Classifying galaxies with deep learning: Using Images from the Sloan Digital Sky Survey for Accurate Classification

Norman Hoever[1]*

[1]Faculty of Communication and Environment, Rhine-Waal University of Applied Sciences

Abstract

The classification of galaxy images is an important step in addressing the growing volume of data generated through astronomical observations. The following paper describes a convolutional neural network architecture that has been designed to categorize images of galaxies into one of ten possible categories. The training was performed on a set of images of galaxies obtained from the Sloan Digital Sky Survey (SDSS), which is one of the largest sky survey projects. The classifications made by volunteers of the Galaxy Zoo Citizen Science project cover all galaxy types, which is why this data is used as the basis for the training. The model is based on a deep neural network, in particular Convolutional Neural Networks (CNNs). It achieves an accuracy of 90.3%. The use of machine learning in astronomy is not only possible but also effective in automating the analysis of astronomical data and reducing the time required for research.

1 Introduction

Morphological classification of galaxies is a basic technique in astronomy since it offers information about the physical characteristics, formation and evolution of galaxies. The first serious effort to categorize galaxies based on their structural features was made by Edwin Hubble in 1926 [1]. His Hubble sequence diagram, also known as Hubble tuning fork diagram, divides galaxies mainly into elliptical, spiral and lenticular galaxies and its impact is still being felt in the field of astronomy.

In the past, the classification of galaxies was done by astronomers by observing the galaxies and comparing them with the different categories laid down by Hubble; however, this is not feasible in the present world due to the large number of galaxies discovered in the recent sky surveys. Manual classification is a time consuming and a subjective process as different astronomers may classify it differently.

In this regard, the use of machine learning in the automation of the classification of galaxies is gaining importance. The use of machine learning, particularly deep learning, provides the opportunity to analyze large amounts of data quickly and accurately. This automation makes it possible to identify patterns and structures in the data that may be unnoticed by human observers, making the results more precise.

One of the most impressive projects that can demonstrate the necessity of and possibility for the automated classification of galaxies is the Galaxy Zoo. Since 2007, hundreds of thousands of volunteers have classified millions of galaxies [2] as part of the Sloan Digital Sky Survey (SDSS). This citizen science initiative has produced a vast amount of classified data that can be used as a training and test dataset for machine learning algorithms.

Automating the classification of galaxies means that more data can be processed and understood in less time, allowing astronomers to better understand the structure and evolution of the universe. This paper demonstrates that AI-based methods are an essential tool to handle the flood of data in modern astronomy.

2 Related Work

The task of the automated classification of galaxy images has attracted more attention in the scientific community in the recent past. Several studies have shown that machine learning can be useful in achieving this goal.

Some studies so far consisted of classifying galaxies into two to five classes [3] [4]. This amount of classes for the classification has the advantage that the accuracy of the model can be higher than models with more classes. Some studies using this amount of classes have achieved a high degree of accuracy, but as a result, galaxy images are only classified into a few classes. This may not be sufficiently detailed for some applications [4].

Some progress in this direction has been made by researchers that seeks to provide a more detailed classification of galaxies based on their morphologies. Specifically, one of the papers employed ten categories of galaxies and obtained the accuracy of 84.73% [5]. This study has therefore also shown that

*Corresponding author: Norman.Hoever@protonmail.com

it is possible to make more detailed categorizations and still achieve an acceptable level of accuracy.

3 Data preparation and model training

3.1 Data Processing

Classified galaxy samples from two different catalogs were used for this project: the Galaxy Zoo DESI catalog [6] [7] [8] and the Galaxy Zoo DECaLS catalog [9]. Some classes were underrepresented in the Galaxy Zoo DECaLS catalog, so galaxy images from the DESI catalog were merged with the same classes. This method increased the number of samples for the underrepresented classes by a factor of two. The galaxy images and the corresponding tables of the voluntary classifications were downloaded from the respective publications. When assigning the galaxy images to the classes, the class with the highest approval among the volunteer ratings was selected. To ensure that the classifications were reliable, an additional condition was introduced: 75% of the volunteers had to vote for at least this class.

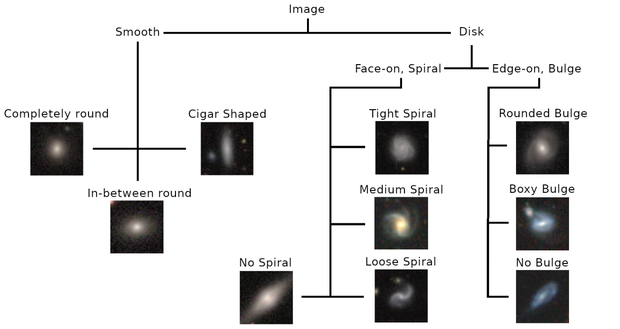


Figure 1: *Tree Diagram of the Classes used from the Datasets*

With the given classifications of the Galaxy Zoo some adjustments still need to be made in order to obtain classes that are comparable with the 10 classes of the study [5] already mentioned (Figure 1). First, a split must be made between smooth and disk shapes. The Smooth category includes galaxies that look more elliptical and even. The Disk category includes galaxies that have a disk-like and flat structure.

The first 3 classes belong to the Smooth category.

- Smooth, Completely round
- Smooth, In-between round
- Smooth, Cigar shaped

The remaining 7 classes are again separated between Bulge and Spiral.

- Disk, Edge-on, Rounded Bulge
- Disk, Edge-on, Boxy Bulge
- Disk, Edge-on, No Bulge
- Disk, Face-on, Tight Spiral
- Disk, Face-on, Medium Spiral
- Disk, Face-on, Loose Spiral
- Disk, Face-on, No Spiral

To ensure the efficiency of the model and the training process, the galaxy images were first cropped to 207x207 pixels and then downscaled to 69x69 pixels (Figure 2).

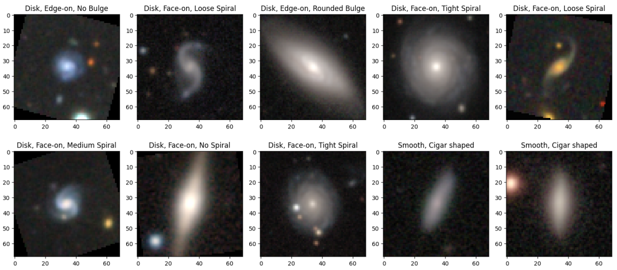

Disk, Edge-on, No Bulge
Disk, Face-on, Loose Spiral
Disk, Edge-on, Rounded Bulge
Disk, Face-on, Tight Spiral
Disk, Face-on, Loose Spiral
Disk, Face-on, Medium Spiral
Disk, Face-on, No Spiral
Disk, Face-on, Tight Spiral
Smooth, Cigar shaped
Smooth, Cigar shaped
0
10
20
30
40
50
60
0
20
40
60

Figure 2: *Example Images with assigned Class in the Dataset*

These steps helped to reduce the size of the input data while providing enough detail to enable accurate classification.

3.2 Data Distribution

The whole data set, consisting of 15,406 galaxy samples, was split into 80% training data and 20% test data to provide a solid basis for model evaluation. This results in 12,325 training samples and 3,081 test samples (Figure 3).

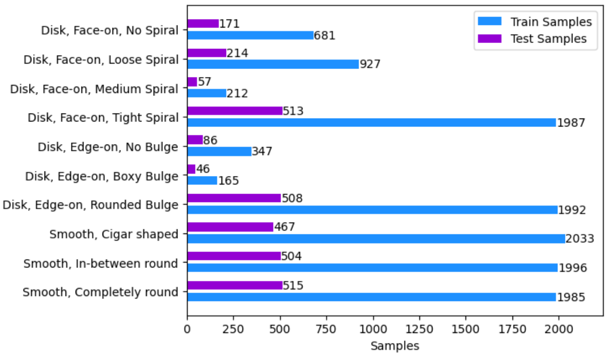


Figure 3: *Number of Samples based on the Distribution of Classes, Training and Test Data*

One of the main issues when it comes to data distribution was the fact that some classes in the dataset had very few samples. This can have a great impact on the performance of the model as some of the classes may not have enough training samples. To overcome this problem, data augmentation was

Table 1: *Model Architecture - In F = Input Features, Out F = Output Features*

Layer No.	Layer Type	Input Channels	Output Channels	Kernel Size	Features	Others
1	Convolution	3	16	5 x 5		
2	Batch Normalization				16	
3	ReLU					
4	Max Pooling			2 x 2		
5	Convolution	16	32	5 x 5		
6	Batch Normalization				32	
7	ReLU					
8	Max Pooling			2 x 2		
9	Convolution	32	64	5 x 5		
10	Batch Normalization				64	
11	ReLU					
12	Max Pooling			2 x 2		
13	Flatten					
14	Dropout					p = 0.5
15	Fully Connect					In F = 1600, Out F = 10

used to add more samples in these classes which were limited in number.

Data augmentation involves creating new images from the existing ones through techniques like rotation, flipping and changes in brightness levels. These methods generate new, synthetic patterns of the original images, artificially enlarging the data and simultaneously increasing the variability of the training data.

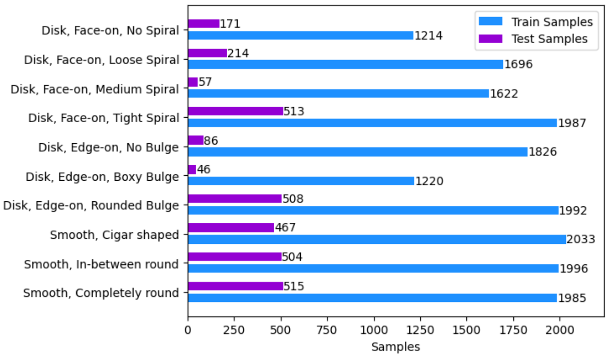


Figure 4: *Number of Samples after Data Augmentation based on the Distribution of Classes, Training and Test Data*

It was effective in increasing the ratio of all classes in the training set to ensure that none was underrepresented (Figure 4). However, it should be noted that augmented data, especially in this high quantity, is not comparable to original samples. With the same amount of samples, the class with a high amount of augmented data can still perform worse. As a result, the number of training samples increased to 17,571. This made it possible to make sure that the model not only identified the classes that occurred most often but also the classes that were less frequent.

3.3 Model

The model (Table 1) used in this paper is based on a feature extraction model and a classifier (Figure 5). The feature extraction model contains several convolutional blocks. With three convolutional blocks, the features of the input images can be extracted and processed. These blocks each consist of four layers. Convolution, Batch Normalization, ReLU and the Max Pooling layer.

Having processed the feature maps, they are converted into a one-dimensional vector using a flatten layer. To avoid overfitting, a dropout layer removes 50% of random input nodes during training.

Lastly, the output is passed to the classifier (Figure 5), which has 10 nodes as output. These represent the 10 classes for the classification.

3.4 Training and Testing

The parameters used for training include a batch size of 512, an initial learning rate of 0.001 and a training duration of 50 epochs. A learning rate scheduler was used to reduce the learning rate over the entire training iterations. This combination of parameters proved to be a good basis for training the model.

The training of the model was performed on an NVIDIA Tesla T4 GPU at Google Colab. This GPU made it possible to train the model efficiently and relatively quickly.

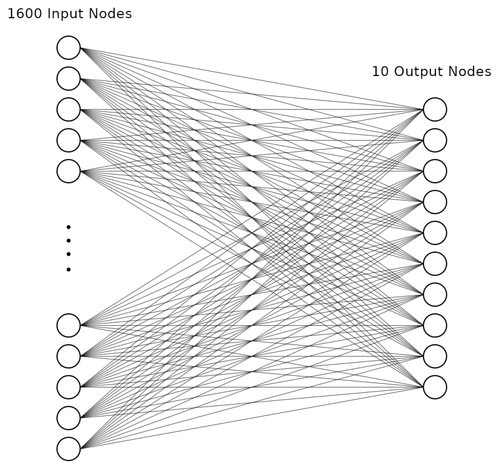


Figure 5: *Illustration of the Classifier Network*

4 Results

Based on the trained model for galaxy image classification, it can be concluded that the model has high accuracy and stability.

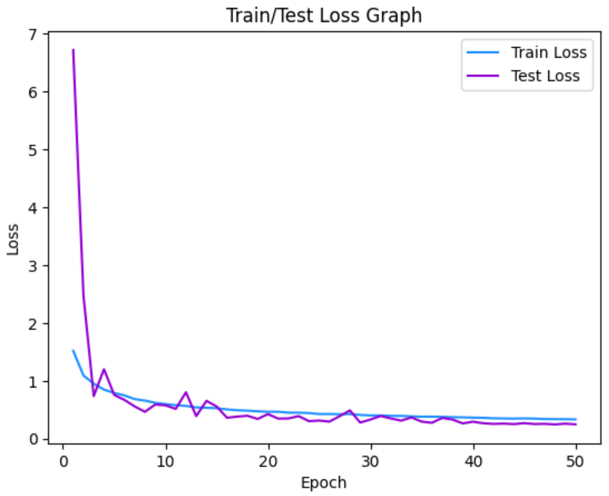


Figure 6: *Graph of Training and Test Loss in Relation to the Number of Epochs*

The loss graph for training and testing is illustrated in Figure 6. Here it can be observed that the training loss (blue line) and the test loss (purple line) are decreasing sharply and then plateau after about 10 epochs. The reduction of the loss values indicates that the model is learning and the mistakes made during the training process are gradually reduced.

Figure 7 presents the accuracy of the training and testing processes for 50 epochs. The training accuracy (blue line) and the test accuracy (purple line) rise gradually, and the test accuracy fluctuates but generally rises. By the 50th epoch, the test accuracy gets to a value of 90.3%.

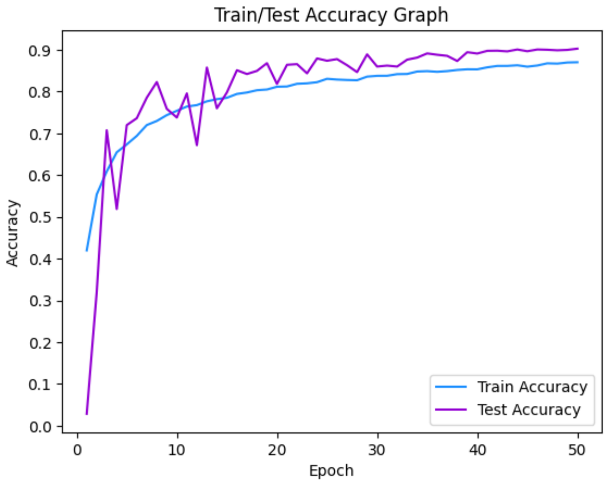
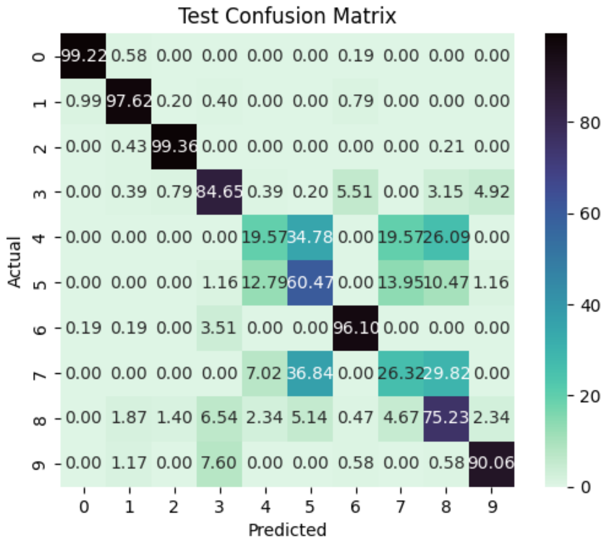


Figure 7: *Graph of Training and Test Accuracy in Relation to the Number of Epochs*



	0	1	2	3	4	5	6	7	8	9
0	99.22	0.58	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00
1	0.99	97.62	0.20	0.40	0.00	0.00	0.79	0.00	0.00	0.00
2	0.00	0.43	99.36	0.00	0.00	0.00	0.00	0.00	0.21	0.00
3	0.00	0.39	0.79	84.65	0.39	0.20	5.51	0.00	3.15	4.92
4	0.00	0.00	0.00	0.00	19.57	34.78	0.00	19.57	26.09	0.00
5	0.00	0.00	0.00	1.16	12.79	60.47	0.00	13.95	10.47	1.16
6	0.19	0.19	0.00	3.51	0.00	0.00	96.10	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	7.02	36.84	0.00	26.32	29.82	0.00
8	0.00	1.87	1.40	6.54	2.34	5.14	0.47	4.67	75.23	2.34
9	0.00	1.17	0.00	7.60	0.00	0.00	0.58	0.00	0.58	90.06

Figure 8: *Confusion Matrix for the Test Dataset*

The confusion matrix of the test set is depicted in Figure 8 with the values expressed in percentage. This matrix provides a clear understanding of the performance of the model in terms of the accuracy of the 10 classes. A high correlation between the actual and predicted values is observed in most classes, which underscores the accuracy of the model. However, some classes, especially class 5, 6 and class 8, have stronger misclassification, which is due to the fact that these three classes have limited samples. More samples for these classes from for instance other datasets could be incorporated in the future work to fine tune the model.

Compared to another paper [5] that also uses 10 classes for the classification of galaxies and obtains an average accuracy of 84.5%, and the model that is presented in this paper achieves even better accuracy, 90.3%. The improvement shown here can be credited to the model architecture and training techniques, including the choice of training parameters and data augmentation.

5 Conclusion

Usually, galaxy images from the Sloan Digital Sky Survey (SDSS) are classified by astronomers and other volunteers, but this is no longer feasible with the present large datasets. The solution to this problem is machine learning, which can classify images of galaxies fast and accurately.

This paper proposes an machine learning model for classifying galaxies into ten types based on images from the Sloan Digital Sky Survey (SDSS). The model is based on a Convolutional Neural Networks (CNNs) and reaches an accuracy of 90.3%.

References

[1] E. P. Hubble. "Extragalactic nebulae". In: *Astrophysical Journal* 64 (Dec. 1926). DOI: 10.1086/143018.

[2] Zooniverse. *Galaxy Zoo The Story So Far*. 2007. URL: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/results (visited on 06/10/2024).

[3] Mike Walmsley et al. "Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning". In: *Monthly Notices of the Royal Astronomical Society* 491.2 (Oct. 2019), pp. 1554–1574. ISSN: 0035-8711. DOI: 10.1093/mnras/stz2816. URL: https://doi.org/10.1093/mnras/stz2816.

[4] Mitchell K. Cavanagh, Kenji Bekki, and Brent A. Groves. "Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs". In: *mnras* 506.1 (Sept. 2021), pp. 659–676. DOI: 10.1093/mnras/stab1552.

[5] Sarvesh Gharat and Yogesh Dandawate. "Galaxy classification: a deep learning approach for classifying Sloan Digital Sky Survey images". In: *Monthly Notices of the Royal Astronomical Society* 511.4 (Feb. 2022), pp. 5120–5124. ISSN: 1365-2966. DOI: 10.1093/mnras/stac457. URL: http://dx.doi.org/10.1093/mnras/stac457.

[6] Mike Walmsley et al. *Galaxy Zoo DESI: Detailed Morphology Classifications for 8.7M Galaxies in the DESI Legacy Imaging Surveys*. Version 1.0.1. Zenodo, Sept. 2023. DOI: 10.5281/zenodo.8360385. URL: https://doi.org/10.5281/zenodo.8360385.

[7] M. Walmsley et al. "Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys". In: *Monthly Notices of the Royal Astronomical Society* 526.3 (2023), pp. 4768–4786. ISSN: 0035-8711. DOI: 10.1093/mnras/stad2919.

[8] A. Dey et al. "Overview of the DESI Legacy Imaging Surveys". In: *Astronomical Journal* 157.5 (2019), p. 168. ISSN: 1538-3881. DOI: 10.3847/1538-3881/ab089d.

[9] Mike Walmsley et al. *Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies*. Version 0.0.2. Zenodo, June 2021. DOI: 10.5281/zenodo.4573248. URL: https://doi.org/10.5281/zenodo.4573248.