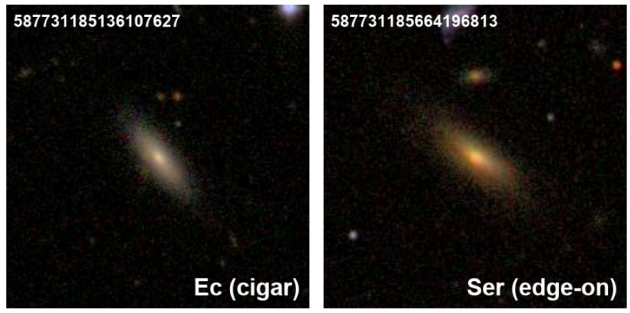

Figure 3: Comparison between cigar and edge-on classes

3 CNN architecture and training

For this project, we trained a total of 2 CNN models.

The first one, which we refer to as "main model" predicts 4 classes, since we merged conflictive classes into one class (cigar+edgeon). This CNN is described as follows:

The chosen architecture can be seen in Figure 4. It uses ReLU as activation function and includes a dropout layer before the output layer with rate 0.5 to reduce overfitting. For the main model, a *softmax* function is used in the last layer to produce a probability distribution over $K = 4$ classes. The model is trained using the Adam optimizer with sparse categorical cross-entropy loss and max 15 epochs (with early stopping).

To perform training, evaluation and testing, the clean dataset is split as follows:

- **Training set**: 70% (70793 images)

- **Validation set**: 15% (15170 images)
- **Test set**: 15% (15170 images)

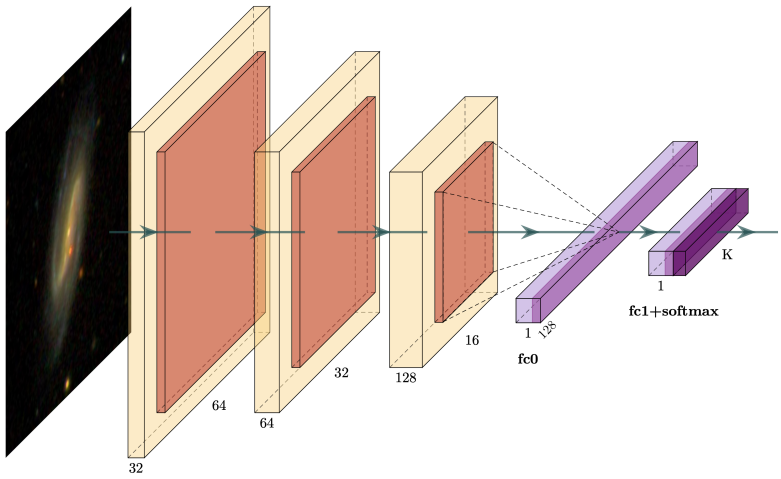

Figure 4: General model architecture. It consists of 3 convolutional layers (yellow) followed by max pooling layers (orange) to reduce spatial dimension. The resulting feature maps are flattened and passed through fully connected layers (purple), with the final layer as the output layer with K neurons.

The second model is designed to separate the two previously merged classes. It is trained exclusively on images from class 2 (cigar-shaped) and class 3 (edge-on).

The main differences, compared to the previous model, are the following: the output layer consists of a single neuron with *sigmoid* activation, producing a binary output (0 for cigar-shaped and 1 for edge-on). Consequently, the loss function is changed to binary cross-entropy. The maximum number of training epochs is increased to 20 to improve convergence. In addition, a data augmentation strategy is applied to artificially increase the number of training samples, by randomly rotating, flipping, and zooming the images.

4 Hierarchical classification

Once the models are trained, we use them to perform inference, which means applying the pre-trained models to make predictions on new data.

Using the already trained models, the inference pipeline is applied to an unseen testing dataset. These images are fed into the main model and the images predicted as "cigar+edgeon" are the input for the binary model. In this specialized model, predictions closer to 0 are associated with the "cigar" class, while values closer to 1 correspond to the "edge-on" class. A decision threshold of 0.5 is applied to convert the predicted probability into a discrete class label.

5 Results and Conclusions

The results of each individual model are presented in the following classification reports: the main model (Table 2) and the binary model (Table 3). Compared to the accuracies reported by Vavilova et al. (2022) (75% for the cigar-shaped class and 83% for the edge-on class), our

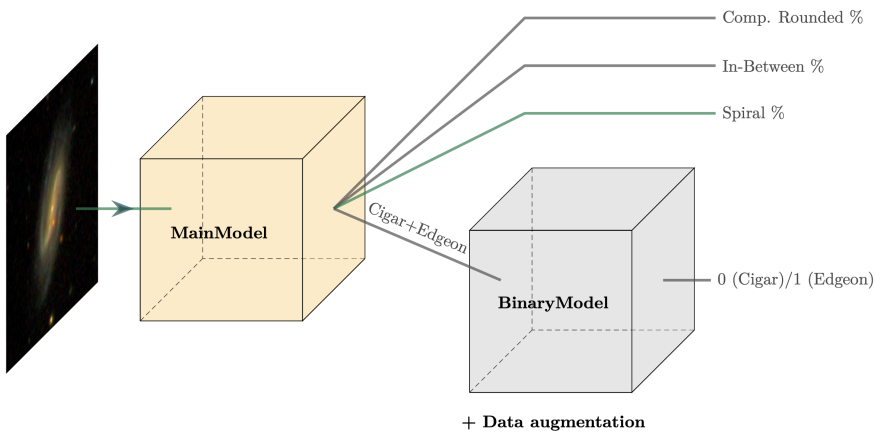

Figure 5: Hierarchical model architecture. The boxes represent the previously described CNNs

models achieve an accuracy improvement of 20% for the cigar-shaped class and 14% for the edge-on class (see Table 3).

Resulting in an **overall accuracy of 95**%

5.1 Comparison with literature

We achieved our goal of improving classification accuracy, reaching 95%. Although Ma et al. (2022) report a slightly higher accuracy (96%), their approach relies on a much more complex seven-stage convloutional layers + extra final layers over more than 100 epochs. In contrast, our results were obtained with a significantly simpler network. We deliberately accepted this minor reduction in accuracy to maintain a lightweight model that requires fewer computational resources, promoting a more ethical and environmentally responsible use of AI.

5.2 Model Attention Visualization

For many people, CNNs are considered to be a black box with weak interpretability. For example, why it predicts as it does and where it focuses are usually unknown. Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2019) is a visualization method based on gradient localization. It uses the gradient of the target to produce a heatmap, showing where the model pays more attention, and proving that our CNN is correctly focusing on galaxy features and not leaning noise, as shown in Figure 6.

Class	Precision	Recall	F1-score	Support
0 - Comp. Rounded	0.96	0.96	0.96	4695
1 - In Between	0.96	0.95	0.95	5488
2 - Cigar+Edge-on	0.95	0.97	0.96	1387
3 - Spiral	0.96	0.97	0.96	3601
Accuracy			**0.96**	15171
Macro avg	0.96	0.96	0.96	15171
Weighted avg	0.96	0.96	0.96	15171

Table 2: Independent classification report of the main model

Class	Precision	Recall	F1-score	Support
0 - Cigar	0.95	0.94	0.95	737
1 - Edge-on	0.94	0.94	0.94	651
Accuracy			**0.94**	1388
Macro avg	0.94	0.94	0.94	1388
Weighted avg	0.94	0.94	0.94	1388

Table 3: Independent classification report of the binary model

Author	Dataset	Accuracy
Murrugarra LL. and Hirata (2017)	GZ1	90%
Vavilova et al. (2022)	GZ2	93%
Ma et al. (2022)	GZ2	96%
Ours	GZ2	95%

Table 4: Comparison of deep learning approaches in galaxy classification.

5.3 Conclusions

This work addresses previously identified challenges such as class similarity and class imbalance in cigar-shaped and edge-on galaxies. We developed a simple CNN model that outperforms others in the literature with similar characteristics by employing a hierarchical approach and data augmentation. Furthermore, using Grad-CAM, we demonstrated that our model learns the actual features of the galaxies and is therefore able to generalize well. The techniques applied in this study can also be extended to a broader range of astronomical data processing problems related to class similarity and imbalanced datasets.

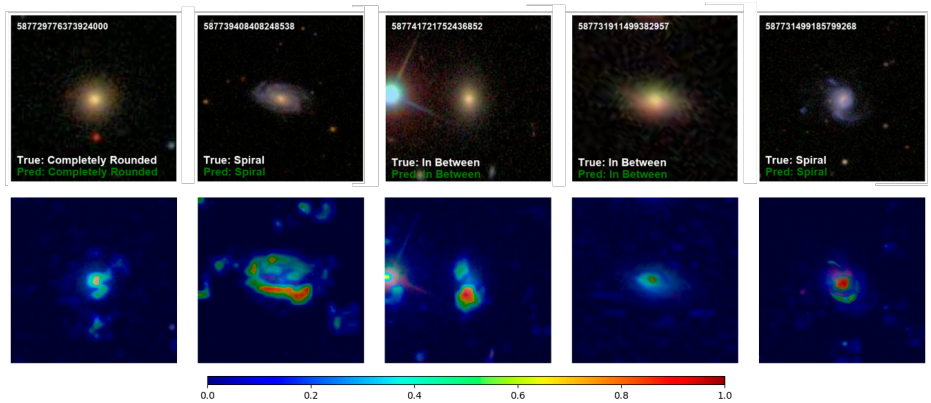

Figure 6: Grad-CAM results for our CNN architecture. The bar indicates Grad-Cam activation intensity (0=low, 1=high).

Bibliography

K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, and et al. The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement*, 182(2):543–558, June 2009.

R. E. Hart, S. P. Bamford, K. W. Willett, K. L. Masters, C. Cardamone, C. J. Lintott, R. J. Mackay, R. C. Nichol, C. K. Rosslowe, B. D. Simmons, R. J. Smethurst, and et al. Galaxy zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, 07 2016. URL *https://doi.org/10.1093/mnras/stw1588*.

X. Ma, X. Li, A. Luo, J. Zhang, and H. Li. Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing. *Monthly Notices of the Royal Astronomical Society*, 519(3):4765–4779, 12 2022. URL *https://doi.org/10.1093/mnras/stac3770*.

J. H. Murrugarra LL. and N. S. T. Hirata. Galaxy image classification. *Astronomy and Astrophysics*, 2017. URL *http://sibgrapi.sid.inpe.br/attachment.cgi/sid.inpe.br/sibgrapi/2017/09.10.23.00/doc/article.pdf*.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019. URL *http://dx.doi.org/10.1007/s11263-019-01228-7*.

I. B. Vavilova, V. Khramtsov, D. V. Dobrycheva, M. Y. Vasylenko, A. A. Elyiv, O. V. Melnyk, and et al. Machine learning technique for morphological classification of galaxies from sdss. ii. the image-based morphological catalogs of galaxies at 0.02¡z¡0.1. *Kosmìčna nauka ì tehnologìâ*, 28(1):03–22, Feb. 2022. URL *http://dx.doi.org/10.15407/knit2022.01.003*.

K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, and et al. Galaxy zoo 2: detailed morphological classifications for 304.122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4): 2835–2860, Sept. 2013. URL *http://dx.doi.org/10.1093/mnras/stt1458*.