Università degli Studi di Padova

Galaxy Classification using Convolutional Neural Networks

G. Bellavia, L. Belli, M. Lera, G. A. Maida

June 9, 2025

Table of Contents

Table of Contents

Introduction

Galaxy morphology is key to understanding galaxy evolution, but

Scaling Problem

With new mission and surveys (JWST, Euclid) automated machine learning approaches are now **indispensable** to scale further.

Goal

Explore DL techniques to automate morphological classification using data from Galaxy Zoo 2.

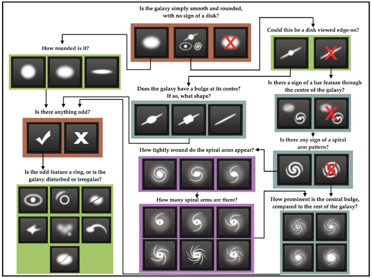
Table of Contents

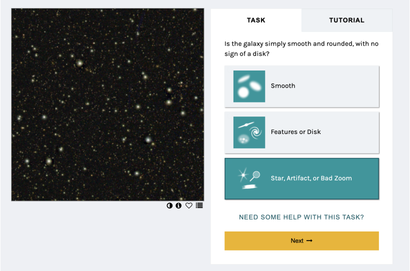
Dataset

We dealt with a subset of GZ2, from the challenge hosted on Kaggle in 2013.
The dataset is composed by 424×424px RGB images, with target centered in it, divided in

- 60 000 labeled images
- 80 000 unlabeled images

Labels consist of 37 probability values corresponding to morphological features defined in the Galaxy Zoo decision tree.





Why CNNs?

- Spatial awareness: preserving spatial structure of the image.
- Efficiency: reduced number of parameters
- Global image processing: analyzing the image as a whole.
- Automatic feature extraction: the learning process of features is data-driven.

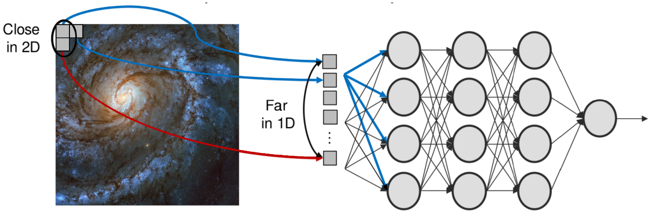
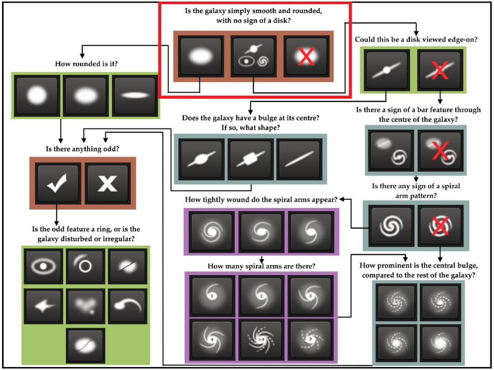


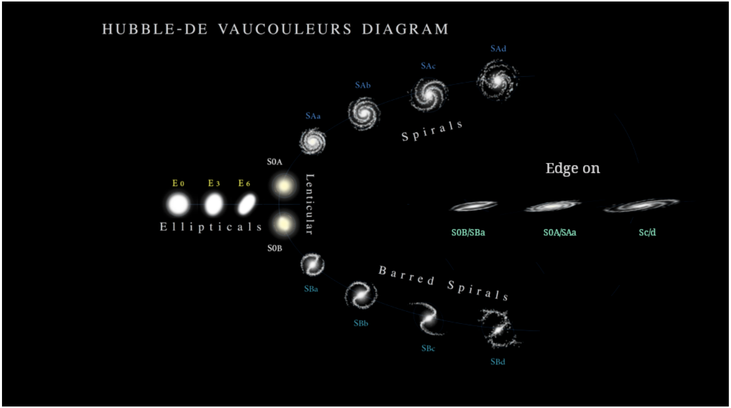
Table of Contents

Original labels



GalaxyID	Class1.1	Class1.2	Class1.3	...
100078	0.693377	0.238564	0.068059	...
100128	0.687783	0.288344	0.023873	...
100143	0.269843	0.730157	0.000000	...
100335	0.165002	0.834998	0.000000	...
100367	0.471429	0.512708	0.015863	...
100402	0.610000	0.390000	0.000000	...
100541	0.445052	0.533256	0.021693	...
100673	0.044846	0.955154	0.000000	...
100724	0.862362	0.137638	0.000000	...
100836	0.418589	0.557830	0.023581	...
101127	0.361844	0.623873	0.014283	...
101151	0.644645	0.222222	0.133133	...
101324	0.444025	0.555975	0.000000	...
101355	0.315310	0.684690	0.000000	...
101371	0.618786	0.312107	0.069107	...
101375	0.394491	0.586077	0.019431	...
101548	0.750104	0.155924	0.093971	...
...	...	...	...	...

Modified Decision Tree



Revised Labels

GalaxyID	Class1.1	Class1.2	Class1.3	Class2.1	Class2.2	Class3.1	Class3.2	Class4.1	Class4.2	...
100078	0.693377	0.238564	0.068059	0.000000	0.238564	0.109493	0.129071	0.189098	0.049466	...
100128	0.687783	0.288344	0.023873	0.000000	0.288344	0.069098	0.219246	0.000000	0.288344	...
100143	0.269843	0.730157	0.000000	0.730157	0.000000	0.000000	0.000000	0.000000	0.000000	...
100335	0.165002	0.834998	0.000000	0.235325	0.599673	0.148674	0.450999	0.378226	0.221447	...
100367	0.471429	0.512708	0.015863	0.028662	0.484046	0.217021	0.267025	0.265796	0.218250	...
...	...	...	...	...	...	...	...	...	...	...

GalaxyID	E0	E3	E6	S0a_eon	SB0a_eon	Scd_eon	SoB	SoA	SAa	...
100078	$\boxed{0.408599}$	0.284778	0.000000	0.000000	0.000000	0.000000	0.022703	0.026763	0.053726	...
100128	0.482483	0.205300	0.000000	0.000000	0.000000	0.000000	0.069098	0.219246	0.000000	...
100143	0.000000	0.000000	0.269843	0.561429	0.000000	0.168728	0.000000	0.000000	0.000000	...
100335	0.000000	0.123752	0.041251	0.078442	0.078442	0.078442	0.054902	0.166545	0.000000	...
100367	0.026338	0.445091	0.000000	0.028662	0.000000	0.000000	0.097852	0.120398	0.023425	...
...	...	...	...	...	...	...	...	...	...	...

GalaxyID	predicted_label	true_label
100078	$\boxed{\text{E0}}$	E0
100128	E0	E0
100143	S0a_eon	S0a_eon
100335	SoB	SoA
100367	E3	E3
100402	SoA	E3
100541	E3	SoA
100673	SAc	SAc
100724	E0	E3
100836	S0a_eon	E6
101127	S0a_eon	Scd_eon
101151	E3	E0
101324	SoA	E3
101355	S0a_eon	S0a_eon
101371	E0	E0
101375	E3	E3
101548	E0	E0
...	...	...

Setup

Università degli Studi di Padova

Preprocessing steps applied to input images:

- Cropped borders to reduce size to 324×324
- Resized to 128×128 pixels
- Applied random rotation
- Converted to grayscale

Preprocessing applied to test images:

- kept only 'clear' images (maximum value higher than 10% than other values)

We set up the problem as a **regression**. The loss used is the **RMSE**.

Table of Contents

Architectures

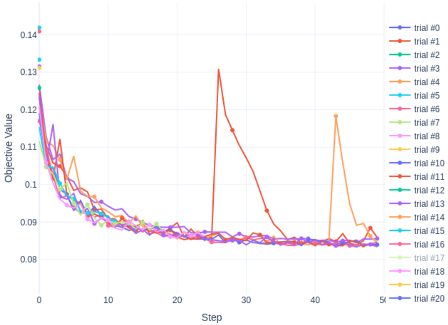
5 architectures:

- **VGG**: little variation on the standard
- **JAGZoo**: 5 convolutional layers (C) + 3 fully connected (FC)
- **PADel**: 7 C + 3 FC
- **PC**: 1 C + 3 FC
- **SKyNet**: 7 C + 4 FC

Tuning with OPTUNA

Automated hyperparameter search using `optuna` framework. We searched

- **Optimizer**: Adam variants, RMSprop, SGD
- **Activation function**: ReLU, Leaky ReLU
- **Learning rate** and momentum when needed
- **Weight initialization**: He initialization or `PyTorch` standard



Loss Curves

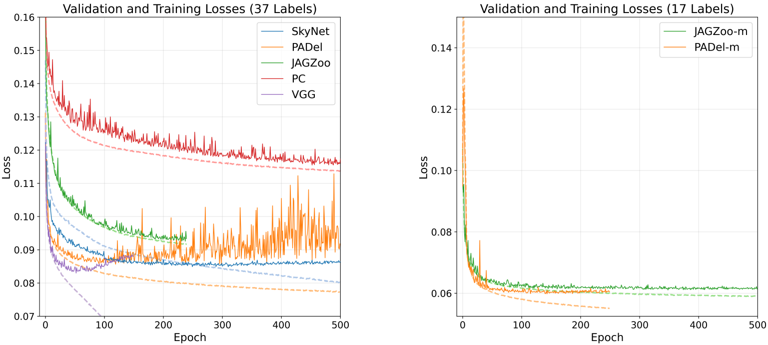


Figure: Training (dashed lines) and validation (solid lines) losses for the trained models.

Table of Contents

Confusion Matrices

UNIVERSITÀ DEGLI STUDI DI PADOVA

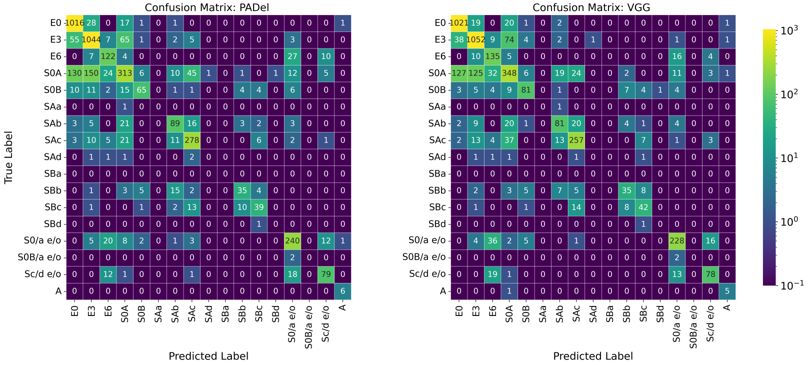
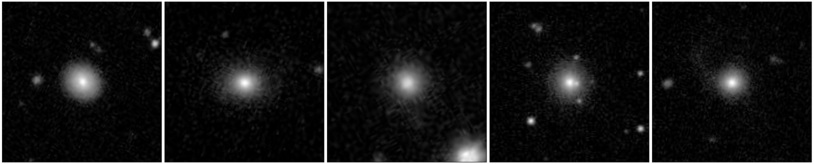
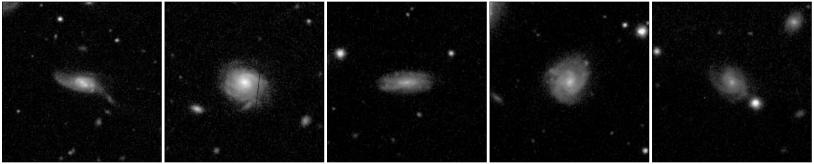


Figure: Confusion matrices for PADel and VGG. In each row is summarized how the images from each class are classified by the CNN. In particular, the diagonal contains the correctly classified ones.

Classification Test



(a) E0 galaxies.



(b) SAc galaxies.

Figure: PADel classification of unlabeled galaxies.

Table of Contents

Conclusion

The CNNs were able to proficiently classify the galaxies. It performed worse in recognising lenticular galaxies and faint features like bars.

Future work

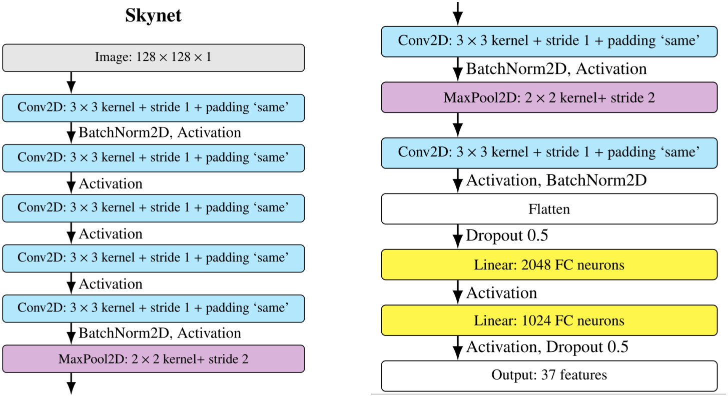
- CNNs architectures
- RGB channels
- Standardization
- Expand the classification

Summary

- We followed a deep learning approach to classify galaxy morphologies using CNNs.
- We designed and compared multiple CNN architectures.
- We used the Galaxy Zoo 2 dataset, with crowd-sourced probabilistic labels, to train the CNNs.
- We predicted the 37 probability values by regression, then mapped them to 17 simplified morphological classes.
- We achieved an overall high accuracy, proving the effectiveness of CNNs as scalable tools for automated morphological classification in astronomy.

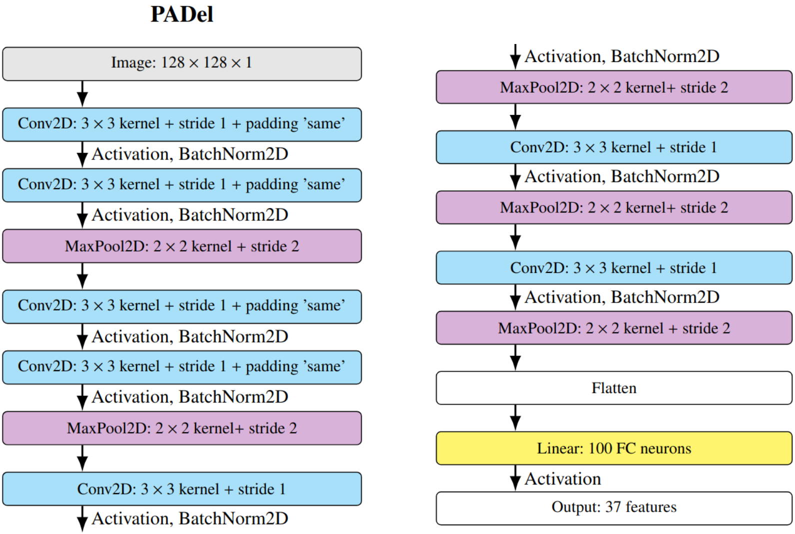
Table of Contents

Architecture diagrams: SkyNet

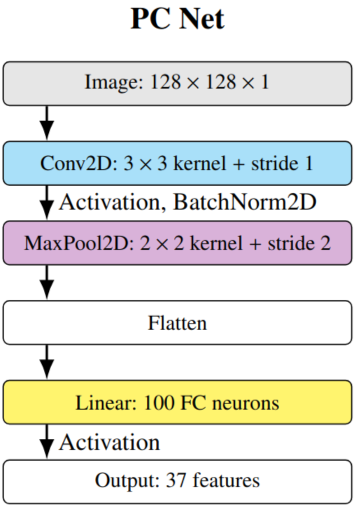


Architecture diagrams: PADel

UNIVERSITÀ DEGLI STUDI DI PADOVA



Architecture diagrams: PC



Architecture diagrams: JAGZoo

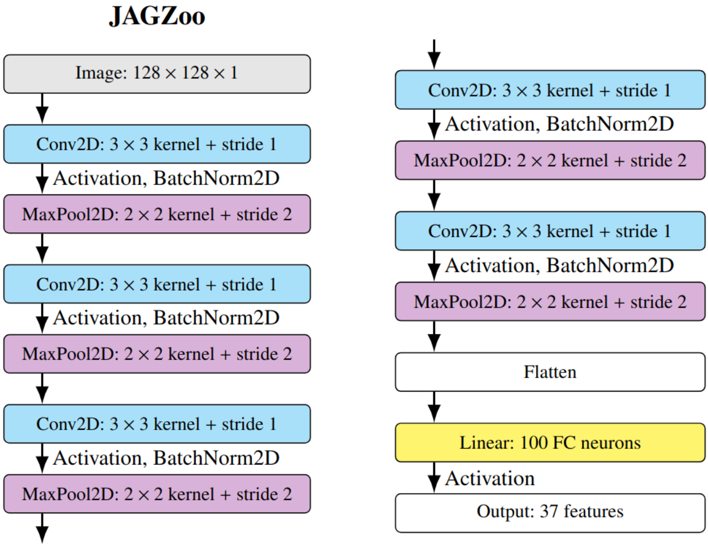
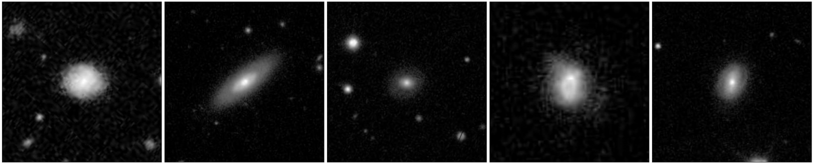


Table of Contents

Uncertain Classes



(a) S0A galaxies.

(b) SAb galaxies.

Figure: PADel classification of unlabeled galaxies.

Table of Contents

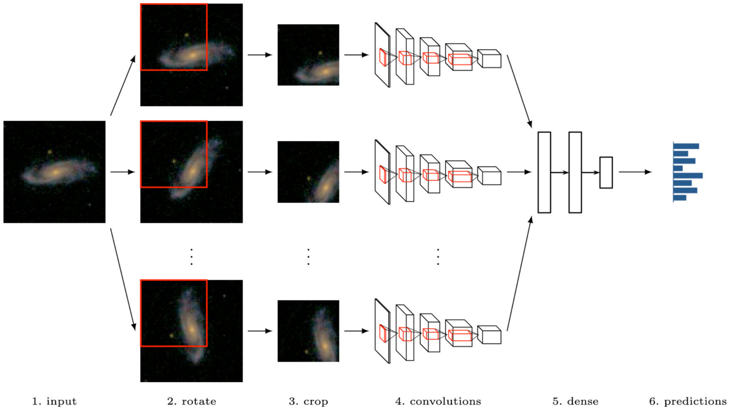
Results

UNIVERSITÀ DEGLI STUDI DI PADOVA

True Label	Samples	JAGZoo		JAGZoo-m		PADel-m		PADel		VGG		SkyNet		PC	
		Misclassified	Score	Misclassified	Score	Misclassified	Score	Misclassified	Score	Misclassified	Score	Misclassified	Score	Misclassified	Score
E0	1064	36	0.97	52	0.95	59	0.94	48	0.95	43	0.96	55	0.95	40	0.96
E3	1182	151	0.87	118	0.90	138	0.88	138	0.88	130	0.89	142	0.88	474	0.6
E6	170	48	0.72	51	0.70	44	0.74	48	0.72	35	0.79	49	0.71	155	0.09
S0A	698	410	0.41	390	0.44	353	0.49	385	0.45	350	0.50	366	0.48	491	0.3
S0B	119	60	0.50	48	0.60	42	0.65	54	0.55	38	0.68	45	0.62	94	0.21
SAa	1	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0
SAb	142	86	0.39	73	0.49	71	0.5	53	0.63	61	0.57	63	0.56	131	0.08
SAc	337	109	0.68	80	0.76	76	0.77	59	0.82	80	0.76	67	0.80	166	0.51
SAd	5	5	0.0	4	0.20	5	0.0	5	0.0	5	0.0	4	0.20	5	0.0
SBa	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SBb	65	30	0.54	30	0.54	30	0.54	30	0.54	30	0.54	27	0.58	46	0.29
SBc	66	38	0.42	24	0.64	25	0.62	27	0.59	24	0.64	19	0.71	58	0.12
SBd	1	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0
S0/a e/o	292	78	0.73	49	0.83	47	0.84	52	0.82	64	0.78	50	0.83	69	0.76
Sc/d e/o	111	27	0.76	29	0.74	28	0.75	32	0.71	33	0.70	27	0.76	33	0.7
SB0/a e/o	2	2	0.0	2	0.0	2	0.0	2	0.0	2	0.0	2	0.0	2	0.0
A	6	5	0.17	0	1.0	1	0.83	0	1.0	1	0.83	1	0.83	4	0.33

Why CNNs?

Università degli Studi di Padova



Dieleman et al 2015