



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES



sid.inpe.br/mtc-m21b/2016/08.05.13.14-TDI

MACHINE AND DEEP LEARNING APPLIED TO GALAXY MORPHOLOGY

Paulo Henrique Barchi

Doctorate Thesis of the Graduate
Course in Applied Computing,
guided by Drs. Reinaldo Roberto
Rosa, and Reinaldo Ramos de
Carvalho, approved in March 09,
2020.

URL of the original document:

<<http://urlib.net/8JMKD3MGP3W34P/3M7RCUE>>

INPE
São José dos Campos
2020

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GBDIR)
Serviço de Informação e Documentação (SESID)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

**BOARD OF PUBLISHING AND PRESERVATION OF INPE
INTELLECTUAL PRODUCTION - CEPPII (PORTARIA Nº
176/2018/SEI-INPE):****Chairperson:**

Dra. Marley Cavalcante de Lima Moscati - Centro de Previsão de Tempo e Estudos
Climáticos (CGCPT)

Members:

Dra. Carina Barros Mello - Coordenação de Laboratórios Associados (COCTE)
Dr. Alisson Dal Lago - Coordenação-Geral de Ciências Espaciais e Atmosféricas
(CGCEA)
Dr. Evandro Albiach Branco - Centro de Ciência do Sistema Terrestre (COCST)
Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia e Tecnologia
Espacial (CGETE)
Dr. Hermann Johann Heinrich Kux - Coordenação-Geral de Observação da Terra
(CGOBT)
Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação - (CPG)
Sílvia Castro Marcelino - Serviço de Informação e Documentação (SESID)

DIGITAL LIBRARY:

Dr. Gerald Jean Francis Banon
Clayton Martins Pereira - Serviço de Informação e Documentação (SESID)

DOCUMENT REVIEW:

Simone Angélica Del Ducca Barbedo - Serviço de Informação e Documentação
(SESID)
André Luis Dias Fernandes - Serviço de Informação e Documentação (SESID)

ELECTRONIC EDITING:

Ivone Martins - Serviço de Informação e Documentação (SESID)
Cauê Silva Fróes - Serviço de Informação e Documentação (SESID)



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES



sid.inpe.br/mtc-m21b/2016/08.05.13.14-TDI

MACHINE AND DEEP LEARNING APPLIED TO GALAXY MORPHOLOGY

Paulo Henrique Barchi

Doctorate Thesis of the Graduate
Course in Applied Computing,
guided by Drs. Reinaldo Roberto
Rosa, and Reinaldo Ramos de
Carvalho, approved in March 09,
2020.

URL of the original document:

<<http://urlib.net/8JMKD3MGP3W34P/3M7RCUE>>

INPE
São José dos Campos
2020

Cataloging in Publication Data

Barchi, Paulo Henrique.

B235m Machine and deep learning applied to galaxy morphology / Paulo Henrique Barchi. – São José dos Campos : INPE, 2020.

xviii + 85 p. ; (sid.inpe.br/mtc-m21b/2016/08.05.13.14-TDI)

Thesis (Doctorate in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2020.

Guiding : Drs. Reinaldo Roberto Rosa, and Reinaldo Ramos de Carvalho.

1. Computational astrophysics. 2. Galaxy morphology.
3. Machine learning. 4. Deep learning. I.Title.

CDU 004.032.2:52



Esta obra foi licenciada sob uma Licença [Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada](https://creativecommons.org/licenses/by-nc/3.0/).

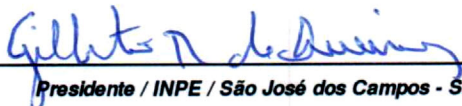
This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](https://creativecommons.org/licenses/by-nc/3.0/).

Aluno (a): **Paulo Henrique Barchi**

Título: "MACHINE AND DEEP LEARNING APPLIED TO GALAXY MORPHOLOGY"

Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de **Doutor(a)** em
Computação Aplicada

Dr. Gilberto Ribeiro de Queiroz

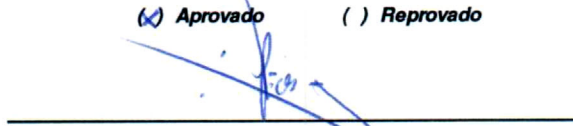


Presidente / INPE / São José dos Campos - SP

Participação por Vídeo - Conferência

Aprovado Reprovado

Dr. Reinaldo Roberto Rosa

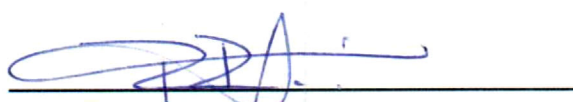


Orientador(a) / INPE / SJC Campos - SP

Participação por Vídeo - Conferência

Aprovado Reprovado

Dr. Reinaldo Ramos de Carvalho

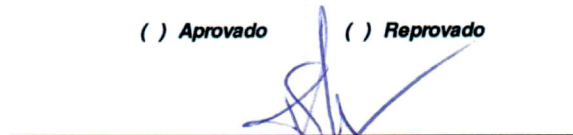


Orientador(a) / UNICID / São Paulo - SP

Participação por Vídeo - Conferência

Aprovado Reprovado

Dr. Thales Sehn Körting

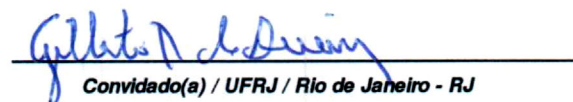


Membro da Banca / INPE / São José dos Campos - SP

Participação por Vídeo - Conferência

Aprovado Reprovado

Dra. Karín Menéndez-Delmestre



Convidado(a) / UFRJ / Rio de Janeiro - RJ

Participação por Vídeo - Conferência

Aprovado Reprovado

Este trabalho foi aprovado por:

maioria simples

unanimidade

São José dos Campos, 09 de março de 2020

Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de *Doutor(a)* em
Computação Aplicada

Dr. Irapuan Rodrigues de Oliveira Filho



Convidado(a) / UNIVAP / São José dos Campos - SP

() Participação por Video - Conferência

Aprovado

() Reprovado

Este trabalho foi aprovado por:

() maioria simples

unanimidade

São José dos Campos, 09 de março de 2020

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my parents, Antonio Claudio Barchi and Maria Lucia Madeira, and my siblings, Claudia Maria Barchi and André Paulo Barchi (extending to their families), for constant support and motivation in my life.

I would like to thank Dr. Reinaldo Roberto Rosa and Dr. Reinaldo Ramos de Carvalho for accepting work with me and proposing this very interesting research theme. I would like to thank Dr. Marcelle Soares-Santos as well, who accepted, welcomed and guided me through my research experience abroad. Their advising, guidance and criticism have been continuously shaping this research project and myself as a researcher.

Special thanks goes to Gerson de Oliveira Barbosa, Carlos Alexandre Romani, Leonardo de Souza Vieira, Asiel Bomfin Junior, Maria Elidaiana da Silva Pereira, Jui-Jen (Ryan) Wang, Johnny Hebert Esteves de Queiroz for all the support, motivation and friendship.

I thank Rubens A. Sautter, Dr. Diego Stalder, Dr. Tatiana C. Moura, Dr. Bjorn Penning, Alyssa Garcia, Luke Korley, Dr. Helena Domínguez Sánchez, and Dr. Sandro B. Rembold for productive discussions and thoughtful comments on several topics related to the present work.

I am grateful to the graduate program professors and students at INPE for all the classes, support and inspiring conversations. I have been delighted with how welcoming and supportive the people from the Physics Department at Brandeis University had been in my PhD sandwich timespan.

This work had been financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

I am also grateful to all people who directly or indirectly helped materialize this work.

ABSTRACT

Morphological classification is a key piece of information to define samples of galaxies aiming to study the large-scale structure of the universe. In essence, the challenge is to build up a robust methodology to perform a reliable morphological estimate from galaxy images. Here, I investigate how to substantially improve the galaxy classification within large datasets by mimicking human classification. I combine accurate visual classifications from the Galaxy Zoo project with machine and deep learning methodologies. I propose two distinct approaches for galaxy morphology: one based on non-parametric morphology and traditional machine learning algorithms; and another based on deep learning. To measure the input features for the traditional machine learning methodology, I and my collaborators have developed a system called CyMorph, with a novel non-parametric approach to study galaxy morphology. The main datasets employed comes from the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7). I also discuss the class imbalance problem considering three classes. Performance of each model is mainly measured by overall accuracy (OA). A spectroscopic validation with astrophysical parameters is also provided for Decision Tree models to assess the quality of our morphological classification. In all of our samples, both Deep and Traditional Machine Learning approaches have over 94.5% OA to classify galaxies in two classes (elliptical and spiral). I compare our classification with state-of-the-art morphological classification from literature. Considering only two classes separation, I achieve 99% OA in average when using our deep learning models, and 82% when using three classes. I provide a catalog with 670,560 galaxies containing our best results, including morphological metrics and classification.

Keywords: Computational Astrophysics. Galaxy Morphology. Machine Learning. Deep Learning.

APRENDIZADO DE MÁQUINA TRADICIONAL E PROFUNDO APLICADO A MORFOLOGIA DE GALÁXIAS

RESUMO

Classificação morfológica é peça chave de informação para definir amostras de galáxias com objetivo de estudar a estrutura do Universo em larga-escala. Em essência, o desafio é construir uma metodologia robusta para produzir uma estimativa morfológica confiável a partir de imagens de galáxias. Aqui, investigo como melhorar substancialmente a classificação automática de galáxias em grandes conjuntos de dados ao imitar a classificação fornecida por humanos. Combino classificações visuais do projeto Galaxy Zoo com metodologias de aprendizado de máquina tradicional e profundo. Proponho duas abordagens distintas para morfologias de galáxias: uma baseada em morfologia não-paramétrica e algoritmos de aprendizado de máquina tradicional; e outra baseada em aprendizado profundo. Para medir as características morfológicas de entrada para algoritmos de aprendizado de máquina tradicional, desenvolvi com meus colaboradores um sistema chamado CyMorph, com uma nova abordagem não-paramétrica para estudar morfologia de galáxias. O principal conjunto de dados explorado provém do Sloan Digital Sky Survey Data Release 7 (SDSS-DR7). Também discuto o problema de desbalanceamento de classes considerando o problema com três classes. A performance de cada modelo é medida principalmente por acurácia global. A validação espectroscópica com parâmetros astrofísicos também é fornecida para os modelos de Árvore de Decisão para avaliar a qualidade de nossa classificação morfológica. Em todas as nossas amostras, tanto com aprendizado de máquina profundo como tradicional, obtenho mais de 94.5% de acurácia global para classificar galáxias em duas classes (elíptica e espiral). Comparo minha classificação com classificações morfológicas do estado-da-arte da literatura. Considerando apenas duas classes, atingi 99% de acurácia global e média usando modelos de aprendizado profundo, e 82% usando três classes. Forneço uma catálogo com 670.560 galáxias contendo nossos melhores resultados, incluindo métricas morfológicas e classificações.

Palavras-chave: Astrofísica Computacional. Morfologia de Galáxias. Aprendizado de Máquina. Aprendizado de Máquina Profundo.

LIST OF FIGURES

	<u>Page</u>
2.1 Hubble sequence of galaxy morphologies.	5
2.2 Galaxy Zoo 1.	7
2.3 Galaxy Zoo 2.	8
2.4 Schematic representation of a neuron.	11
2.5 Illustration within a 2-D environment of a simple straightforward convo- lution operation example.	13
2.6 Representation of typical (complex) Convolutional Layer.	13
2.7 Inception Module.	14
2.8 GoogleNet architecture.	15
3.1 Examples of ETG (left) and LTG in black and white from SDSS-DR7. . .	17
3.2 Preprocessing example.	21
3.3 Traditional Machine Learning schema.	28
3.4 Example of convolutions applied to a galaxy for illustration purposes. . .	31
4.1 Illustrative sketch of traditional machine learning and deep learning flows.	35
4.2 Optimization process for morphological metrics configuration.	40
4.3 Results on galaxy morphology using Classic CAS (CONSELICE, 2003; LOTZ et al., 2004).	42
4.4 Results on galaxy morphology using CyMorph.	43
4.5 Panel about class imbalance in galaxy morphology.	48
4.6 Performance metrics plots.	50
4.7 Sample of misclassified galaxies among traditional machine learning ap- proach and GZ1.	53
4.8 Sample of misclassified galaxies among our deep learning approach and GZ2.	53
4.9 Spectroscopic validation.	55
4.10 Star formation acceleration (SFA) as a function of NUV-r colours.	57
4.11 Histograms presenting classifications for Nair and Abraham (2010)'s sam- ple.	59
4.12 Histograms presenting classifications for Sánchez et al. (2018) sample. . .	60
4.13 Sample classified as spiral galaxies by our classifier with $-2.25 \leq$ T-Type ≤ -2 by Sánchez et al. (2018).	61
4.14 Sample of galaxies classified by our deep learning approach.	64

LIST OF TABLES

	<u>Page</u>
3.1 Number of galaxies for the main samples in this work from each database (SDSS, GZ1 and GZ2).	19
3.2 Experiments with different configurations for CNN architectures.	30
4.1 Parameter ranges explored in the optimization process. Asymmetry is omitted since it depends only on d_σ . Concentration(*) does not depend on d_σ	39
4.2 Overall Accuracy (OA in percentage) for all approaches considering GZ1 classification (elliptical and spiral galaxies separation).	46
4.3 Overall Accuracy (OA in percentage) for all approaches considering GZ2 classification. The darker the green colour of a cell, the better OA obtained.	46

LIST OF ABBREVIATIONS

AUC	–	Area Under the Curve
ANN	–	Artificial Neural Network
A	–	Asymmetry (non-parametric galaxy morphology metric)
BBH	–	Binary Black Holes
BNS	–	Binary Neutron Stars
C	–	Concentration (non-parametric galaxy morphology metric)
CAS system	–	Concentration, Asymmetry, Smoothness system
CNN	–	Convolutional Neural Network
CV	–	Cross-Validation
CyMorph	–	Non-parametric galaxy morphology system
DR7	–	(SDSS) Data Release 7
DT	–	Decision Tree
DL	–	Deep Learning
ETGs	–	Early-Type Galaxies (ellipticals)
FN	–	False Negative
FP	–	False Positive
FWHM	–	Full Width at Half Maximum
GHS	–	Geometric Histogram Separation
GPA	–	Gradient Pattern Analysis
GS	–	Grid Search
GZ1	–	Galaxy Zoo 1
GZ2	–	Galaxy Zoo 2
GW	–	Gravitational Waves
G_2	–	GPA's second moment
H	–	Entropy (non-parametric galaxy morphology metric)
K (parameter)	–	The area of the galaxy's Petrosian ellipse divided by the – area of the Full Width at Half Maximum
LTGs	–	Late-Type Galaxies (spirals)
MLP	–	Multilayer Perceptron
NUV	–	Near-ultraviolet
OA	–	Overall Accuracy
P	–	Precision
R	–	Recall
ROC curve	–	Receiver Operating Characteristic curve
SDSS	–	Sloan Digital Sky Survey
S	–	Smoothness (non-parametric galaxy morphology metric)
SExtractor	–	Source Extractor
SFA	–	Star Formation Acceleration
SVM	–	Support Vector Machine
(T)ML	–	(Traditional) Machine Learning

TN – True Negative
TP – True Positive

CONTENTS

	<u>Page</u>
1 INTRODUCTION	1
2 THEORETICAL FOUNDATIONS	5
2.1 Galaxy morphology	5
2.1.1 Galaxy zoo	5
2.2 Machine learning	8
2.2.1 Performance metrics — overall accuracy	9
2.2.2 Decision Tree	9
2.2.3 Support Vector Machine	10
2.2.4 Artificial neural networks	10
2.3 Deep learning	11
2.3.1 Deep convolutional neural network	12
2.3.2 GoogleNet Inception	14
2.4 Machine learning applied to astrophysics	14
3 MATERIAL AND METHODOLOGY	17
3.1 Sloan Digital Sky Survey	17
3.2 Sample and data	18
3.3 CyMorph - non-parametric galaxy morphological system	19
3.3.1 Preprocessing	20
3.3.2 Error detection	21
3.3.3 Concentration	22
3.3.4 Asymmetry	24
3.3.5 Smoothness	25
3.3.6 Gradient pattern analysis applied to galaxy morphology	26
3.3.7 Entropy	27
3.4 Machine learning applied to galaxy morphology	27
3.5 Deep learning applied to galaxy morphology	29
4 MACHINE AND DEEP LEARNING APPLIED TO GALAXY MORPHOLOGY - A COMPARATIVE STUDY	33
4.1 Introduction	33
4.2 Advances in non-parametric galaxy morphology	36

4.2.1	Geometric histogram separation (δ_{GHS})	38
4.2.2	Optimizing morphological metrics configuration	38
4.2.3	Results on morphology	41
4.3	Machine learning applied to galaxy morphology	44
4.4	Results on classification and discussion	45
4.4.1	Classifier’s performance by overall accuracy (OA)	45
4.4.2	Class imbalance in galaxy morphology	47
4.4.3	Classifier’s performance by ROC curve and AUC	49
4.4.4	Learning about differences between TML and DL from misclassifications	51
4.4.5	Validating classification with spectroscopic data	54
4.4.6	Case study: star formation acceleration and morphologies	56
4.5	Comparison to other available catalogs	58
4.6	Final catalog (paper appendix)	61
5	GENERAL DISCUSSION	65
5.1	Computational aspects	65
6	CONCLUSION	69
6.1	Summary	69
6.2	Concluding remarks	70
	REFERENCES	73
	ANNEX A - PUBLICATIONS.	81
A.1	Publication 1	81
A.2	Publication 2	81
A.3	Publication 3	81
A.4	Publication 4	82
A.5	Publication 5	82
	ANNEX B - EXTRACTING THE SLIME MOLD GRAPH FROM THE COSMIC WEB	85

1 INTRODUCTION

¹Astronomy has become extremely data-rich with the advancement of new technologies in recent decades. New observation instruments such as satellites and telescopes provide massive datasets. Such datasets are not only voluminous but also complex, since these data often have spatial and temporal components, and are collected at different frequencies and resolutions (WAY et al., 2012; IVEZIĆ et al., 2014; FEIGELSON; BABU, 2006; TAN et al., 2005). In observational astronomy, one of the main data resources is photometric: each image corresponds to a field of view of specific area in the sky, in a determined frequency-band (or multi-band), which admits noise and can contain multiple (tens, hundreds or millions of) objects. Concurrently, hardware and software technologies related to Machine and Deep Learning have been constantly improving for handling and generating more value from such huge datasets in different research and industry contexts (AL-JARRAH et al., 2015).

One of the key aspects of any extragalactic investigation is the definition of an unbiased sample that includes reliable morphological types. Galaxy morphological properties result from not only the internal formation and evolution processes but also from interaction with the environment. Galaxies in groups or clusters may have diverse evolutionary paths compared to isolated ones, which is clearly reflected in their morphology. Therefore, classification of galaxies into a meaningful taxonomy system is of paramount importance for galaxy formation and evolution studies. The main differentiation of galaxies is between Early-Type Galaxies (ETGs), which are elliptical galaxies with basically one single structural feature, and Late-Type Galaxies (LTGs) which have a prominent disk (HUBBLE, 1926; HUBBLE, 1936; MO et al., 2010) — see Chapter 2 and 4 for more details on this subject.

Computational Astrophysics and applied Machine Learning (ML) provide two of the most used approaches for automatically classifying galaxies by their morphology:

- a) To extract non-parametric morphology features with a system especially developed for this task, and, use such features as input for a Traditional Machine Learning (TML) algorithm, which generates the classification output — a few examples with regard to this approach: Rosa et al. (2018), Barchi et al. (2016), Ferrari et al. (2015), Conselice (2003), Lotz et al. (2004), Abraham et al. (1996).

¹Since this thesis is in the alternative format, with separated chapters for Literature Review, Material and Methodology, and Chapter 4 has an adapted version of the main publication of this research, this introduction chapter is brief to avoid being repetitive.

- b) To apply Deep Learning (DL) algorithms directly upon the images. Here, thousands of features are internally extracted by the network, and it produces the classification output as its last processing step — examples: Khalifa et al. (2017), Sánchez et al. (2018).

This PhD project is part of the FAPESP thematic project entitled “*What Drives the Stellar Mass Growth of Early-Type Galaxies? Born or made: the saga continues...*”. One of the main goals here is to develop a highly accurate and robust methodology for automatically identifying Early-Type Galaxies, so that our research group can perform the main study around the mass growth of such galaxies. This thesis is written around the main publication by its author (BARCHI et al., 2020) — see Chapter 4 for an adapted version — which compares the application of the two methodologies enumerated above. The main hypotheses explored here are:

- a) **Hypothesis 1:** *It is hypothesized that TML and DL approaches can reliably perform galaxy morphology classification, considering the separation between ETGs and LTGs.* Although one can say this is a safe statement, it had not being shown previously in literature the application and comparison of both methodologies. We develop a specialized system to extract non-parametric morphology features, apply different TML methods and test different configurations for DL experiments. We use labeling provided by humans as supervision to all learning processes (more details in Subsection 2.1.1). So, by reliably, we mean a high degree of agreement between the machines and their supervision. We explain performance metrics in Subsections 2.2.1 and 4.4.3 and show results in Section 4.4.
- b) **Hypothesis 2:** *DL achieves higher standards of performance than TML for visual classification, however, TML has a similar performance (considering two classes) while preserving meaningful features.* By hand-engineering a system to extract features in the TML flow, it is much more understandable how galaxies are classified in this approach than using DL. This hypothesis states that although DL certainly obtain higher performance than TML, TML still has its value for further classification analysis.

Both hypotheses (and derivatives) are explored further throughout this document. Chapters 2 and 3 complement the theoretical foundations and the methodology presented in the main publication of this research, respectively. Chapter 4 presents an adapted version of the main publication from this research. Chapter 5 provides

a further discussion of Section 4.4 and presents a computational analysis of the developed systems. Chapter 6 concludes this document. Annex chapters present a parallel project and publications in the course of this PhD.

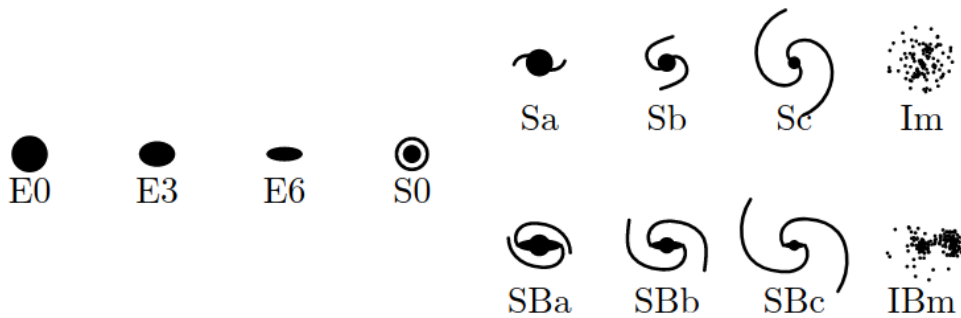
2 THEORETICAL FOUNDATIONS

This chapter complements the theoretical foundations for the paper presented in Chapter 4. Here, we cover topics of Galaxy Morphology and Machine Learning which are not covered in details in the full paper, and related works as well.

2.1 Galaxy morphology

In observational research, the most basic process is the classification of objects into a taxonomy system. The challenge is to build a robust methodology to perform a reliable classification. The first classification of galaxies by [Hubble \(1926\)](#), [Hubble \(1936\)](#) — Figure 2.1 — distinguishes galaxies with dominant bulge component — also known as Early-Type Galaxies (ETGs) — from galaxies with a prominent disk — called Late-Type Galaxies (LTGs). LTGs are commonly referred to as spiral galaxies because of their prominent spiral arms, while ETGs are commonly referred to as elliptical galaxies as they have a more simple elliptical structure, with less structural differentiations. A more refined classification divides ETGs by their ellipticities (with prefix E), while LTGs fork into two groups: barred (with a barred shaped central structure — SB) and non-barred (S)galaxies. These two groups can be also more refined by their spiral arms strength.

Figure 2.1 - Hubble sequence of galaxy morphologies.



Source: [Mo et al. \(2010\)](#).

2.1.1 Galaxy zoo

Even with a fully automated process, in a given moment we would need some kind of human supervision to guide the machine and give proper names to the patterns

it has been finding. Galaxy Zoo is a citizen science project for classifying galaxies by their morphologies. In its first phase — Galaxy Zoo 1 (GZ1) — the user had six options in three main categories (see Figure 2.2):

a) Spiral galaxy (LTG):

- clockwise arms;
- counterclockwise arms;
- edge-on / unclear.

b) Elliptical galaxy (ETG).

c) Undefined:

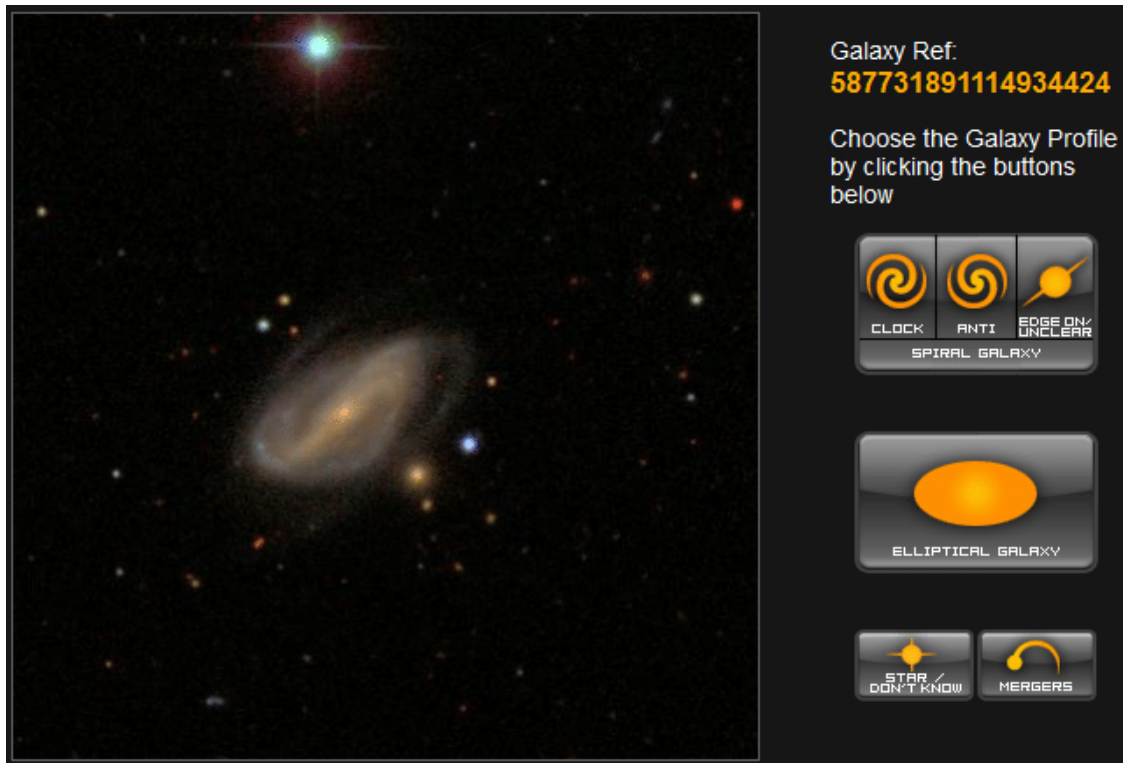
- Star / unrecognizable object;
- Mergers.

In its second phase, Galaxy Zoo introduced more options and classification stages (see Figure 2.3), which provided labeled data for galaxy morphology with an unprecedented level of detail.

I use GZ1 as supervision for distinguishing ETGs from LTGs, and, GZ2 when attacking problems with more classes. We report experiments with 11 classes (Ei, Ec, Er, Sa, Sb, Sc, Sd, SBa, SBb, SBc, SBd — which roughly corresponds to the scheme in Figure 2.1), 9 classes (same as previous but one class for all early-type united) and 7 classes (same as previous but discarding the faintest galaxy types: Sd and SBd). For clarity, I itemize each of the problems with different number of classes below:

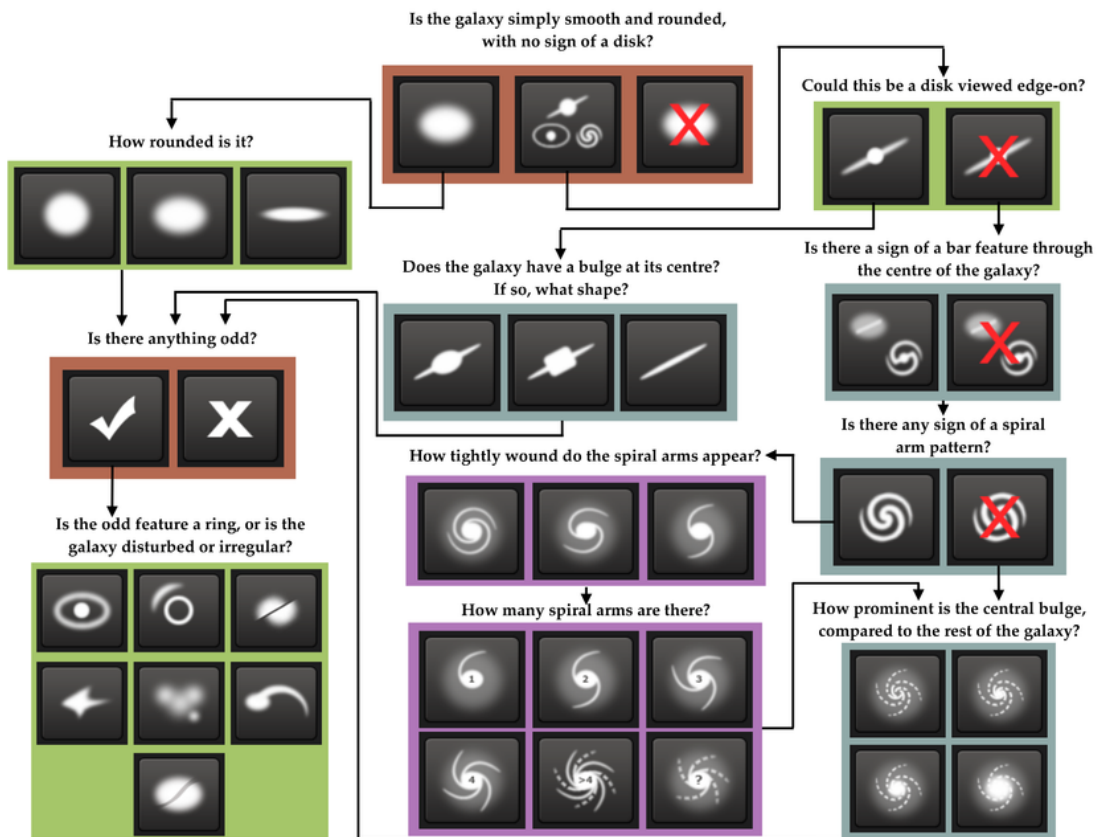
- 2 classes: ETGs, LTGs.
- 3 classes: ETGs, barred spirals and non-barred spirals.
- 7 classes: ETGs (grouped together), Sa, Sb, Sc, SBa, SBb, SBc.
- 9 classes: Ei, Ec, Er, Sa, Sb, Sc, SBa, SBb, SBc.
- 11 classes: Ei, Ec, Er, Sa, Sb, Sc, Sd, SBa, SBb, SBc, SBd.

Figure 2.2 - Galaxy Zoo 1.



Source: Lintott et al. (2008), LINTOTT et al. (2011).

Figure 2.3 - Galaxy Zoo 2.



Source: Willett et al. (2013).

2.2 Machine learning

The ML field is governed by the central question: “How can we build computers which can automatically improve their performances with experience? Furthermore, which are the fundamental laws which drive the learning process?” (MITCHELL, 1997). A machine learns with respect to a specific task, a performance metric and an experience. The evolution of this learning process is noted if the system reliably improves its performance with an specific task, following the experience. We specify ML in this work as follows:

- a) Task: morphologically classify galaxies;
- b) Performance measure: overall accuracy, precision, recall, Receiver Operating Characteristic (ROC) curve and Area Under the ROC curve — AUC (BRADLEY, 1997). We will cover these subjects in detail in Subsections

2.2.1 and 4.4.3, respectively;

- c) Training experience: galaxies which have both, Galaxy Zoo classification and non-parametric morphological features.

2.2.1 Performance metrics — overall accuracy

To empirically guide the experiments and the learning processes themselves, also to analyze if determined approach has reached its goal, it is necessary to define and employ consistent validation metrics, also called performance metrics. In this subsection, we describe the main performance metric used in this work: Overall Accuracy. We calculate overall accuracy (OA) in terms of true positives (TP: correctly classified galaxies), false positives (FP: misclassification; objects which are not from this class and were classified as such), true negatives (TN: objects correctly not classified for such class), and false negatives (FN: galaxies that should be labeled for such class, but were not). It indicates the probability that a individual galaxy will be correctly classified — see Equation 2.1.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

However, OA alone is not enough to characterize the performance. A model can have high OA but poorly perform for one or more specific classes. For the most interesting cases (experiments applying DL for the problem with three classes, for example) we cover more performance metrics: precision (P) and recall (R). In astronomy, P and R are known as purity and completeness. P and R are also calculated by means of TP, FP, TN and FN:

$$P = \frac{TP}{TP + FP} \quad (2.2)$$

$$R = \frac{TP}{TP + FN} \quad (2.3)$$

2.2.2 Decision Tree

One of the most used methods for classification and regression is the Decision Tree (DT). The model is adjusted through the learning process to predict the classification by simple decision rules inferred from the dataset. Among the different versions and variations of DTs, we use the optimized version of Classification and Regression Tree

(CART) algorithm. CART builds up binary trees using feature and threshold that yield the largest information gain at each node (QUINLAN, 1986; GÉRON, 2019).

2.2.3 Support Vector Machine

Another influential method for supervised classification is the Support Vector Machine (SVM) which finds the optimal hyperplane that divides the target classes. SVM performs this task by drawing infinite different hyperplanes for separating target classes aiming to get the minimum error rate (HEARST et al., 1998; CORTES; VAPNIK, 1995). The hyperplane which maximizes the separation margins among the classes is the optimal hyperplane, i.e., the hyperplane provides a unique optimal solution for the problem (HEARST et al., 1998; CORTES; VAPNIK, 1995).

Generally, the input data is not linearly separable. SVM performs the kernel-trick in order to find the optimal hyperplane. The kernel-trick maps the original input space into some high dimensional space through a dot-product in the feature space by a N -dimensional vector function – polynomial function or radial basis function, for example (CORTES; VAPNIK, 1995).

2.2.4 Artificial neural networks

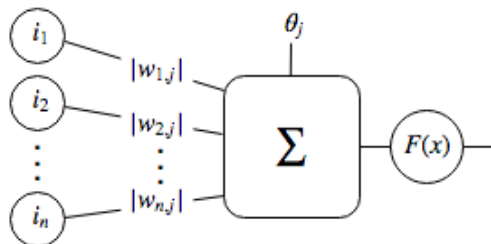
A standard Neural Network (NN) consists of many simple, connected neurons, each one being a computing unit which outputs a sequence of real-valued activations. A artificial neural network architecture is organized in layers: input, hidden (which may be one or many) and output. A Multilayer Perceptron (MLP) has at least three layers (one input, one hidden and one output layer). With information from previous layers, weighted connections activates neurons of the next layer. Each neuron has n inputs i , weights (w), bias (b), an activation function ($F(x)$) and output (y) Formally, as represented in Figure 2.4, the output y of the j -th neuron is given by the Equation 2.4:

$$y_j = F \left(b + \sum_{k=1}^n w_{k,j} i_k \right). \quad (2.4)$$

Weighted inputs and bias are adjustable parameters which makes the neural network a parameterized system. Among the nonlinear activation functions, logistic function and Rectified Linear Unit (ReLU) are two of the most widely used. The logistic activation function is heavily applied to predict probabilities since it ranges from 0 to 1. ReLU is present in most of the convolutional neural networks and it is given by the following equation: $R(x) = \max(0, x)$, i.e., $R(x)$ is zero for $x < 0$ and x when $x > 0$. The training process optimizes the weights for each neuron by minimizing the

error of predictions and reaching a specified level of accuracy. One epoch of training is one forward pass and one backward pass in all training examples through the whole network. The batch size is the number of training examples in one forward/backward pass (MITCHELL, 1997; GÉRON, 2019; GOODFELLOW et al., 2016).

Figure 2.4 - Schematic representation of a neuron.



Source: Author's production.

2.3 Deep learning

A Neural Network (NN) can be extremely complex when using an architecture with many layers. Deep Learning (DL) methods are built through a deep neural network with multiple layers of non-linear transformations. A detailed explanation of deep learning is out of the scope of this thesis, but we provide here a brief introduction. Multiple layers ensure the deep characteristic while multiple neurons represent its width. Its quintessence is many parametric functions composed of many other parametric functions. Each of these parametric functions has multiple inputs and, possibly, multiple outputs (GOODFELLOW et al., 2016; GÉRON, 2019). DL models typically benefit from large amounts of training data. The training process can be accelerated dramatically using parallel processors such as Graphics Processing Unit (GPU). This approach is impracticable if a huge amount of data is not available to be extracted for optimizing the features and weights in each layer/neuron. To have such amount of data processed in a reasonable time, we need specifically dedicated hardware. With the advance of GPU and hardware in general, it is well-established that DL is nowadays the state of the art approach for classification tasks (GOODFELLOW et al., 2016; GÉRON, 2019).

2.3.1 Deep convolutional neural network

Deep Convolutional Neural Network (CNN) or simply Convolutional Networks (Lecun, 1989) are a special kind of neural network for processing data with grid-like topology. CNN uses a hierarchy of layers to recognize the desired patterns from the input data. A CNN necessarily has a convolutional layer and uses a variation of the MLP (Subsection 2.2.4). Convolution preserves the spatial relationship between pixels by using submatrices from the input data for learning features. The convolution operation has two arguments: one often referred as input (simply the input data or the output from a previous layer) and the kernel, filter, or feature detector. There are convolutional models using 1D, 3D and even higher-dimensional data, however, generally, the input data has two-dimensions (images). The convolution operation allows CNN to be deeper with much fewer parameters than a non-convolutional network. Considering the input as a two-dimensional matrix, the kernel generally is a submatrix. The output from the convolution operation can be named the feature map. Considering a two-dimensional image χ as input, a suitable method for performing the convolution operation consists in using a two-dimensional kernel φ . Equation 2.5 is one possible formulation of this operation and Figure 2.5 presents a simple theoretical example to clarify the operation — both, the equation and figure, inspired by Goodfellow et al. (2016).

$$S(i, j) = \sum_m \sum_n \chi(m, n) \varphi(i - m, j - n). \quad (2.5)$$

Figure 2.5 - Illustration within a 2-D environment of a simple straightforward convolution operation example.

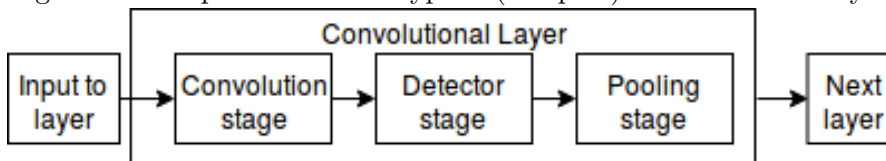
$$\begin{array}{c}
 \text{Input} \\
 \left[\begin{array}{cccc}
 i_1 & i_2 & i_3 & i_4 \\
 i_5 & i_6 & i_7 & i_8 \\
 i_9 & i_{10} & i_{11} & i_{12} \\
 i_{13} & i_{14} & i_{15} & i_{16}
 \end{array} \right] \\
 \underbrace{\hspace{10em}}_{4 \times 4}
 \end{array}
 \quad
 \begin{array}{c}
 \text{Kernel} \\
 \left[\begin{array}{cc}
 k_1 & k_2 \\
 k_3 & k_4
 \end{array} \right] \\
 \underbrace{\hspace{2em}}_{2 \times 2}
 \end{array}
 =
 \begin{array}{c}
 \text{Output} \\
 \left[\begin{array}{ccc}
 \boxed{i_1k_1 + i_2k_2} & i_2k_1 + i_3k_2 & i_3k_1 + i_4k_2 \\
 \boxed{+ i_5k_3 + i_6k_4} & + i_6k_3 + i_7k_4 & + i_7k_3 + i_8k_4 \\
 i_5k_1 + i_6k_2 & i_6k_1 + i_7k_2 & i_7k_1 + i_8k_2 \\
 + i_9k_3 + i_{10}k_4 & + i_{10}k_3 + i_{11}k_4 & + i_{11}k_3 + i_{12}k_4 \\
 i_9k_1 + i_{10}k_2 & i_{10}k_1 + i_{11}k_2 & i_{11}k_1 + i_{12}k_2 \\
 + i_{13}k_3 + i_{14}k_4 & + i_{14}k_3 + i_{15}k_4 & + i_{15}k_3 + i_{16}k_4
 \end{array} \right] \\
 \underbrace{\hspace{10em}}_{3 \times 3}
 \end{array}$$

The output is restricted to positions where the kernel lies entirely within the input matrix. Inside the dashed rectangles, we have the input (left) and kernel (middle) operands responsible to produce the first element of the output (right). Below each matrix, we present its dimension. We can imagine the dashed window (kernel) sliding through the input matrix to produce the output.

Source: Author's production.

A typical convolutional layer of a CNN has three stages, as represented in Figure 2.6. In the first stage, several convolutions are performed in parallel to produce a set of linear activations. The second is the detector stage: a nonlinear activation function process each of these linear activations. In the third stage, the pooling layer reduces the spatial dimension which lower the number of parameters. Furthermore, it summarizes the statistic of the neighboring outputs and reduces overfitting occurrences (GOODFELLOW et al., 2016; GÉRON, 2019). Typically, CNNs use the softmax exponential function (Softmax Activation) in the output layer for classification problems, since classes are mutually exclusive and the softmax layer produces a probability for each class.

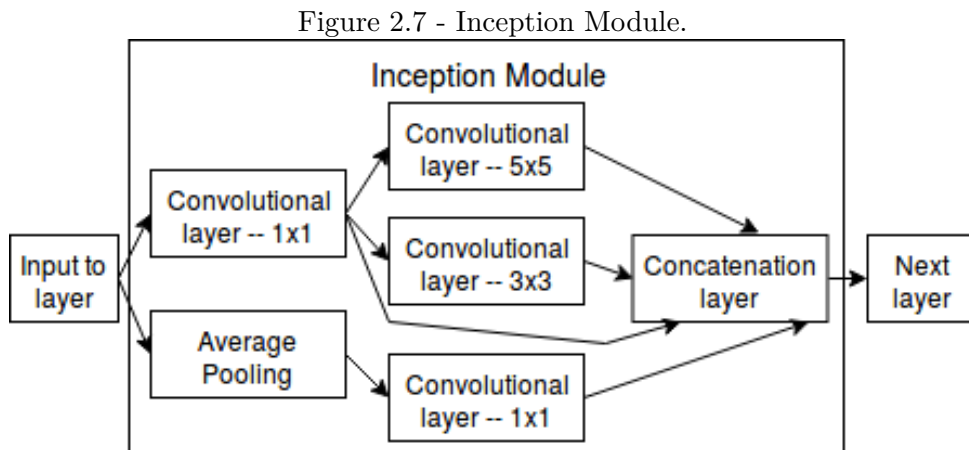
Figure 2.6 - Representation of typical (complex) Convolutional Layer.



Source: Author's production.

2.3.2 GoogleNet Inception

Here, we briefly summarize the main characteristics of the selected deep neural network architecture, GoogleNet Inception, the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. The success of this architecture is mostly due to its deepness (22 complex layers) and the presence of nine Inception Modules, sub-networks that compose the main network. Inception Module captures complex patterns at various scales. It convolves different sizes in parallel, from the smallest possible (1x1) to 5x5 kernel matrices, thus, the network is capable of detecting the presence of thin and rough structural patterns. The output layer of the module is a max pooling to summarize the information from the previous layer by concatenating all results from previous layer to pass them forward to the next one (SZEGEDY et al., 2015; GÉRON, 2019). Figure 2.7 is a representation of the Inception Module and Figure 2.8 presents the whole architecture of this network.



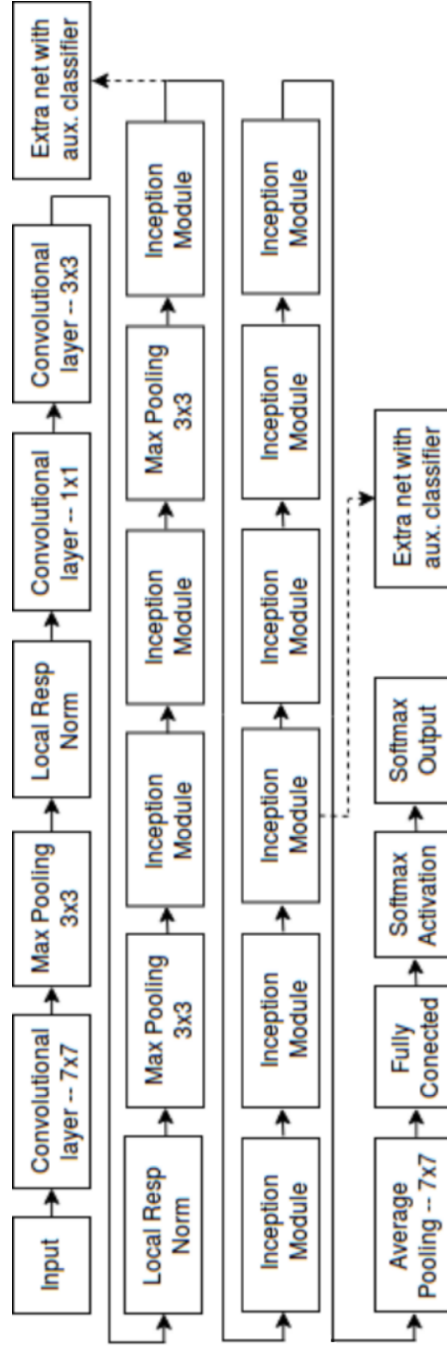
Inspired by Szegedy et al. (2015).

Source: Author's production.

2.4 Machine learning applied to astrophysics

Large astrophysical surveys and the never-ending improving hardware and software related to machine learning have been heating up publications with such interdisciplinary synergy. Ball and Brunner (2010) survey a long list of projects applying data mining for analyzing astronomical data. Ivezić et al. (2014) provides modern statistical methods for analyzing astronomical data. Vasconcellos et al. (2011) employ decision tree classifiers for star/galaxy separation.

Figure 2.8 - GoogleNet architecture.



Convolution (blue), pooling (red), softmax (yellow) and concatenation/normalization (green) layers.
 Source: Szegedy et al. (2015).

Among the different astrophysical knowledge branches, Galaxy morphology plays a crucial role for developing the interdisciplinary synergy between computer science and astrophysics. [Schawinski et al. \(2017\)](#) use Generative Adversarial Networks (GAN) to recover features in astrophysical images of galaxies. [Khalifa et al. \(2017\)](#) reviews the literature and achieves unprecedented success for classifying galaxies into three morphological classes: Elliptical, Spiral and Irregular types. By using a Convolutional Neural Network (CNN) architecture with 8 layers (1 for input, 1 for output and 6 hidden layers), they achieved an accuracy of 97.27% using as dataset 13,000 images from the EFIGI (*Extraction de Formes Idealisées de Galaxies en Imagerie*) catalog ([BAILLARD et al., 2011](#)).

In an innovative approach, [HUERTAS-COMPANY et al. \(2018\)](#) use CNN upon data from simulations to train the neural network for identifying blue nuggets in real-world data, i.e., an astrophysical phenomenon has been predicted by cosmological simulations.

Inside the scope of this PhD, [Barchi et al. \(2016\)](#) present preliminary results for classifying galaxies using the traditional machine learning approach. With this research in an early stage, I achieved significant success in separating ETGs from LTGs with overall accuracy $\geq 97\%$ for all supervised methods explored.

Applying Deep Learning methodologies in galaxy morphology, [Dieleman et al. \(2015\)](#), [Huertas-Company et al. \(2015\)](#), [Sánchez et al. \(2018\)](#) present different catalogs of galaxies. Highlight to [Sánchez et al. \(2018\)](#) who uses Galaxy Zoo 2 questions and answers to try to replicate the answers from the users and presents classification by T-Type. T-Type is a number for determining morphogogical types: ETGs have $T\text{-Type} \leq 0$ and LTGs have $T\text{-Type} > 0$ ([VAUCOULEURS, 1963](#)). T-Type considers ellipticity and spiral arms strength but does not reflect the presence or absence of the bar feature in spirals.

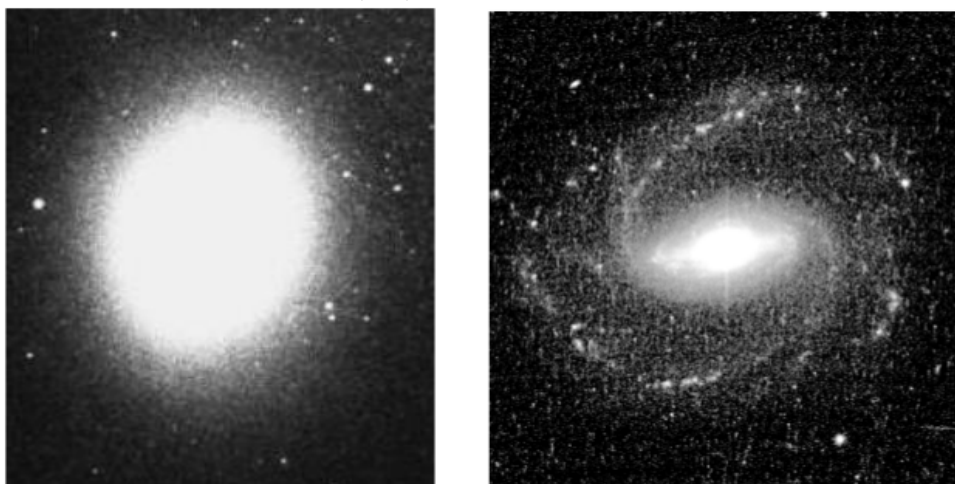
3 MATERIAL AND METHODOLOGY

Analogously to Chapter 2, this chapter presents additional information about material and methodology used for publishing the paper presented in Chapter 4. Here, we briefly introduce Sloan Digital Sky Survey (SDSS), provide more details on the system for extracting non-parametric galaxy morphology features, and more information on Machine and Deep Learning experiments as well.

3.1 Sloan Digital Sky Survey

One of the most handled astronomical datasets is the Sloan Digital Sky Survey — SDSS (EISENSTEIN et al., 2011) — which is acquiring photometry from the northern sky since 1998. After its first two phases, its Data Release 7 (DR7) has publicly released photometry for 357 million unique sources, and it is expected to have nearly 15 terabytes of data when the survey is complete (IVEZIĆ et al., 2014). This massive data set is just one of hundreds of surveys that are being produced continuously by several institutions. In view of their voluminous size, much of these data are never looked at, and therefore the potential extraction of information from these collected data is partially realized. Even though many answers for questions of the contemporary science depends on the processing of such data (FEIGELSON; BABU, 2006; TAN et al., 2005). Figure 3.1 shows examples of galaxies from SDSS-DR7. Section 3.2 presents more details about samples and data used in this work.

Figure 3.1 - Examples of ETG (left) and LTG in black and white from SDSS-DR7.



Source: Author's production.

3.2 Sample and data

This work uses data acquired from the SDSS-DR7 (EISENSTEIN et al., 2011) and Galaxy Zoo catalogs (LINTOTT et al., 2008; LINTOTT et al., 2011; WILLETT et al., 2013) for measuring morphology and training the classification models. The samples are composed of galaxies in r-band from SDSS-DR7 in the redshift range $0.03 < z < 0.1$, Petrosian magnitude in r-band brighter than 17.78 (spectroscopic magnitude limit), and $|b| \geq 30^\circ$, where b is the galactic latitude.

For supervised learning purposes, we consider the defined classification from Galaxy Zoo 1 — GZ1 hereafter (LINTOTT et al., 2008; LINTOTT et al., 2011) between E and S galaxies, and the classification from Galaxy Zoo 2 — GZ2 (WILLETT et al., 2013) with prefixes in one of 11 following classes: Er, Ei, Ec, Sa, Sb, Sc, Sd, SBa, SBc, SBd (see Figures 2.2 and 2.3, respectively). Other three different scenarios are explored with GZ2 supervision. Classification considering 9 classes (same as 11 classes except that we have one class for all elliptical galaxies united), 7 classes (same as previous but disregarding the faintest galaxy types: Sd and SBd) and three classes: E, S and SB.

We study the impact of different datasets on the training process, varying the number and size of objects in the samples. We define a parameter K as the area of the galaxy’s Petrosian ellipse divided by the area of the Full Width at Half Maximum (FWHM). Equation 3.1 presents how to calculate K , where R_P is the Petrosian radius — see Petrosian (1976), Eisenstein et al. (2011) for more details about R_P . By restricting the samples to a minimum K , we limit the number and size of objects in the dataset. The number of galaxies for the three main samples we explore ($K \geq 5$, $K \geq 10$ and $K \geq 20$) are presented in Table 3.1.

$$K = \left(\frac{R_P}{\text{FWHM}/2} \right)^2 \quad (3.1)$$

Table 3.1 - Number of galaxies for the main samples in this work from each database (SDSS, GZ1 and GZ2).

Restriction	Number of galaxies in		
	SDSS	GZ1	GZ2
$K \geq 5$	239,833	104,787	138,430
$K \geq 10$	175,167	89,829	110,163
$K \geq 20$	96,787	58,030	67,637

Source: Barchi et al. (2020).

With smaller values of K we have more but smaller objects, while samples restricted by bigger values of K have less but bigger objects. To properly check the impact of the number and sizes of objects in the samples, we explore the Deep Learning approach for three classes problem in detail with other restrictions: $K \geq 7$, $K \geq 9$, $K \geq 11$, $K \geq 14$ and $K \geq 17$.

For Machine and Deep Learning experiments, we split the datasets from GZ1 e GZ2 into training-validation-test subsets in the proportion 80-10-10. In all experiments, each of these subsets are constrained to the same restriction (the model trained and validated with a subset restricted to $K \geq 20$ is also tested with the subset restricted to $K \geq 20$). We should keep in mind that the data used in this work, SDSS-DR7, does not have a proper spatial resolution (0.396 arcsec pixel⁻¹) and not adequate PSF FWHM (~ 1.5 arcsec). For comparison, the Dark Energy Survey — DES (ABBOTT et al., 2016), has a pixel size of (0.27 arcsec) and PSF FWHM of ~ 0.9 arcsec. This is why we study the quality of our classification as a function of K .

3.3 CyMorph - non-parametric galaxy morphological system

CyMorph is the non-parametric galaxy morphological system which process Concentration (C), Asymmetry (A), Smoothness (S), Entropy (H) and Gradient Pattern Analysis (GPA) metrics¹.

We perform a query in SDSS DR7 database to obtain the data table which is input to the system. This data table has the required information to download each field of view (which is a fits image file — example in Figure 3.2.a). It must contain the

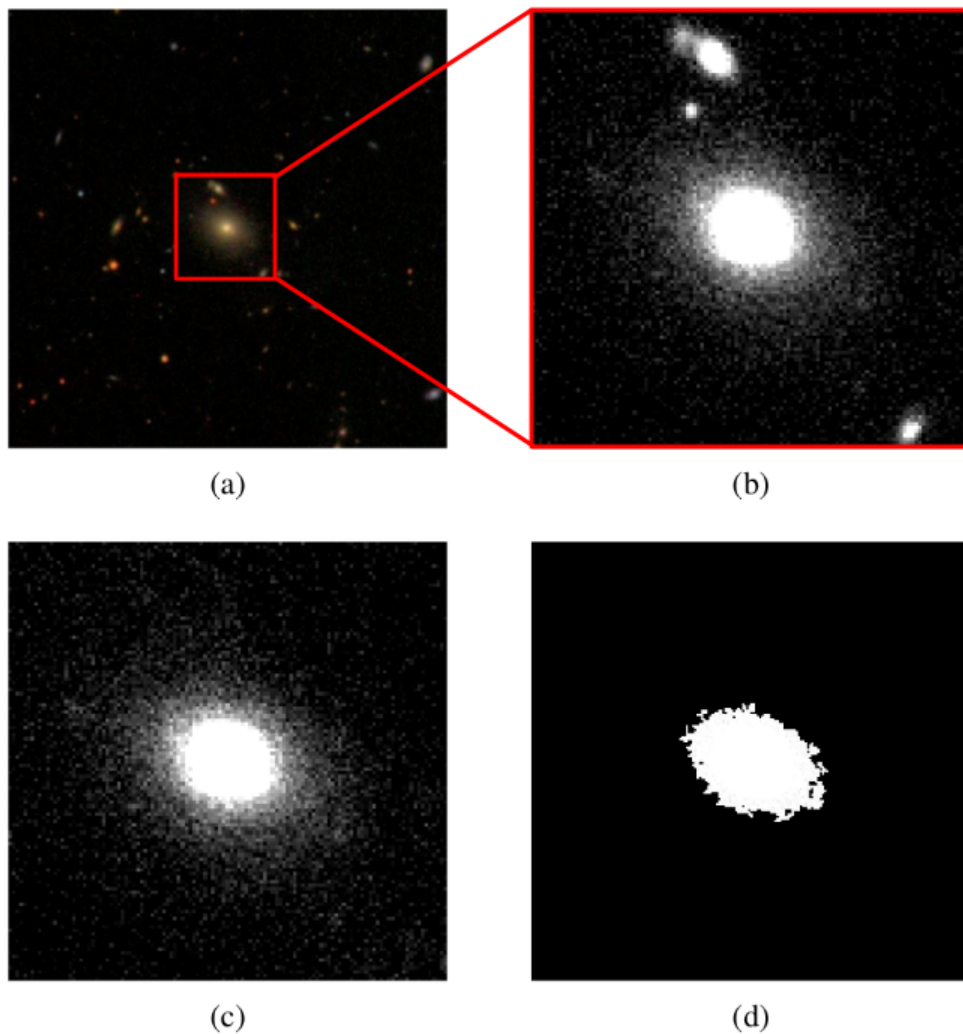
¹The whole system code is available on this repository: <https://github.com/paulobarchi/CyMorph>.

coordinates (RA and DEC) for each galaxy of interest as well. Alternatively, the system accepts the path to the directory with all fits files as input.

3.3.1 Preprocessing

Before starting to process the non-parametric morphological metrics, incisive image preprocessing techniques are mandatory for morphology. Preprocessing ensures the consistency of parameters and improves the feature extraction. For the morphological analysis, there are three major issues in preprocessing: cut the stamp, remove secondary objects, and generate the segmented image. Source Extractor — SExtractor (BERTIN; ARNOUITS, 1996) — is employed for photometry and automated detection of sources in fits image files. In the first step, the desired galaxy is selected from the field of view image. Then, an interpolation is performed: pixels from secondary objects that remain inside the stamp are replaced by the target’s isophotal level. For each pixel, the isophotal level is obtained using a random value from a Gaussian distribution on the aimed object expanded ellipse that intersects the pixel. For all metrics except Concentration, we use as input the segmented image, which is generated by applying the mask (obtained by region growing algorithm from the center of the galaxy of interest) upon the cleaned image — see Figure 3.2. Concentration does not use the segmented image, instead, it uses the original galaxy stamp to get the whole accumulated flux profile of each object.

Figure 3.2 - Preprocessing example.



From original field of view (a), the stamp is cut (b), cleaned (c) and segmented by mask (d).

Source: Author's production.

3.3.2 Error detection

It is possible to verify if the galaxy can have problems when calculating non-parametric metrics. The first verification is to assert that the stamp does not have n objects inside the area from the center of the galaxy to 2 times the galaxy radius ($4 \times \pi \times R^2$). If the galaxy has more objects than n in this area, this galaxy has the error flag 1 assigned to it. We set $n = 10$, empirically.

The next verification is associated with the flux profile computed from the galaxy.

If convergence is not reached in this calculus, an error flag 2 is assigned to the galaxy. If none of these problems are identified, the galaxy has error flag equals 0 at the end of preprocessing phase. Galaxies with error flags different from 0 are not processed by the non-parametric morphological system to avoid galaxy images with possible several problems: central double peak; galaxy at the edge of the field; many objects of similar brightness superimposed in the field; and merging galaxies. These problems are mostly identified when calculating the Petrosian radius (R_p). Galaxies with $Error \neq 0$ are disconsidered to build up the classification models and to be classified in the traditional machine learning approach.

All error flags are mapped as follows: $Error = 0$: success (no errors); $Error = 1$: many objects of significant brightness inside $2 \times R_p$ of the galaxy; $Error = 2$: not possible to calculate the galaxy's R_p ; $Error = 3$: problem calculating GPA; $Error = 4$: problem calculating H; $Error = 5$: problem calculating C; $Error = 6$: problem calculating A; $Error = 7$: problem calculating S.

3.3.3 Concentration

Morgan and Mayall (1957) proposed the first concentration index, which is given by the ratio of the distance that contains 80% brightness² ($R_{80\%}$) of the observed galaxy, and the distance that contains 20% brightness ($R_{20\%}$) of the observed galaxy as shown in Equation 3.2. Notice that fixing the brightness value, the ratio between this distances indicates the galaxy profile slope.

$$C_1 = \log_{10} \left(\frac{R_{80\%}}{R_{20\%}} \right) \quad (3.2)$$

Further improvement in this index associate other ratios of brightness proportion radius, since the distance measurement is affected by the image sky, and seeing effects in the center of the galaxies (FERRARI et al., 2015). In order to avoid the smoothing effect from galactic center, Kent (1985) proposed another concentration index (C_2) based on a ratio between $R_{90\%}$ and $R_{50\%}$.

In literature, empirical techniques are usually applied for the accumulated flux profile estimation (CONSELICE, 2003; FERRARI et al., 2015), another approach is to associate to a galaxy parametric model, as the Sérsic index (GRAHAM et al., 2001). Both techniques have issues, for instance empirical techniques have problems associated

²A galaxy image brightness is the integrated flux in a given region.

to seeing effects (FERRARI et al., 2015), whereas parametric models are inaccurate for late-type galaxies (GRAHAM et al., 2001).

Following the basic methods from these authors and after an empirical analysis process, we propose to calculate the Concentration index with the following steps: (1) calculate the flux profile, (2) calculate the η_P function, (3) obtain the Petrosian Radius (R_P) and (4) calculate the log ratio of the relative Petrosian flux. Each step is explained below.

(1) The accumulated flux profile is given by the sum of light flux within each circular radius. For each circular radius R , we calculate the flux inside it as follows:

$$F(R) = \sum_{i,j=0,0}^{N,N} (m_{i,j} - B) \times P \quad (3.3)$$

where

- N is the dimension of the matrix (image).
- $m_{i,j}$ is the intensity of the pixel from the matrix in coordinates (i, j) .
- B is the median background intensity value. We have defined four background corners as the outer most submatrices (top left, top right, bottom left and bottom right) with dimensions $width/5 \times height/5$ for defining B . B is the median of all these regions.
- P is the percentage of points within the pixel that are inside the galaxy ellipse. If the pixel is totally inside the galaxy ellipse, $P = 1.0$; if the pixel is totally outside the galaxy ellipse, $P = 0.0$; else, given 1000 random points inside the pixel, P is the fraction of these points which are inside the galaxy ellipse.

The accumulated flux profile is the calculus of the above equation for each radius.

(2) The $\eta_P(R)$ function provides the average intensity within some projected radius R divided by the intensity at that radius. In order to objectively calculate the $\eta_P(R)$ function in a ring area inner centered in R , we adopted a modified form of the Petrosian system (PETROSIAN, 1976) as Sloan Digital Sky Survey (SDSS) does (EISENSTEIN et al., 2011).

$$\eta_P(R) = \frac{\frac{F(1.25 \times R) - F(0.8 \times R)}{\pi(1.25^2 - 0.8^2)R^2}}{\frac{F(R)}{\pi \times R^2}} \quad (3.4)$$

At the end of this step, we have the accumulated flux profile and the η_P for each radius (R) of the image.

(3) Still within the modified form of the Petrosian system we adopted, the Petrosian radius R_P is given by the radius R which have a corresponding $\eta_P(R) = 0.2$. In most cases, where there is not an exact R_p for which $\eta_P(R) = 0.2$, we used linear interpolation to obtain the desired R_P .

(4) Following the literature (FERRARI et al., 2015), we primarily test two configurations for the Concentration index:

$$C_1 = \log_{10} \left(\frac{R_{80\%}}{R_{20\%}} \right) \quad (3.5)$$

$$C_2 = \log_{10} \left(\frac{R_{90\%}}{R_{50\%}} \right) \quad (3.6)$$

An optimization work is performed in order to define which ratio better characterizes the observed galaxies. This process is described in Subsection 4.2.2: a new methodology for setting the best configuration parameters for this galaxy morphological system. Among all configurations tested, the best configuration is:

$$C = \log_{10} \left(\frac{R_{75\%}}{R_{35\%}} \right) \quad (3.7)$$

3.3.4 Asymmetry

The Asymmetry index is simply given by the comparison of the source image with its π -rotated variant. Since this metric is commonly applied to characterize high-redshift galaxies (CONSELICE, 2003), a diversity of equations and enhancement processes were proposed. A popular version of Asymmetry is given by the Equation 3.8 (ABRAHAM et al., 1996). In this equation, $I_{i,j}^\pi$ is the (i, j) pixel intensity after subtracting the background, and π is the angle of rotation in radians. Notice that each term of the sum is weighted by $|I_{i,j}^0|$. This weight enhances the spiral disk region,

since it usually has lower intensity value than the galaxy central region. However, the sky is also enhanced in this process since it has a low flux intensity. Consequently, one of the main tasks is how to segment the images. We may expect higher values of Asymmetry for Late-Type Galaxies using Equation 3.8.

$$A_A = \sum_{i,j}^{N,N} \frac{|I_{i,j}^0 - I_{i,j}^\pi|}{|I_{i,j}^0|} \quad (3.8)$$

Among several previous works which contributed with this morphological metric — with similar computation to Equation 3.8 (ABRAHAM et al., 1996; CONSELICE, 2003; LOTZ et al., 2004), we follow the strategy by (FERRARI et al., 2015). The Asymmetry index is measured using correlation coefficients — see Equations 3.9 and 3.10. The functions $r()$ and $s()$ are, respectively, the Pearson rank and the Spearman rank (PRESS, 2005). The advantage of the correlation coefficients is the robustness to the visual effects and the interference of the sky in the measurement.

$$A_r = 1 - r(I^0, I^\pi) \quad (3.9)$$

$$A_s = 1 - s(I^0, I^\pi) \quad (3.10)$$

From now on, we refer to asymmetry as A , defined by Equation 3.10 (details on metric configuration and selection in Subsection 4.2.2).

3.3.5 Smoothness

Classically, the Smoothness is measured as the weighted difference between the image (composed by elements $I_{i,j}^0$) and its smoothed version (composed by elements $I_{i,j}^s$), according to the Equation 3.11, where $B_{i,j}$ is the background intensity value (CONSELICE, 2003).

$$S_C = 10 \times \sum_{i,j}^{N,N} \frac{(I_{i,j}^0 - I_{i,j}^s) - B_{i,j}}{I_{i,j}} \quad (3.11)$$

In recent works, this parameter has been improved by adopting correlation coefficient (FERRARI et al., 2015). The advantages to this approach are (a) to characterize levels of flux intensity and (b) the robustness to local noise. The smoothness parameters

are measured according to Equations 3.12 and 3.13, where I^0 is the flux intensity in the original image, and I^s is the flux intensity in the smoothed image.

$$S_r = 1 - r(I^0, I^s) \quad (3.12)$$

$$S_s = 1 - s(I^0, I^s) \quad (3.13)$$

3.3.6 Gradient pattern analysis applied to galaxy morphology

Gradient Pattern Analysis (GPA) is a novel metric to galaxy morphological analysis introduced by Rosa et al. (2018). Given a matrix, the local gradient is calculated as the first partial difference of $I(x_i, y_i)$ with respect to each neighbour element in the matrix. The operation returns the x and y components of the two-dimensional numerical gradient, ΔM , that can be described in terms of its symmetry, and the local vector characteristics (norm and orientation). ΔM can be represented as a composition of the following four gradient patterns (GPs):

- GP1: the matrix representation of the total vector distribution ΔM ;
- GP2: the matrix of the respective norms;
- GP3: the matrix of the respective phases;
- GP4: the matrix of the respective complex numbers.

For each type of matrix pattern from the set $GP1, GP2, GP3, GP4$ we can calculate specific parameters as the respective gradient moments G_1, G_2, G_3, G_4 , where each is extracted from its respective matrix pattern, namely vector, norm, phase and complex representations.

Traditionally, G_1 and G_2 Equations are presented as follows:

$$G_1 = \begin{cases} \frac{T_A - V_A}{V_A} & \text{if } T_A \geq V_A, \\ 0 & \text{otherwise.} \end{cases} \quad (3.14)$$

$$G_2 = \frac{V_A}{V} \left(2 - \frac{|\sum_i^{V_A} v_i|}{\sum_i^{V_A} |v_i|} \right). \quad (3.15)$$

where V is the total amount of gradient vectors, V_A is the amount of asymmetric vector after removing all the symmetric pairs, T_A is the amount of edges (after performing a Delaunay triangulation), and v_i is i^{th} asymmetrical vector norm.

In G_1 , all vectors that belongs to the sky (background) are forced to be symmetrical, since G_1 depends only on the ratio between the number of Delaunay connections and the number of asymmetrical vectors. However, in this work we do not apply the mask for G_1 , as it has been observed better results for G_1 without segmentation mask.

Concerning GPA performance for characterizing galaxy images, a small change on G_2 is proposed to adapt it to work with segmentation masks. In symmetry detection step, if a pixel in a position (i, j) is detected as sky (background), then the gradient at position (i, j) is ignored, decreasing the total number of vectors N (ROSA et al., 2018).

3.3.7 Entropy

Entropy is a measurement of the distribution of information in the object of analysis. In digital image processing, it measures the distribution of pixel values in the image. We adopt the Shannon entropy, given by Equation 3.16, already in use in galaxy morphology by Ferrari et al. (2015). Assuming the galaxy flux as the random variable, this measurement shows the heterogeneity degree in pixel distribution.

$$H = - \frac{\sum_k^K p(I_k) \log(p(I_k))}{\log(K)} \quad (3.16)$$

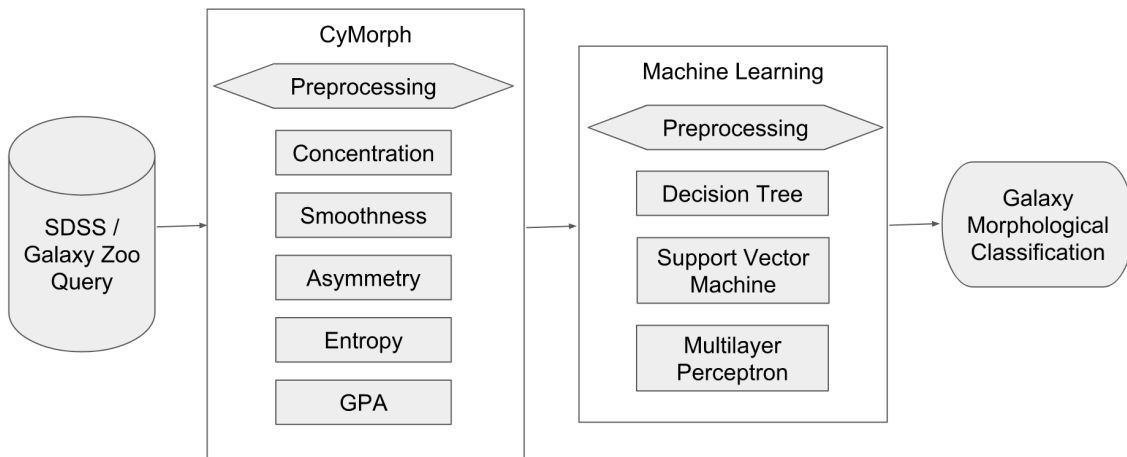
To calculate the Equation 3.16, we must separate the flux intensity in bins: k is the k^{th} bin. Thus, it is fundamental to establish the number of bins (K) to use in this process. We apply an objective function to determine the optimal number of bins. Section 4.2.1 and Subection 4.2.2 describe the objective method and experiments performed, respectively.

3.4 Machine learning applied to galaxy morphology

Traditionally, the main division in ML is with regard to the supervision (or absence of it) in the learning process. Supervised learning approaches have a guide for building up the model, while unsupervised learning do not have a supervision to guide the learning process (MITCHELL, 1997; GÉRON, 2019; GOODFELLOW et al.,

2016). Supervised learning algorithms learn to associate input (features) with output (labels or floating values) given the training dataset. In this work, we focus on supervised learning as our main goal is to reproduce the human eye in morphological classification of galaxies and we have Galaxy Zoo 1 and 2 catalogs for guiding the learning processes. In this work, we focus on supervised learning as our main goal is to reproduce the human eye in morphological classification of galaxies, using morphological metrics and catalogs from Galaxy Zoo 1 and 2 to guide the learning process (as described in Section 3.2). We maintain the restriction related to the area of the galaxies to build up different classification models: (1) $K \geq 5$; (2) $K \geq 10$; and (3) $K \geq 20$, i.e., the area of the galaxy is at least five (model 1), ten (model 2) and twenty (model 3) times larger than the Full Width at Half Maximum (FWHM) area for each corresponding object, respectively. We build up Decision Tree (DT), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) models to classify galaxies considering different numbers of classes. We use `scikit-learn` (PEDREGOSA et al., 2011) python library to perform the experiments and procedures reported in this document (see more about computational details in Chapter 5).

Figure 3.3 - Traditional Machine Learning schema.



Source: Author's production.

For each classification model, it is necessary to address how to split the dataset into training, validating and testing sets. If no refined strategy is adopted to perform this partitioning, the results may be biased to the sets selected. Cross-Validation

(CV) is the most common procedure used to overcome this challenge. CV addresses the trade-off between bias and variance by slicing the dataset in k -folds ($k = 10$, 10 folds with 10% of the data), and the learning process is repeated k times, each time combining different folds to compose training and testing sets, maintaining the defined proportions. In this work, first we split the dataset in a 90/10 proportion for training and testing sets, respectively. CV is applied in the 90% portion of the dataset. CV imposes the learning process to occur ten times, each time using all possible different combinations of folds. The resulting model is then validated on the remaining part (10%) of the data. Besides CV to properly perform the dataset partitioning, we use Grid Search (GS) to exhaustively generate and test values for parameter candidates in Decision Tree (DT) models. This process is also known as hyperparameter optimization. Hyperparameter values are not adjusted by the learning process itself. The hyperparameter optimization tests all possible combinations of parameter values by automatically performing the CV step (described above), fitting the model and computing the score on the validation set. The best configuration is used for the definitive model. DT parameters are associated with the tree depth. The training process is feasible when applying GS to DT and the results improve significantly. For SVM and MLP, there are many more parameter values to test in GS. Preliminary tests exploring problems with two and three classes show that the results are equivalent whether or not we use GS. We do not use GS in SVM and MLP experiments because the computational cost is very high and it would have a minimal impact in the results.

3.5 Deep learning applied to galaxy morphology

In this work, we focus on two notable robust CNN architectures, Residual Networks –ResNet (HE et al., 2016) — and GoogleNet (SZEGEDY et al., 2015), judging overall accuracy performance and training time. We split the datasets into training-validation-test subsets in the proportion 80-10-10. The training phase for all experiments consist of 30 epochs, which is enough since we always reach convergence with 30 or less epochs. The batch size is limited by available hardware — enough to reach full capacity of 2 TESLA P100 GPUs (more hardware details in 5). ResNet consumes more memory since it has a deeper architecture with 50 layers. The batch size for ResNet is 32 and, for GoogleNet, 128. With these characteristics defined, we perform experiments to answer the other questions. We first test ResNet and GoogleNet on raw data (no preprocessing). We perform experiments on three different datasets restricted by $K \geq 5$: two classes from GZ1, Lintott et al. (2008), LINTOTT et al. (2011); and three classes from GZ2, Willett et al. (2013) with

Table 3.2 - Experiments with different configurations for CNN architectures.

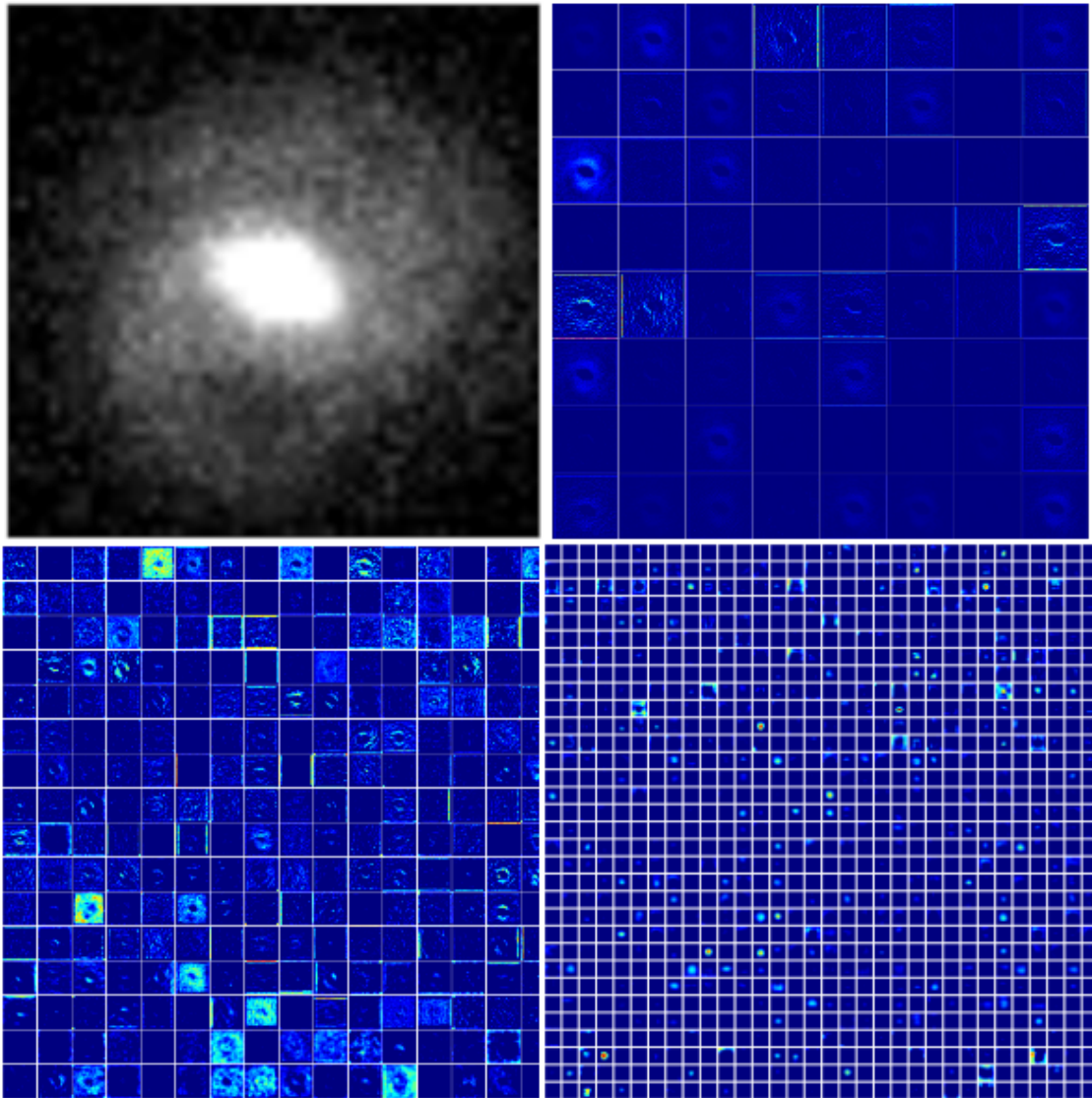
	ResNet		GN (r)		GN (p)	
	t	OA	t	OA	t	OA
2c unb	3:40	85.4	2:33	98.7	2:33	98.7
3c unb	4:17	51.6	1:36	78.2	2:01	80.8
3c bal	2:54	34.0	0:45	73.6	0:45	75.0

Summary of experiments with different configurations for CNN architectures: ResNet (HE et al., 2016), GoogleNet — GN (SZEGEDY et al., 2015) — without pre-trained model (r) and GN using pre-trained model (p). The datasets are: imbalanced dataset considering two classes (2c unb), imbalanced dataset considering three classes (3c unb), balanced dataset considering three classes (3c bal). We show processing time t (format: h:mm) and Overall Accuracy (OA in percentage).

Source: Author’s production.

2 variations, imbalanced dataset and balanced dataset (see Subsection 4.4.2 about the class imbalance problem). Table 3.2 presents the results regarding these experiments to establish the network configuration. Considering the same datasets, we address the issue of using or not pre-trained models. We compare the performance of the default GoogleNet network (without pre-trained weights) and with weights from a network already trained for another problem. We obtain the same OA, but pre-trained models present better results considering processing time and number of epochs. Finally, we set up an experiment to answer the question about using raw data or clean images (preprocessed data). We apply GoogleNet on the dataset limited by $K \geq 20$ for the two classes problem. Both models perform equivalently regarding OA ($\sim 99.5\%$), but we choose to use raw images since we avoid the preprocessing step. After these tests, we select GoogleNet with pre-trained weights and raw data for building up the classification models. Just to give examples of what kind of features deep neural networks extract from these galaxies, Figure 3.4 presents the output of convolutions performed by Inception Module through different stages of the network.

Figure 3.4 - Example of convolutions applied to a galaxy for illustration purposes.



Top left: the input image of a galaxy in r-band. Top right: the output of the first convolution performed. Bottom left: the output of the first Inception Module. Bottom right: the output of the last Inception Module of the neural network.

Source: Author's production.

4 MACHINE AND DEEP LEARNING APPLIED TO GALAXY MORPHOLOGY - A COMPARATIVE STUDY

4.1 Introduction

¹In observational cosmology, the morphological classification is the most basic information when creating galaxy catalogs. The first classification system, by [Hubble \(1926\)](#), [Hubble \(1936\)](#), distinguishes galaxies with dominant bulge component — also known as Early-Type Galaxies (ETGs) — from galaxies with a prominent disk component — named Late-Type Galaxies (LTGs). LTGs are commonly referred to as spiral galaxies because of their prominent spiral arms, while ETGs are commonly referred to as elliptical (E) galaxies as they have a simpler ellipsoidal structure, with less structural differentiation (less information). More refined classifications fork spirals into two groups: barred (SB) and unbarred (S) galaxies. These two groups can also be refined even further by their spiral arms strength. A number known as T-Type can be assigned to the morphological types: ETGs have T-Type ≤ 0 and LTGs have T-Type > 0 ([VAUCOULEURS, 1963](#)). T-Type considers ellipticity and spiral arms strength but does not reflect the presence or absence of the bar feature in spirals.

Morphology reveals structural, intrinsic and environmental properties of galaxies. In the local universe, ETGs are mostly situated in the center of galaxy clusters, have a larger mass, less gas, higher velocity dispersion, and older stellar populations than LTGs, which are rich star-forming systems ([ROBERTS; HAYNES, 1994](#); [BLANTON; MOUSTAKAS, 2009](#); [POZZETTI et al., 2010](#)). By mapping where the ETGs are, it is possible to map the large-scale structure of the universe. Therefore, galaxy morphology is of paramount importance for extragalactic research as it relates to stellar properties and key aspects of the evolution and structure of the universe.

Astronomy has become an extremely data-rich field of knowledge with the advance of new technologies in recent decades. Nowadays it is impossible to rely on human classification given the huge flow of data attained by current research surveys. New telescopes and instruments on board of satellites provide massive datasets. Therefore, in view of their voluminous size, much of the data are never explored. The

¹This chapter is an adapted version of the paper: [BARCHI, P.; CARVALHO, R. de; ROSA, R.; SAUTTER, R.; SOARES-SANTOS, M.; MARQUES, B.; CLUA, E.; GONÇALVES, T.; SÁ-FREITAS, C. de; MOURA, T. Machine and deep learning applied to galaxy morphology-a comparative study. *Astronomy and Computing*, v. 30, p. 100334, 2020. I have incorporated content from the original paper in the previous and next chapters, and suppressed it here for better comprehension of the thesis.](#)

potential extraction of knowledge from these collected data is only partially accomplished, even though many answers of the contemporary science critically depend on the processing of such large amount of data (WAY et al., 2012; IVEZIĆ et al., 2014; FEIGELSON; BABU, 2006). Automatic classification can address this bottleneck of observational research.

One of the most used astronomical datasets is the Sloan Digital Sky Survey — SDSS, which has been acquiring photometry from the northern sky since 1998. After its first two phases, SDSS Data Release 7 has publicly released photometry for 357 million unique sources, and it is expected to be around 15 terabytes of data when the survey is complete (EISENSTEIN et al., 2011). This massive dataset is just one of hundreds of surveys that are currently underway.

One effort to overcome the challenge to classify hundreds of thousands of galaxies depends on the laborious engagement of many people interested in the subject. Galaxy Zoo is a citizen science project which provides a visual morphological classification for nearly one million galaxies in its first phase (Galaxy Zoo 1) distinguishing elliptical from spiral galaxies. With general public help, this project has obtained more than 4×10^7 individual classifications made by $\sim 10^5$ participants. In its second phase, Galaxy Zoo 2 extends the classification into more detailed features such as bars, spiral arms, bulges, and many others, providing a catalog with nearly 300 thousand galaxies present in SDSS. Throughout this work, we use Galaxy Zoo (LINTOTT et al., 2008; LINTOTT et al., 2011; WILLETT et al., 2013) classification as supervision and validation (ground truth) to our classification models.

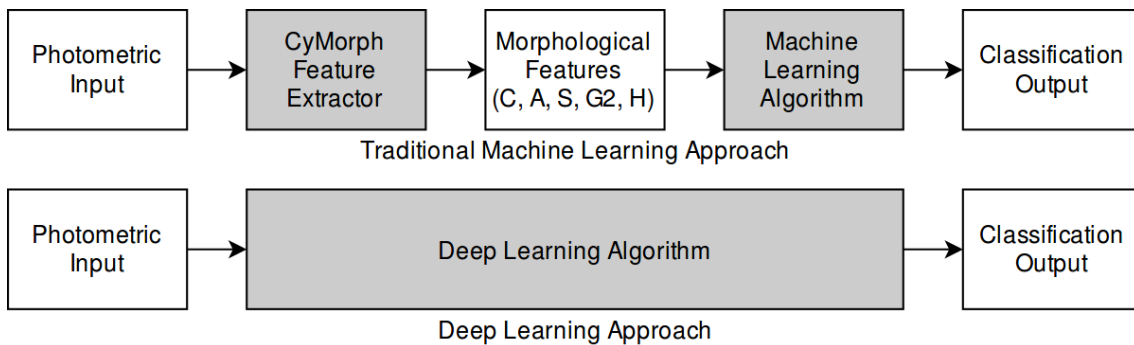
Several authors (ABRAHAM et al., 1996; CONSELICE et al., 2000; CONSELICE, 2003; LOTZ et al., 2004) studied and presented results about objective galaxy morphology measures with Concentration, Asymmetry, Smoothness, Gini, and M20 (also known as CASGM or CAS system). Ferrari et al. (2015) introduced the entropy of information H (Shannon entropy) to quantify the distribution of pixel values in the image. Rosa et al. (2018) introduced the Gradient Pattern Analysis (GPA) technique to separate elliptical from spiral galaxies by the second moment of the gradient of the images. This whole system used by Rosa et al. (2018) — called CyMorph — is described in this paper (Section 4.2).

It is not trivial to determine the success of each non-parametric morphological parameter to perform this classification task. Considering the separation between elliptical and spiral galaxies, for example, a morphological parameter is more reliable if it maximizes the separation of the distributions of these two types. Rosa et al.

(2018) described the evaluation technique proposed and adopted to measure the success of metrics to separate elliptical from spiral galaxies — see also Subsection 4.2.1 and Sautter and Barchi (2017).

The main purpose of this investigation is to answer the question “How to morphologically classify galaxies using Galaxy Zoo (LINTOTT et al., 2008; LINTOTT et al., 2011; WILLETT et al., 2013) classification through non-parametric features and Machine Learning methods?” We also apply Deep Learning techniques directly to images to overcome the same challenge and compare results from both approaches. Deep Convolutional Neural Network (CNN) is a well-established methodology to classify images (GOODFELLOW et al., 2016). Without the need of a feature extractor, the network itself adjusts its parameters in the learning process to extract the features. Figure 4.1 shows both flows for each approach used in this work: Traditional Machine Learning (TML) and Deep Learning (DL).

Figure 4.1 - Illustrative sketch of traditional machine learning and deep learning flows.



Source: Barchi et al. (2020).

The huge amount of photometric astrophysical data available and the highly increasing advancements on hardware and methods to perform automatic classifications has been leveraging related publications (LAW et al., 2007; FREEMAN et al., 2013; KHALIFA et al., 2017; HUERTAS-COMPANY et al., 2018; BARCHI et al., 2016; DIELEMAN et al., 2015; KHAN et al., 2018; HUERTAS-COMPANY et al., 2015; SÁNCHEZ et al., 2018). Highlight to Sánchez et al. (2018) who use questions and answers from Galaxy Zoo 2 for replicating the answers from the users, and provide morphology classification

by T-Type in their final catalog.

The approach used in this paper is different from the one used in [Sánchez et al. \(2018\)](#). Instead of using questions and answers from Galaxy Zoo 2, we use the classifications and images themselves. Also, we revisit issues not touched upon in previous studies dealing with morphological parameters ([ABRAHAM et al., 1996](#); [CONSELICE et al., 2000](#); [CONSELICE, 2003](#); [LOTZ et al., 2004](#)); namely, threshold dependence in the use of the segmented image. We study the impact of that on the parameters that ultimately will be used in the TML approach.

Although it is already a well-established observation that for perception tasks (which galaxy morphology is) Deep Learning is likely to outperform machine learning models trained on hand engineered features ([RUSSAKOVSKY et al., 2015](#)), this subject is in its infancy in galaxy morphology and such comparison of these two approaches have never been presented in the same work in the literature. Also, deep learning methods need huge amounts of data to learn from and huge computational resources to make it effective. Deep learning models can be hard to tune and tame, and the prediction time can take much longer than other models because of the complexity ([GOODFELLOW et al., 2016](#)). The traditional machine learning approach is still relevant.

4.2 Advances in non-parametric galaxy morphology

Methodologies for computing non-parametric morphological metrics have been presented by several authors ([MORGAN; MAYALL, 1957](#); [KENT, 1985](#); [ABRAHAM et al., 1996](#); [TAKAMIYA, 1999](#); [CONSELICE, 2003](#); [LOTZ et al., 2004](#); [FERRARI et al., 2015](#); [ROSA et al., 2018](#)). In this section, we present CyMorph — a non-parametric galaxy morphology system which determines Concentration (C), Asymmetry (A), Smoothness (S), Entropy (H) and Gradient Pattern Analysis (GPA) metrics.

We perform image preprocessing techniques to ensure consistency and improve feature extraction. CyMorph achieves this goal in three major steps: producing the galaxy stamp, removing secondary objects, and generating the segmented image. To remove secondary objects inside the stamp we replace their pixels by the median value of the isophotal level that cross that object.

Concentration is the only metric we calculate using the clean galaxy stamp since we want the whole accumulated flux profile of the galaxy. For all other metrics, we use the segmented image as input which we obtain by applying a mask upon the clean

image. The mask is computed by a region growing algorithm (PEDRINI; SCHWARTZ, 2007). We summarize CyMorph metrics as follows:

- a) Concentration is defined as $C = \log_{10}(R_1/R_2)$, where R_1 and R_2 are the outer and inner radii, respectively, enclosing some fraction of the total flux (CONSELICE, 2003; LOTZ et al., 2004; FERRARI et al., 2015). We use an optimization process for setting up the best configuration parameters for CyMorph. The best configuration by this method is: $C = \log_{10}(R_{75\%}/R_{35\%})$.
- b) Asymmetry is measured using the correlation between the original and rotated image: $A = 1 - s(I^0, I^\pi)$, where I^0 and I^π are the original and the π -rotated images. The function $s()$ is the Spearman’s rank correlation coefficient (PRESS, 2005), which has been proved to be a stable and robust correlation coefficient (SAUTTER, 2018).
- c) Smoothness describes the flux dispersion of an image, namely how the gradient varies over the entire image. This can be measured as the correlation between the original image and its smoothed counterpart (ABRAHAM et al., 1996; CONSELICE, 2003; FERRARI et al., 2015). We apply the Butterworth filter for smoothing the images. This filter provides the advantage of a continuous adaptive control of the smoothing degree applied to the image — see Kaszynski and Piskorowski (2006), Pedrini and Schwartz (2007), Sautter (2018) for more details. We use the the Spearman’s rank correlation coefficient to compute smoothness, following the same reasoning as for asymmetry. We define smoothness as $S = 1 - s(I^0, I^s)$, where I^0 is the flux intensity of the original image, and I^s is the flux intensity of the smoothed image.
- d) Gradient Pattern Analysis (GPA) is a well-established method to estimate the local gradient properties of a set of points, which is generally represented in a two-dimensional (2D) space (ROSA et al., 1999; RAMOS et al., 2000; ROSA et al., 2003). We use the improved version of GPA developed for galaxy morphology — see Rosa et al. (2018) and references therein for more details.
- e) In digital image processing, the entropy of information, H — Shannon entropy Bishop (2006) — measures the distribution of pixel values in the image. In galaxy morphology, we expect high values of H for clumpy galaxies because of their heterogeneous pixel distribution, and low H for smooth galaxies — see Ferrari et al. (2015), Bishop (2006) for more details.

For more specific details about how to compute each of these metrics, see Sautter (2018) and references therein.

4.2.1 Geometric histogram separation (δ_{GHS})

For a given sample of galaxies, CyMorph measures C, A, S, H and GPA and these parameters depend on some quantities. Our main goal is to choose the best quantities possible for reaching a maximum performance in classifying galaxies. Using an independent morphological classification (from GZ1, e.g.), we have elliptical and spiral distributions for each parameter. All we need is a simple and reliable method to objectively assign a value for the separation between elliptical and spiral distributions. Here, we measure the geometric distance between the distributions with the GHS (Geometric Histogram Separation) algorithm — see[Sautter and Barchi (2017), Rosa et al. (2018) for more details.

4.2.2 Optimizing morphological metrics configuration

CyMorph has configurable parameters that we have to fine tune for better distinction between different morphological types. One specific configuration is the threshold parameter used in SExtractor (BERTIN; ARNOUITS, 1996) to detect objects on an image: *DETECT_THRESH* (hereafter d_σ). SExtractor detects an object if a group of connected pixels is above the background by a given d_σ . Thus, we want to find the minimum d_σ value, sufficiently above the background limit, for which we do not lose information when computing each metric. Most of the configurable parameters are related to each morphological metric. It is important to stress these possibilities to obtain the best performance out of CyMorph. Asymmetry only depends on d_σ . For the other metrics, we exhaustively explore the combinations of configurable parameters: outer (R_1) and inner (R_2) radii for Concentration; control parameter c of Butterworth Filter for Smoothness; modular (m_{tol}) and phase tolerance (p_{tol}) for G_2 ; and, number of bins β for Entropy. Table 4.1 summarizes parameters and ranges explored. The optimization process may be approached in different ways. One of them consists of optimizing all variables at once by maximizing a metric which is output from the application of a local two-sample test (KIM et al., 2019). In this work, we focus on a variable-by-variable optimization which not only enables to select the best configuration and input metrics to TML methods but also leaves to the same accuracy in morphology as that obtained in GZ1 (see Section 4.4). In the optimization experiments reported here, we randomly select a sample with 1,000 ellipticals and 1,000 spiral galaxies. Figure 4.2 presents the results for all optimization experiments. In each plot, all lines are dashed except the red one which

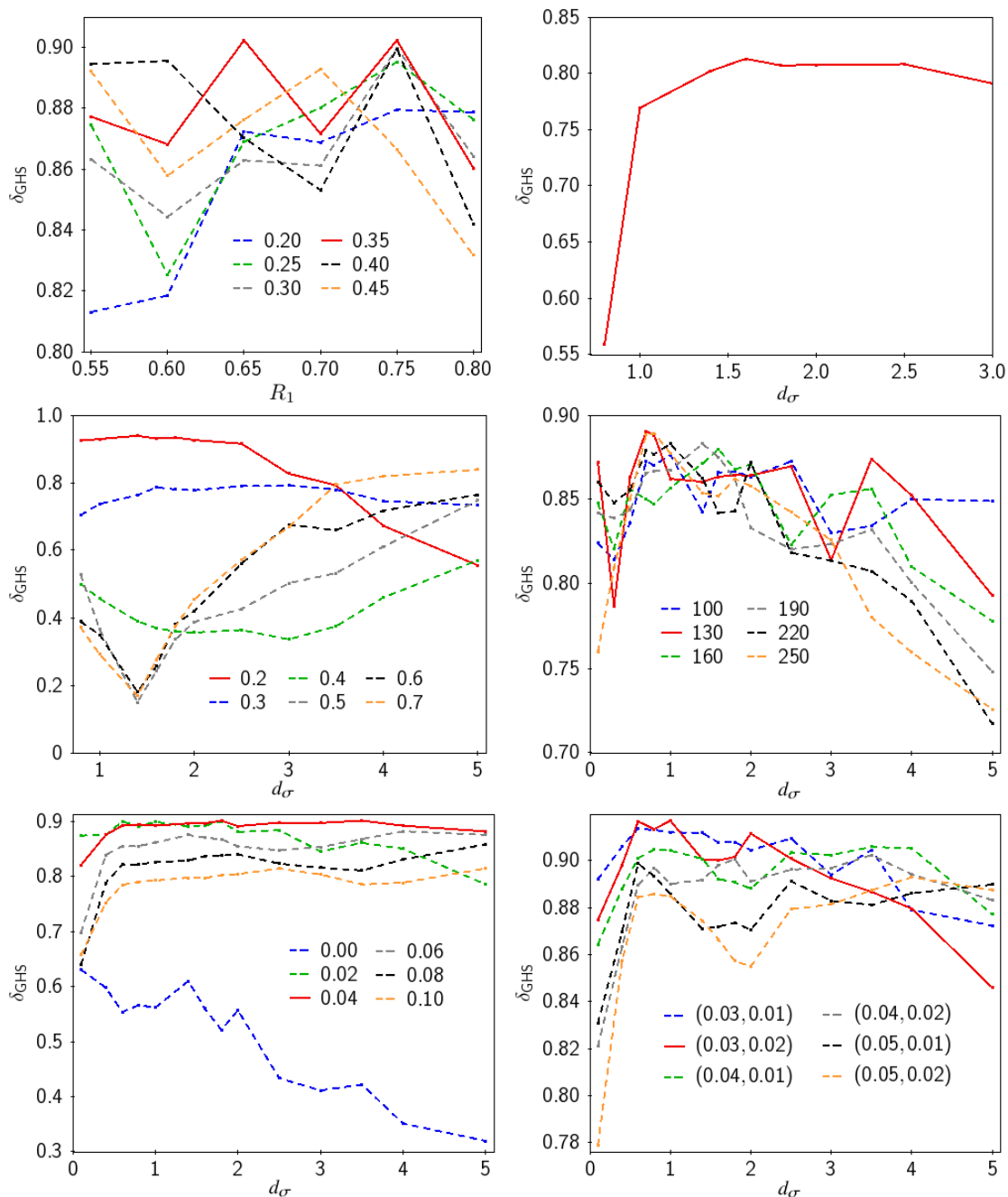
contains the best configuration for a given metric. The y -axis has GHS separation values (δ_{GHS}) in every panel. In the following Subsection, we interpret the results displayed in Figure 4.2.

Table 4.1 - Parameter ranges explored in the optimization process. Asymmetry is omitted since it depends only on d_σ . Concentration(*) does not depend on d_σ .

Sextractor	C*	S	G₂	H
$0.1 \leq d_\sigma \leq 5.0$	$0.55 \leq R_1 \leq 0.80$ $0.20 \leq R_2 \leq 0.45$	$0.2 \leq c \leq 0.8$	$0.00 \leq m_{tol} \leq 0.20$ $0.01 \leq p_{tol} \leq 0.04$	$100 \leq \beta \leq 250$

Source: Barchi et al. (2020).

Figure 4.2 - Optimization process for morphological metrics configuration.



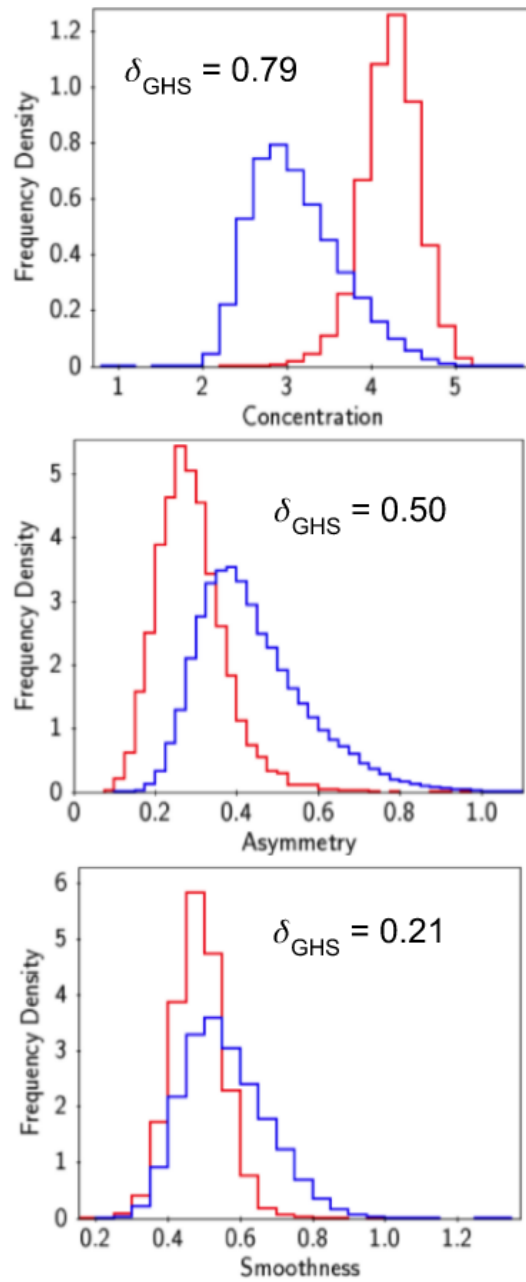
Respective metrics' plots from top left to bottom right: Concentration, Asymmetry, Smoothness, Entropy, and GPA (modular tolerance and fine tuning). Red continuous lines (not dashed) have the best configuration for the given parameter. See the explanation for the experiments and best configuration results obtained in this Subsection 4.2.2.

Source: Barchi et al. (2020)

4.2.3 Results on morphology

In this subsection, we compare the results obtained by computing the classic CAS system (CONSELICE, 2003; LOTZ et al., 2004), which are presented in Figure 4.3 and the optimal results obtained by CyMorph system, exhibited in (Figure 4.4). Conselice (2003), Lotz et al. (2004) estimate Concentration and Asymmetry without significant differences among them. However, Smoothness is implemented in different ways. We present Smoothness as in Lotz et al. (2004), which gives the most consistent results. For each non-parametric morphological index, we display a binomial distribution histogram with elliptical galaxies (in red) and spiral galaxies (in blue). In each panel we also list the δ_{GHS} value.

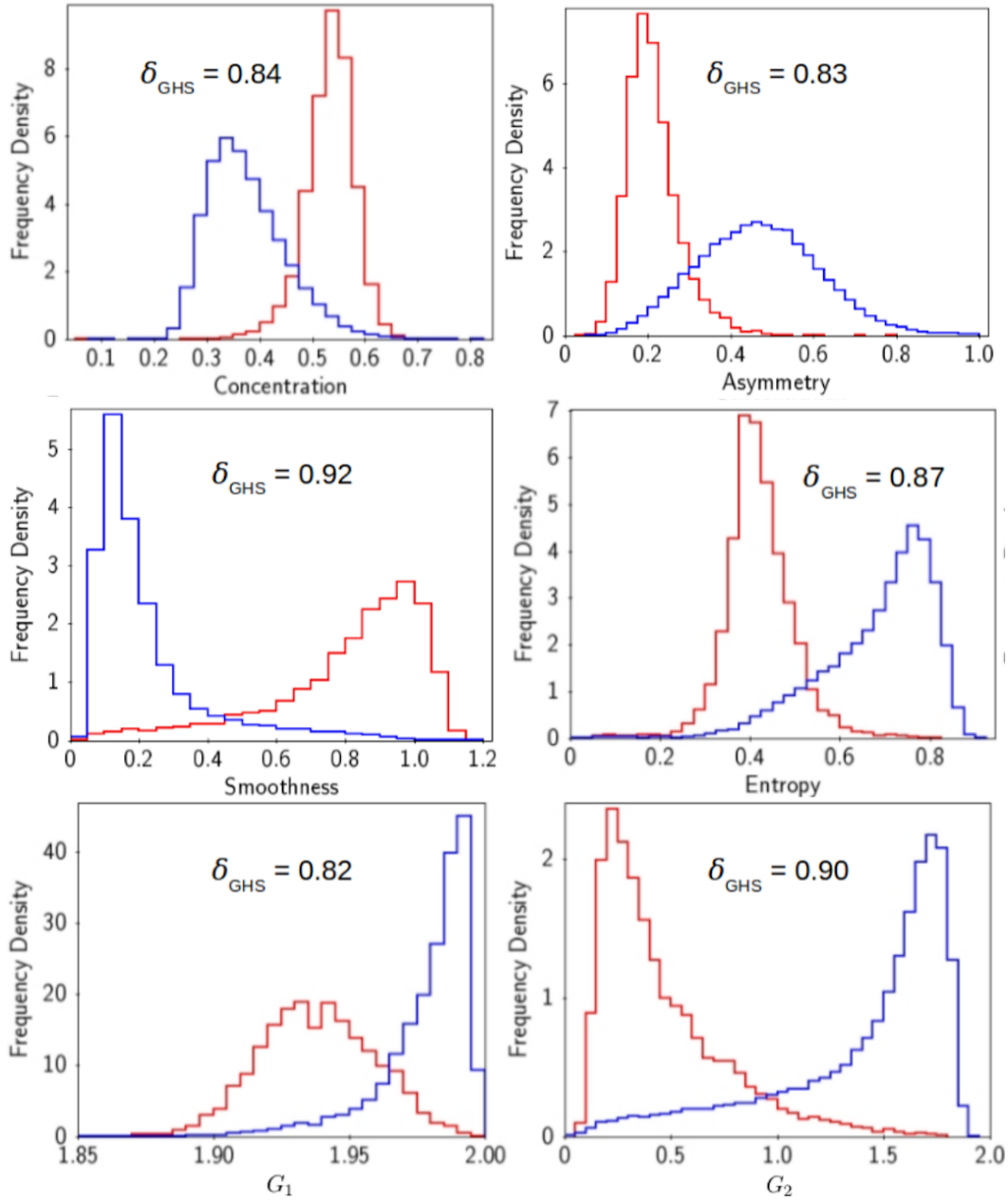
Figure 4.3 - Results on galaxy morphology using Classic CAS (CONSELICE, 2003; LOTZ et al., 2004).



From top to bottom: Concentration, Asymmetry, Smoothness. Elliptical galaxies in red and spiral galaxies in blue.

Source: Barchi et al. (2020).

Figure 4.4 - Results on galaxy morphology using CyMorph.



Elliptical galaxies in red and spiral galaxies in blue.

Source: Barchi et al. (2020).

The classic CAS system has the best result with Concentration ($\delta_{\text{GHS}} = 0.79$), however, this is still lower than the lowest performance obtained by CyMorph metrics, which is Asymmetry with $\delta_{\text{GHS}} = 0.83$ (see Figure 4.4). With this improvement in CAS metrics within CyMorph (Smoothness, for instance, the best result: $\delta_{\text{GHS}} = 0.92$), and the adoption of Entropy ($\delta_{\text{GHS}} = 0.87$) and Gradient Pattern

Analysis (G_2 : $\delta_{\text{GHS}} = 0.90$), we have satisfactory non-parametric morphology metrics to serve as input features to the Traditional Machine Learning algorithms. G_2 and H , two of the best metrics by δ_{GHS} , have highly correlated results: the greater the Entropy value, the more asymmetric gradient patterns, and vice versa, lower entropy values correspond to more symmetric gradient patterns.

The reasons for the improvement upon classic metrics can be summarized as: (1) the three-step preprocessing, (2) Butterworth filter to smooth the image (concerning Smoothness metric), (3) usage of correlation coefficients for Asymmetry and Smoothness, and (4) optimization process to better configure each metric.

4.3 Machine learning applied to galaxy morphology

CyMorph presents a consistent non-parametric morphological system. By employing Machine Learning (ML) methods with CyMorph metrics as features, we can value the best morphological information and obtain reliable and consistent classification results in galaxy morphology. An alternative would be to test logistic regression and other regression methods, which is beyond the scope of this paper. The five input features for the learning process are the best morphological metrics (given by δ_{GHS}) computed by CyMorph: C, A, S, G_2 and H. We maintain the restriction related to the area of the galaxies to build up different classification models: (1) $K \geq 5$; (2) $K \geq 10$; and (3) $K \geq 20$, i.e., the area of the galaxy is at least five (model 1), ten (model 2) and twenty (model 3) times larger than the FWHM area for each corresponding object, respectively.

We build up Decision Tree (DT), Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) models to classify galaxies considering different numbers of classes. We use `scikit-learn` (PEDREGOSA et al., 2011) python library to perform the experiments and procedures reported in this Section. We use Cross-Validation (CV) to split the dataset in training-validation-testing to address the trade-off between bias and variance (MITCHELL, 1997; GÉRON, 2019). First, we split the dataset in a 90/10 proportion for training and testing sets, respectively. CV is applied in the 90% portion of the dataset.

Consistent performance validation metrics are crucial to guide the learning process and to objectively measure the performance of each model. No metric is designed to perform this task alone. We employ the Overall Accuracy (OA) as the figure of merit to compare all the different models. Additionally, we employ other performance metrics: Precision (P) and Recall (R) — see Mitchell (1997), Géron (2019),

Goodfellow et al. (2016) for more details about OA, P, and R. For a further analysis on the problem with two classes, we use the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC curve — AUC (BRADLEY, 1997).

4.4 Results on classification and discussion

4.4.1 Classifier’s performance by overall accuracy (OA)

As we have shown in previous sections, there are several parameters driving the final classification and an appropriate figure of merit is needed to establish which setup/method works best. In Tables 4.2 and 4.3, we present the Overall Accuracy (OA) achieved by all the experiments carried-out in this work. The main goal here is to distinguish between an Early-Type Galaxy (ETG), elliptical (E), and a Late-Type Galaxy (LTG), spiral (S). In the case of TML, using the $K \geq 20$ sample, all methods reached over 98% of OA. In this training set, there are many more S galaxies ($\sim 87\%$) than E galaxies ($\sim 13\%$). This difference in the number of examples between classes is called class imbalance. We discuss class imbalance in Subsection 4.4.2. Despite of the imbalance, we have at least 95% precision and 96% recall for E systems. Since most of the training set is constituted by S galaxies, it is not surprising that we reach $\sim 99\%$ precision and recall, establishing a model with $\sim 99\%$ OA for this dataset.

Overall, CNN is the best approach to establish morphological classification of galaxies. We can safely assert that starting from the classes E and S from Galaxy Zoo 1, we can reproduce the human eye classification with all methods and samples (OA $> 94.5\%$). When trying to distinguish among 11 classes, the problem is much more complex, as it would be for the human eye, and the best result is OA $\sim 65.2\%$ using CNN with $K \geq 20$. However, if we use only three classes we find an OA $> 80\%$ with CNN, for all samples, namely elliptical (E), unbarred (S) and barred spiral (SB) galaxies.

Table 4.2 - Overall Accuracy (OA in percentage) for all approaches considering GZ1 classification (elliptical and spiral galaxies separation).

	$K \geq 5$				$K \geq 10$				$K \geq 20$			
	DT	SVM	MLP	CNN	DT	SVM	MLP	CNN	DT	SVM	MLP	CNN
two classes	94.8	94.6	94.6	98.7	95.7	95.8	95.6	99.1	98.5	98.6	98.6	99.5

Source: Barchi et al. (2020).

Table 4.3 - Overall Accuracy (OA in percentage) for all approaches considering GZ2 classification. The darker the green colour of a cell, the better OA obtained.

	$K \geq 5$				$K \geq 10$				$K \geq 20$			
	DT	SVM	MLP	CNN	DT	SVM	MLP	CNN	DT	SVM	MLP	CNN
11 classes	49.3	48.8	49.4	63.0	51.6	51.6	51.7	63.0	57.7	57.4	57.7	65.2
9 classes	60.9	63.2	63.0	70.2	60.5	63.8	63.6	75.7	63.5	66.4	66.2	67.4
7 classes	63.0	62.5	63.3	72.2	62.9	62.6	63.0	77.6	65.9	65.8	66.0	70.0
3 classes	71.9	71.2	71.2	80.8	71.9	74.6	74.9	81.8	78.7	78.5	78.8	82.7

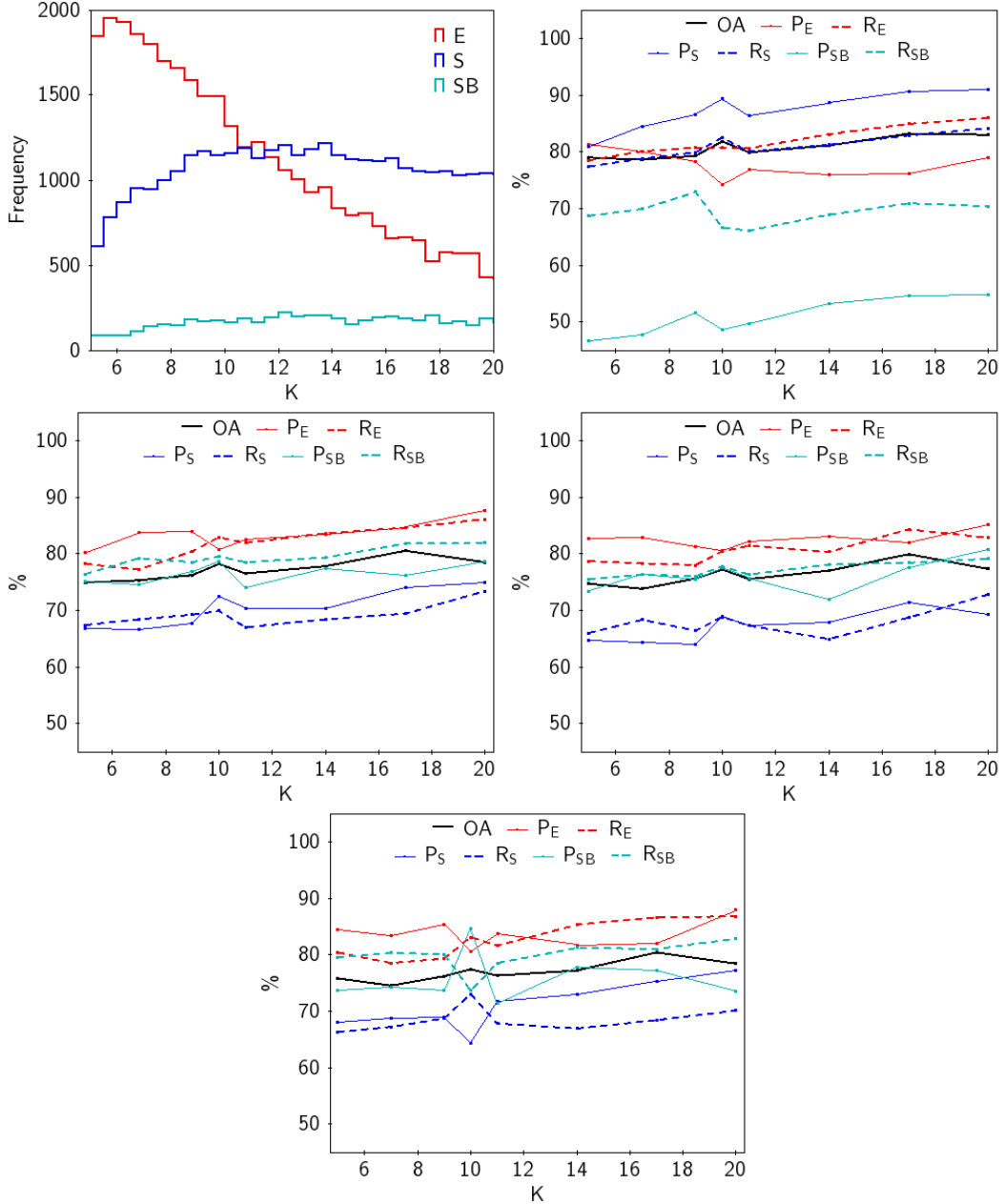
Source: Barchi et al. (2020).

4.4.2 Class imbalance in galaxy morphology

The class imbalance problem is one of the top in data mining, data science, and pattern recognition (YANG; WU, 2006). It arises when at least one of the classes has considerably fewer examples than the other(s). This problem is inherent in galaxy morphology, as the number of examples among classes will never be equal. Applying the restriction of $K \geq 20$ in the dataset from Galaxy Zoo 1 (LINTOTT et al., 2008; LINTOTT et al., 2011), for example, there are $\sim 87\%$ of galaxies classified as spiral and only $\sim 13\%$ as elliptical. Balancing the dataset generally improves the performance for minority classes (since we increase the number of examples of such classes for training), and thus increases precision and recall for these classes (GOODFELLOW et al., 2016). The first plot of Figure 4.5 shows the number of examples from three classes (E, S, SB) in Galaxy Zoo 2 - SDSS DR7 in different bins of K . The bin size is 0.5 and K varies from 5 to 20. SB is the minority class with the number of galaxies approximately constant — the bar component is a feature identified in all resolutions explored. The number of S galaxies increases until $K = 10$, approximately where the numbers of S and E galaxies are equal.

We investigate the impact of the imbalance class problem on the morphological classification testing four different datasets: imbalanced, undersampling, oversampling and Synthetic Minority Over-sampling Technique (SMOTE). The imbalanced dataset is the original query. In the undersampling dataset all classes have the same number of examples as SB originally have. For oversampling, we sample the minority class set with replacement. Using SMOTE, we synthetically generate more examples for SB and we consider the smaller of either the number of E galaxies or double the number of SB galaxies to be the number of examples for each class (PEDREGOSA et al., 2011).

Figure 4.5 - Panel about class imbalance in galaxy morphology.



The first plot shows number of elliptical (E), unbarred spiral (S) and barred spiral (SB) galaxies from GZ2 classification varying K. The other four plots are related to the class imbalance problem considering three classes: E (in redder colours), S and SB — in bluer colours. The black lines indicate the Overall Accuracy (OA). For each of the three classes, continuous lines represent Precision (P) and the dashed lines indicate Recall (R), considering the original imbalanced dataset (panel b), the dataset generated with SMOTE (panel c), the undersampled balanced dataset (panel d) and the oversampled balanced dataset (panel e). Plots from top left to bottom right: (a) Number of examples as a function of K, for different classes; (b) Imbalanced (original dataset); (c) Balanced — SMOTE; (d) Balanced — undersampling; (e) Balanced — oversampling.

Source: Barchi et al. (2020).

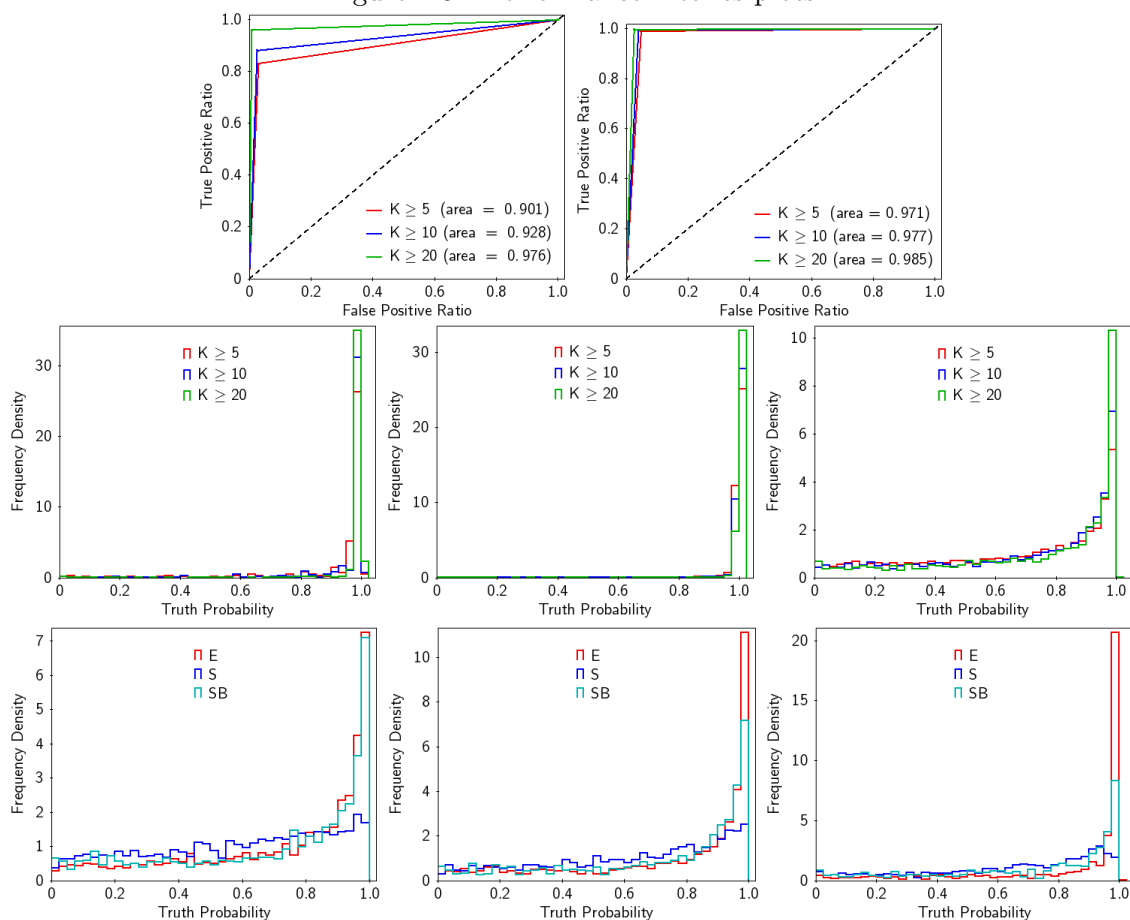
From the second to the last plots of Figure 4.5, we exhibit OA, P, and R for all experiments exploring the imbalance class problem considering the three classes described above. The minority class (SB) is the one more affected by imbalanced datasets, with low P (51%) and R (69%), in average. By employing balancing techniques, we improve to P (76%) and R (79%) for the minority class, thus reducing the misclassification for the SB class. All balancing strategies have similar performances. In all strategies, there is a $\Delta\text{OA} \sim 2\%$ when K varies from 5 to 20. From panels (b) to (e) of Figure 4.5, we notice that OA weakly increases with K , a trend that would imply that restricting the sample to bigger objects reduces classification problems, but the impact is not very significant. Thus, our model built up with the sample restricted by $K \geq 5$ can safely be used to classify an unknown dataset as it classifies smaller objects with a similar OA compared to bigger objects (such as $K \geq 20$).

In the remaining of this paper we continue analysing three methods: Traditional Machine Learning (TML) and Deep Learning (DL) approaches using imbalanced dataset for discriminating between two classes; and DL using SMOTE dataset to classify into three classes. For the TML approach, we choose the Decision Tree (DT) algorithm as it is the simplest solution (compared to Support Vector Machine and Artificial Neural Network) and the results have $\Delta\text{OA} \sim 0$ among them (Tables 4.2 and tab:resultsGZ2) — Occam’s razor (BLUMER *et al.*, 1987). For three classes, we select the model trained with SMOTE dataset because it is in the middle ground between under and overbalancing techniques, and the results using different balanced datasets are equivalent (Figure 4.5). These are the selected classifiers to build up the catalog — see details about the final catalog in 4.6.

4.4.3 Classifier’s performance by ROC curve and AUC

One of the most important issues in machine learning is performance measurement. A very popular method is the ROC (Receiver Operating Characteristics) curve and the Area Under the ROC curve — AUC (BRADLEY, 1997). In our particular case, ROC is the probability curve and AUC represents a measure of separability. It indicates how a model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting E’s as E’s and S’s as S’s. Based on data presented in Tables 4.2 and 4.3, Figure 4.6 displays the ROC curves, the histograms with ground truth probabilities given by the models using different datasets as well. Also, deeper into the three classes problem, Figure 4.6 exhibit histograms with ground truth probabilities given by the models for each class.

Figure 4.6 - Performance metrics plots.



The first row presents ROC Curve and the Area Under the ROC Curve (AUC — area) for each approach and different dataset restrictions considering the two classes problem (panels a and b). Such plots consider the ground truth and predicted labels. The dotted black line represents a random guess. The second row shows histograms with ground truth probabilities given by the models for each class (panels c, d, e). The third row presents histograms with ground truth probabilities given by the models for each class in regard to the three classes problem (panels f, g, h). Plots from top left to bottom right: (a) ROC curves - TML - 2 classes; (b) ROC curves - DL - 2 classes; (c) Truth Probability - TML - two classes; (d) Truth Probability - DL - two classes; (e) Truth Probability - DL - three classes; (f) Truth Probability - DL - three classes - $K \geq 5$; (g) Truth Probability - DL - three classes - $K \geq 10$; (h) Truth Probability - DL - three classes - $K \geq 20$.

Source: Barchi et al. (2020).

ROC curves are typically used in binary classification to study the output of a classifier (BRADLEY, 1997; GÉRON, 2019). Figure 4.6 show ROC curves considering the ground truth and predicted labels (no probabilities). These ROC curves and area values confirm what Table 4.2 shows with OA: all models have high standards for

acting upon the two classes problem with $\text{AUC} > 0.90$; restricting to the Deep Learning (DL) approach we improve it to $\text{AUC} > 0.97$. By experimenting with different dataset restrictions and approaches we can draw some interesting conclusions. The restriction on the dataset has more impact on TML approach than using DL. The ROC curves are closer to each other on the second plot of Figure 4.6 ($\Delta\text{AUC} = 0.014$) when compared to the first one ($\Delta\text{AUC} = 0.075$). One example is to compare TML using $K \geq 20$ and DL using $K \geq 10$ ($\Delta\text{OA} \sim 0.5\%$ and $\Delta\text{AUC} \sim 0$ among them). Using smaller objects, DL can achieve a very similar performance as TML using bigger objects.

The output probabilities given by these models with regard to the ground truth from Galaxy Zoo are explored in the two bottom rows of Figure 4.6). These histograms do not distinguish each class. We consider the output probability from each model for the ground truth of each galaxy. Again, we confirm that: (1) both approaches have a very high performance considering two classes — very high concentration of frequency density for truth probability > 0.9 , and (2) DL (fourth plot of Figure 4.6) improves the results when comparing to TML (third plot) by reducing the frequency density with low truth probability values. The impact of the dataset restriction continues as well: the higher we set the threshold for K , the denser the frequency for truth probability > 0.9 .

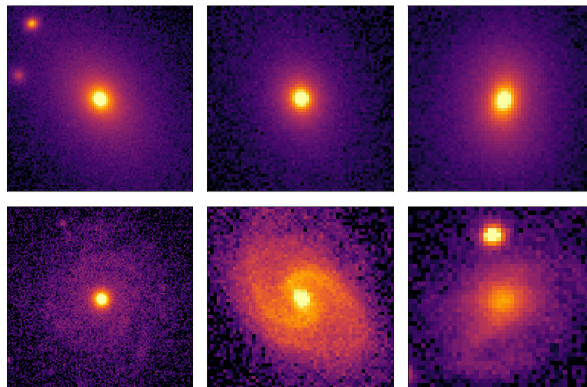
Although the fifth plot of Figure 4.6 also presents a high frequency density for truth probability > 0.9 , it is natural to see a higher frequency density for lower truth probabilities when comparing to the first and second plots, since this problem is more complex when having one more class to consider. Exploring further, the bottom row of Figure 4.6 shows different dataset restrictions employed to train models on the three classes problem. Once more, we can see clearly the impact of using bigger (but less) objects: the frequency density for lower truth probabilities decreases from left to right and gets even lower in the last plot. Non-barred spiral galaxies does not have a very high concentration for truth probability > 0.9 . However, the other two classes have a high frequency density for truth probability > 0.9 .

4.4.4 Learning about differences between TML and DL from misclassifications

The two approaches used in this work, Traditional Machine Learning — TML; and Deep Learning — DL, achieve almost the ideal performance considering the Overall Accuracy ($\text{OA} \sim 99\%$) for the two classes problem with the sample restricted by $K \geq 20$. However, it is still worthwhile to further investigate what causes misclas-

sification even at a low percentage. We remind the reader that misclassification is always established using Galaxy Zoo 1 as the ground truth. Figure 4.7 presents some examples of misclassification using TML. In the first and the last image we see that the preprocessing phase was not able to properly clean the image or discard such examples as bright objects remain close to the central galaxy. The other cases reflect the variance in the parameters used by Decision Tree and the natural uncertainty of the process. In Figure 4.8 we display some misclassified galaxies by DL. Here, the absence of preprocessing allows galaxies to be too close to the border (second and fourth images) and as before the other examples are simple misclassifications imposed by the method itself, namely the galaxies are easy to be misclassified — a bright central structure which gradually fades away to the outer part of the galaxy, which, in a more detailed classification (visual), could be considered as a S0 galaxy. We should stress that this misclassification is very low. Using a sample of 6,763 galaxies selected from the grand total of 58,030 galaxies listed in Table 1 ($K \geq 20$, from GZ1), not used in the training process, TML misclassifies only 72 galaxies (1%) while DL gets 0.5% misclassified galaxies. Also, we noticed that none of the galaxies misclassified by TML are in the list of misclassifications by DL. These results seem innocuous, however they remind us of how important is to treat objects close to the border and those near a very bright source as an specific set since this will always be present in any sample. It also reinforces how important it is that we treat independent methodologies along the process of establishing a final morphology attached to an object. The examples presented here show how visual inspection is still an important source of learning about morphology (the problem is not the eye but the quality of the image placed in front of you), although inefficient for large catalogs currently available and the ones coming up in the near future.

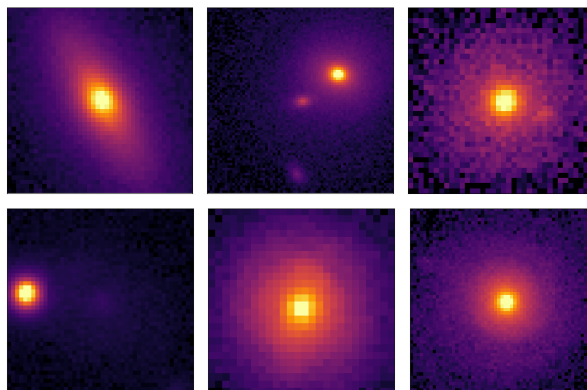
Figure 4.7 - Sample of misclassified galaxies among traditional machine learning approach and GZ1.



Sample of misclassified galaxies comparing the classification of Galaxy Zoo 1 (GZ1) and our Traditional Machine Learning (TML) approach trained with the sample restricted by $K \geq 20$. Under each galaxy image, we present the object ID number from SDSS-DR7 and the classification given by GZ1 and TML (0: Elliptical; 1: Spiral). Top row: galaxies classified as elliptical by GZ1 and as spiral by our TML approach. Bottom row: galaxies classified as spiral by GZ1 and as elliptical by our TML approach.

Source: Barchi et al. (2020).

Figure 4.8 - Sample of misclassified galaxies among our deep learning approach and GZ2.



Sample of misclassified galaxies comparing the classification of Galaxy Zoo 1 (GZ1) and our Deep Learning (DL) approach trained with the sample restricted by $K \geq 20$. Under each galaxy image, we present the object ID number from SDSS-DR7 and the classification given by GZ1 and DL (0: Elliptical; 1: Spiral). Top row: galaxies classified as elliptical by GZ1 and as spiral by our DL approach. Bottom row: galaxies classified as spiral by GZ1 and as elliptical by our DL approach.

Source: Barchi et al. (2020).

4.4.5 Validating classification with spectroscopic data

The performance analysis presented in the previous section reflects our ability to establish a morphological classification using a given method among several that might in principle work properly, and that's why finding a robust figure of merit is of paramount importance. However, an independent validation is even more essential when presenting a catalog with reliable morphology, namely, we have to show that our new classes recover well known relations. Figure 4.9 presents histograms of Age, stellar mass (M_{stellar}), Metallicity ($[Z/H]$) and central velocity dispersion (σ) — for more details on how these parameters were obtained and errors, see DE CARVALHO et al. (2017). In every panel we show the distribution for ellipticals (in red) and spirals (in blue). Also, we display the parameter δ_{GHS} which measures how distant these two distributions are (see Section 3.8). This validation procedure was done using only galaxies from GZ1 classified as "Undefined". The classification here is provided by DT (TML). We remind an important characteristic of the samples: $K \geq 5$, which has more but smaller objects; $K \geq 10$; and $K \geq 20$, which has less but bigger objects. Although we have a bigger dataset with $K \geq 5$, the presence of smaller objects impairs our classification. The degradation of the quality of our classification as we go to smaller galaxies is evident from Figure 4.9 where δ_{GHS} decreases for smaller K for all quantities except for Age where only a small fluctuation is noticed.

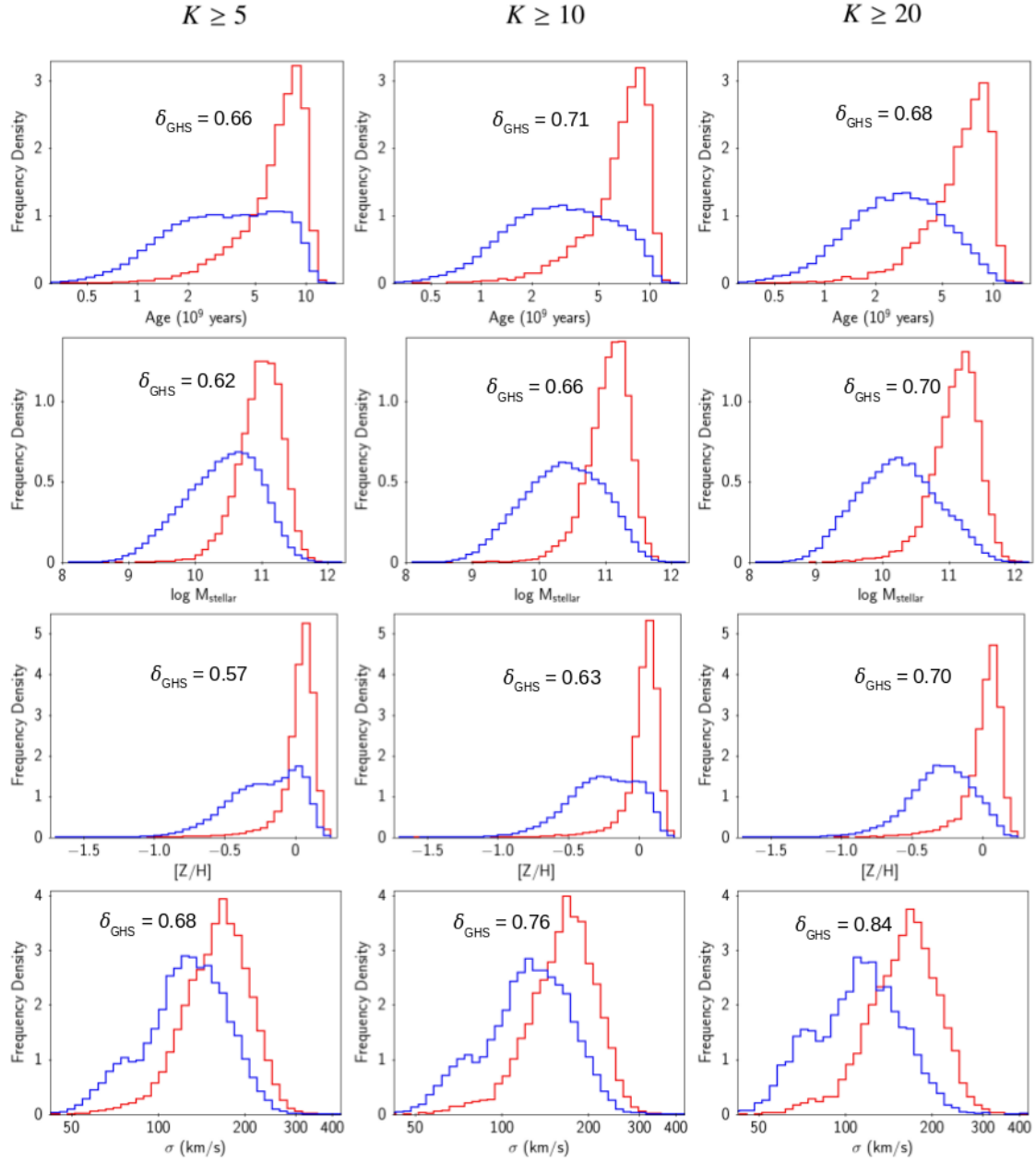
The number of galaxies for each histogram from Figure 4.9 is as follows:

- a) $K \geq 5$: 13,373 ellipticals; 87,095 spirals; Total: 100,468.
- b) $K \geq 10$: 9,030 ellipticals; 59,096 spirals; Total: 68,126.
- c) $K \geq 20$: 6,390 ellipticals; 24,988 spirals; Total: 31,378.

Figure 4.9 shows how our classification recovers well known properties of galaxies like in the first row we see that ellipticals have ages peaked around 9 Gyr while spirals are younger and the distribution is more spread, probably due to the contamination by S0 galaxies. The second row exhibits the stellar mass distribution and again ellipticals have larger M_{stellar} compared to spirals with a difference of ~ 0.9 dex, peak to peak. In the third row it is also evident the difference in metallicity (~ 0.4 dex), ellipticals are more metal rich than spirals, specially for larger systems. Finally, the distributions of central velocity dispersion show larger values for ellipticals and for spirals we even see a bimodality which reflects the disk to bulge ratio in this

morphological type. These distributions attest credibility to our final classification using DT (TML).

Figure 4.9 - Spectroscopic validation.



Spectroscopic validation for the “Undefined” galaxies from Galaxy Zoo 1 which here are classified by our Machine Learning approach using Decision Tree. Elliptical galaxies are displayed in red and spirals in blue. In each panel we give the geometric histogram separation, δ_{GHS} .

Source: Barchi et al. (2020).

4.4.6 Case study: star formation acceleration and morphologies

In this section we describe an application of the method presented here to study the relation between morphologies and galaxy evolution. More specifically, we use the method to classify a sample of galaxies between disks and spirals and measure quenching timescales for each group separately.

It has been established that galaxies show a bimodal distribution in colors, with two distinct peaks with young (blue) and old (red) stellar populations — e.g., Baldry et al. (2004), Wyder et al. (2007) — and a minimum in the distribution commonly known as the green valley. Although it is generally accepted that galaxies move from blue to red, the physical processes associated with this transition are not completely understood, i.e., we do not know which phenomena are responsible for accelerating the decline in star formation rates, whether a single one or a combination of effects.

Using galaxy colours and stellar population synthesis models, Schawinski et al. (2014) has shown that galaxy quenching can be divided into two distinct processes depending on morphology: elliptical galaxies quench faster, probably through merger activity. Nogueira-Cavalcante et al. (2018) have reached a similar conclusion with more precise measurements from spectral indices (the 4000Å break in the spectral continuum and the equivalent width of the H_δ absorption line). Nevertheless, both works rely on assuming specific exponentially declining star-formation histories.

To circumvent this limitation, Martin et al. (2017) have developed a method using the same spectral indices but with no restraints regarding a parametric star formation history. The authors have shown, by comparisons with results from cosmological simulations, that one could infer the instantaneous time derivative of the star-formation rates, denominated the *star-formation acceleration*. Formally, this is defined as

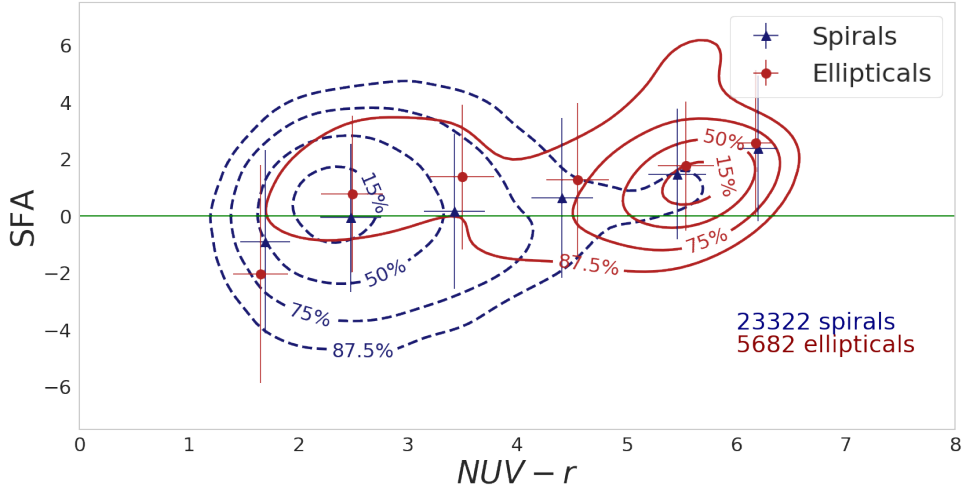
$$SFA \equiv \frac{d}{dt}(\text{NUV} - r), \quad (4.1)$$

with higher values representing stronger quenching.

In Sá-Freitas et al. (in prep) we apply this methodology to a sample of galaxies out to $z = 0.120$, divided by morphology. We only consider galaxies brighter than $M_r = -20.4$ for the sake of sample completeness. When compared to previous works, we are able to measure SFA in galaxies according to morphology for *all* objects, regardless of colour and assumed quenching histories. In that sense, the learning techniques presented here are fundamental to our analysis: by classifying a much larger number of galaxies (almost 30,000 galaxies in total), we are able to bin

our sample by colours and draw conclusions based on smaller subsamples of objects according to their morphologies.

Figure 4.10 - Star formation acceleration (SFA) as a function of NUV-r colours.



Higher SFA values represent faster quenching, while more negative values indicate strong bursts of star formation, with the green line showing no current variation in SFR. Data are binned in colour, with blue triangles for spiral galaxies and red circles for ellipticals. Error bars show the standard deviation within each bin. Contours show the number of galaxies in the diagram as percentage of the total count for each morphological type. Red galaxies are statistically indistinguishable, while ellipticals in the green valley are quenching significantly faster than spirals. At the blue end, the difference is not large enough for this sample to draw any conclusions.

Source: Barchi et al. (2020).

In Figure 4.10 we show our results: as expected, the bluest galaxies are currently undergoing strong bursts, while red galaxies are typically quenching. More importantly, we detect a significant distinction between SFA values for spirals and ellipticals in the green valley. Elliptical galaxies are quenching more strongly, while spirals appear to be moving gradually towards redder colours. We perform Kolmogorov-Smirnov and Anderson-Darling tests to test for the null hypothesis that the distributions for spirals and ellipticals are drawn from the same parent sample in each bin, ruling this out ($p < 0.05$) only for $2 \lesssim (NUV - r) \lesssim 5$. We therefore conclude that this is an effect distinguishable primarily within the green valley, which means that the star

formation histories of spirals and ellipticals are only significantly different during their transition to the red sequence.

In the near future, we expect the large upcoming imaging and spectroscopic surveys such as Euclid and DESI to increase our samples significantly, and deep learning techniques will yield reliable morphological classification of millions of objects. This will in turn allow us to further divide our galaxy sample, correlating morphologies with other phenomena such as AGN activity and environment in order to narrow our studies to the specific impact of each on the star formation histories of spiral and elliptical galaxies.

4.5 Comparison to other available catalogs

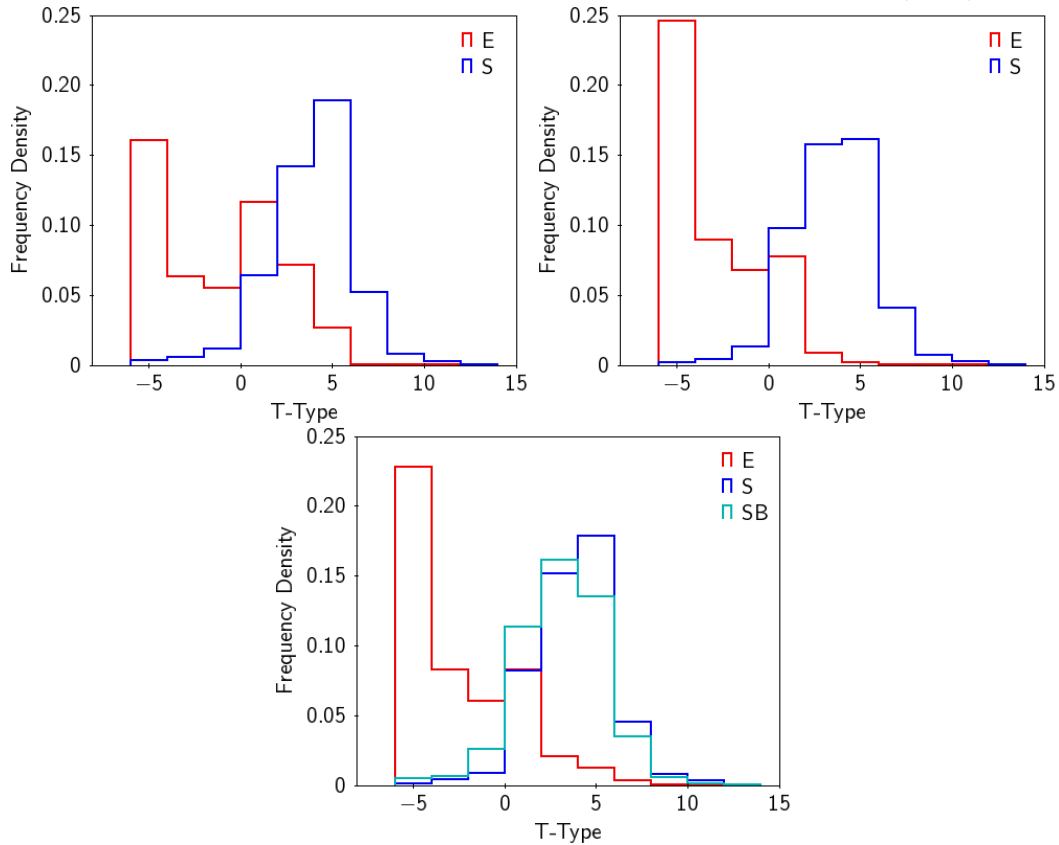
To attest the reliability of the morphological classification we provide in this work (see 4.6 for details about our catalog), it is of paramount importance to do external comparisons. There are currently two reliable catalogs that serve this purpose. First, [Nair and Abraham \(2010\)](#) provide T-Type information for 14,034 galaxies visually classified by an expert astronomer. Second, [Sánchez et al. \(2018\)](#) lists 670,722 galaxies also with T-Type available.

Figure 4.11 presents the histogram of T-Type provided by [Nair and Abraham \(2010\)](#) for the elliptical and spirals classes from our work. In general, the distributions are as expected - ellipticals peak around T-Type = -5 and spirals around T-Type = 5. In all three cases we notice an extension of the histogram for ellipticals towards larger T-Types, like a secondary peak around T-Type = 1, which may be associated to S0 galaxies. In the second plot (top right) of Figure 4.11, we note an improvement in using DL over TML by observing the decrease of the fraction of elliptical galaxies and the corresponding increase of spiral galaxies with T-Type > 0. Such behavior is also there in the last plot (bottom center) of Figure 4.11, considering three classes, with elliptical galaxies mostly with T-Type ≤ 0 and spirals (S and SB) primarily with T-Type > 0. The general comparison to the classification obtained by [Nair and Abraham \(2010\)](#) exhibit a 87% agreement.

Figure 4.12, analogous to Figure 4.11, shows how our classification performs in comparison with that provided by [Sánchez et al. \(2018\)](#). In all panels, we see a striking difference *wrt* the comparison with [Nair and Abraham \(2010\)](#) - a considerable amount of spirals around T-Type ~ -2 . Along with T-Type, [Sánchez et al. \(2018\)](#) provide also the probability of each galaxy being S0: P_{S0} . They define elliptical galaxies as those with T-Type ≤ 0 and $P_{S0} < 0.5$; S0 galaxies have T-Type ≤ 0

and $P_{S_0} > 0.5$; and spiral galaxies have T-Type > 0 . We plot the elliptical galaxies, following their definition, as a filled histogram in orange, and this shows a higher peak at T-Type ~ -2 wrt the distribution of the ellipticals with no T-Type restriction. Therefore, restricting the definition we get much higher concordance, namely higher fraction of systems that we classify as ellipticals, which translates into a higher peak around T-Type ~ -2 . Not only this, but as we can see from panel (c), using the three classes morphologies, the fraction of ellipticals with no T-Type restriction gets lower and the barred spirals appear more prominently around T-Type ~ 4 . In the same way we did when comparing to Nair and Abraham (2010), here we find a 77% agreement when comparing only elliptical and spiral galaxies.

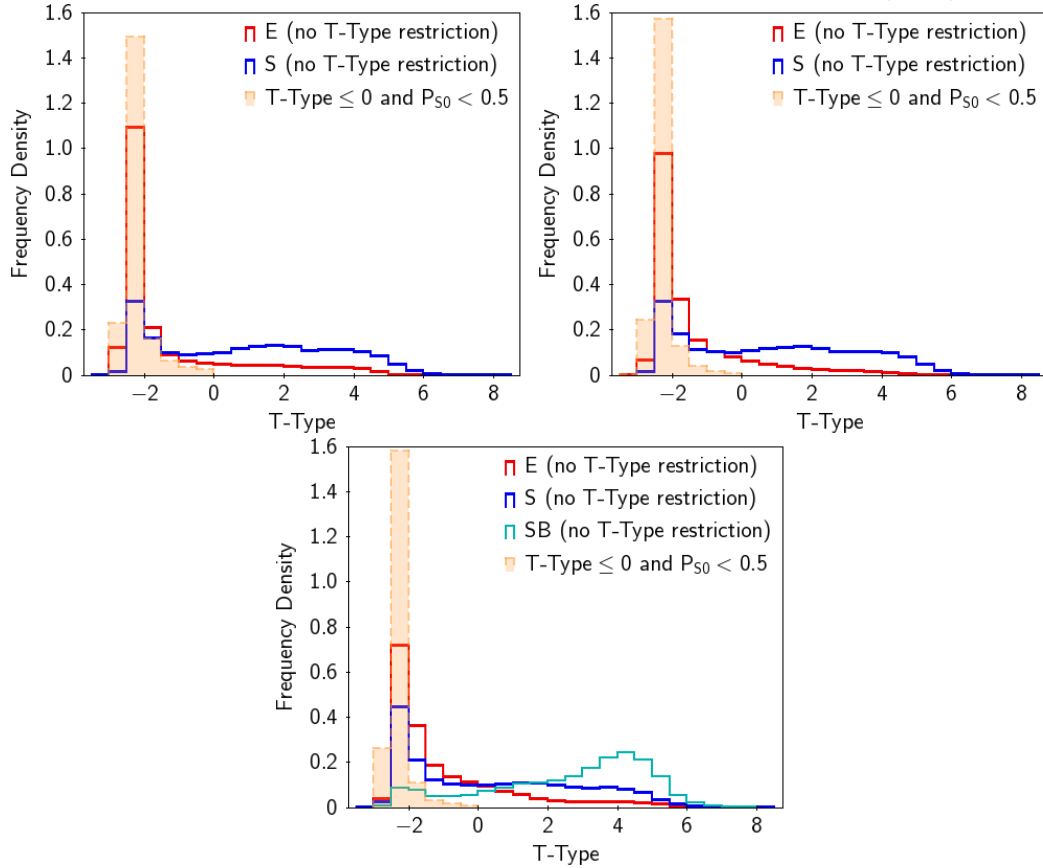
Figure 4.11 - Histograms presenting classifications for Nair and Abraham (2010)'s sample.



Histograms (normalized by area) presenting classifications for Nair and Abraham (2010)'s sample by T-Type using Traditional Machine Learning classification with two classes (panel a) and classification with two (panel b) and three classes (panel c) from deep CNN.

Source: Barchi et al. (2020).

Figure 4.12 - Histograms presenting classifications for Sánchez et al. (2018) sample.



Histograms presenting classifications for Sánchez et al. (2018) sample by T-Type using Traditional Machine Learning classification with two classes (panel a), and, classification with two (panel b) and three classes (panel c) from deep CNN (normalized by area).

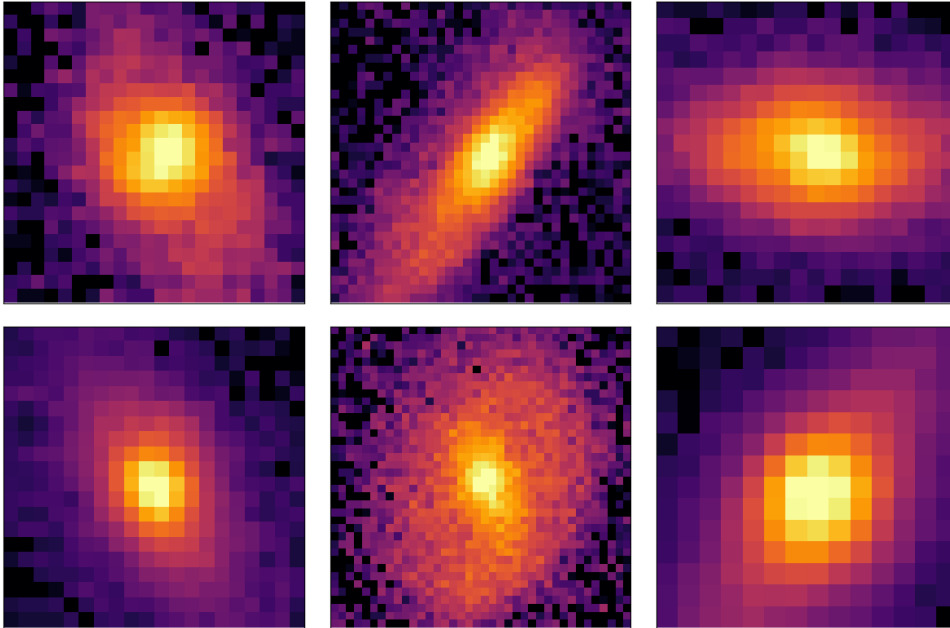
Source: Barchi et al. (2020).

A final note on the comparison with Sánchez et al. (2018) is related to the S0 class, in which we see a prominent bulge and a disk. They classify 230,217 as S0 and 27.96% of these systems (64,380) have $K < 5$, i.e., $\sim 28\%$ of galaxies classified as S0 are very small objects. Visually, it is easy to misclassify galaxies with predominant, oval and bright structure; and the task becomes even more difficult if the objects are small. Figure 4.13 shows a sample of galaxies with $-2.25 \leq T\text{-Type} \leq -2$ according to Sánchez et al. (2018) that we classify as spiral galaxies. As our classifiers do not discriminate the S0 morphology it is not surprising that we classify as spiral if the galaxy has a prominent disk.

Finally, we note that since Nair and Abraham (2010) and Sánchez et al. (2018) present their classification as T-Type, a proper comparison is difficult to make.

However, the agreement displayed in Figures 4.11 and 4.12, together with the global concordance when comparing elliptical and spiral galaxies, give us confidence that the classifications obtained here in this work are consistent and robust.

Figure 4.13 - Sample classified as spiral galaxies by our classifier with $-2.25 \leq \text{T-Type} \leq -2$ by Sánchez et al. (2018).



The object ID number from SDSS-DR7 is presented under each galaxy image.

Source: Barchi et al. (2020).

4.6 Final catalog (paper appendix)

The final product of this work is a catalog with morphological information for 670,560 galaxies (available at [this public link](#)²). The input data comes from SDSS-DR7 and the sample is restricted by the redshift range $0.03 < z < 0.1$, Petrosian magnitude in r-band brighter than 17.78, and $|b| \geq 30^\circ$. We provide morphological classification using TML and DL approaches for distinguishing elliptical (0) from spiral (1) galaxies. Furthermore, using DL approach, we release classification considering three classes: ellipticals (0), unbarred spirals (1) and barred spirals (2). For DL classification, we exhibit the classes ordered by probability and respective confidence percentages. We provide our best morphological non-parametric parame-

²Full link for the final catalog: <https://www.sciencedirect.com/science/article/pii/S2213133719300757?via%3Dihub#ec-research-data>.

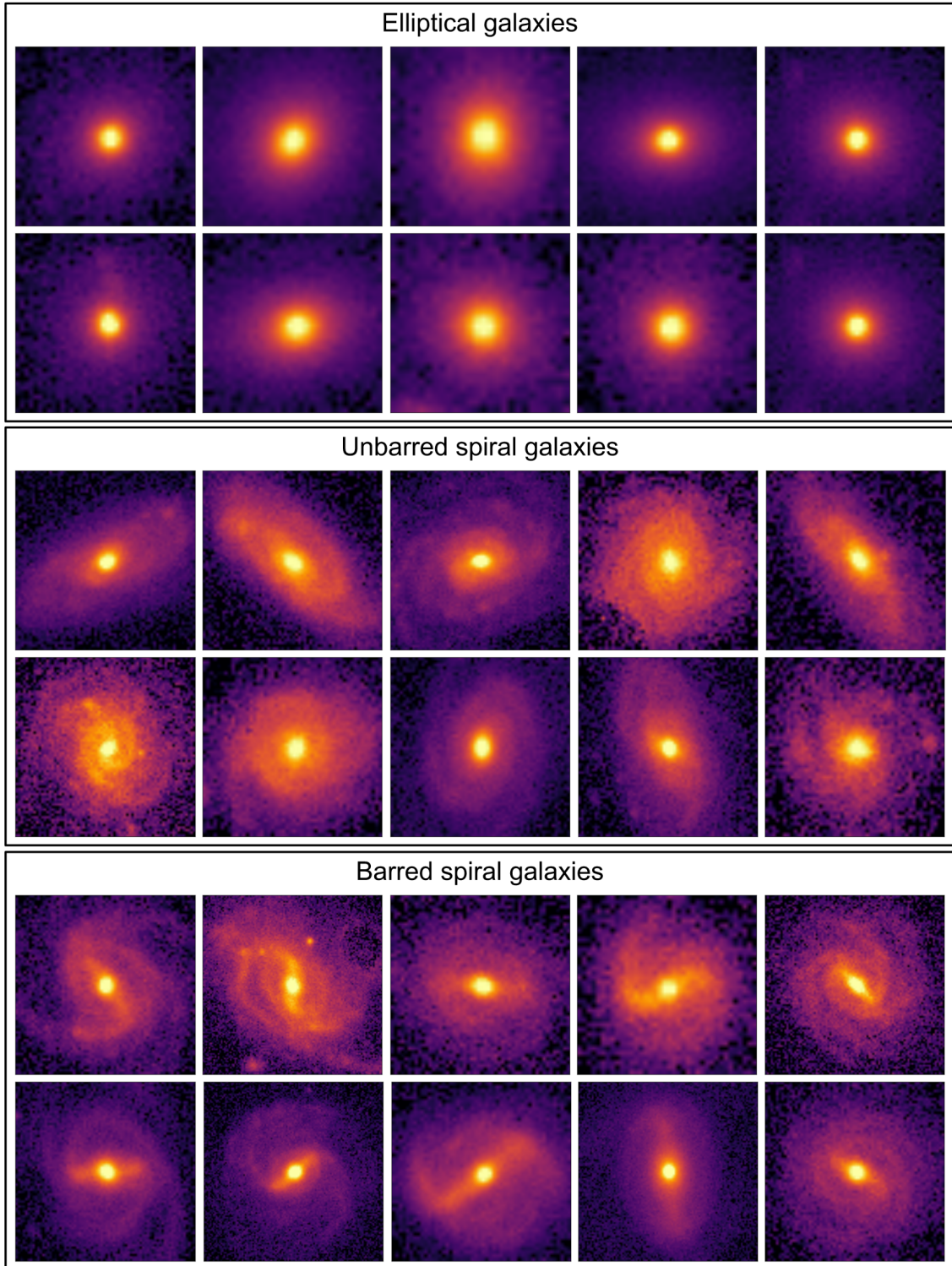
ters as well: Concentration (C), Asymmetry (A), Smoothness (S), Gradient Pattern Analysis parameter (G_2) and Entropy (H). The columns we provide are: the value of the parameter K , CyMorph metrics (5 columns), CyMorph Error, TML classification considering two classes, DL classes considering two classes and their respective percentages (4 columns), DL classes considering 3 classes, and their respective percentages (6 columns). In detail:

- a) K is the area of the galaxy’s Petrosian ellipse divided by the area of the Full Width at Half Maximum (FWHM).
- b) C , A , S , G_2 and H are the non-parametric morphological parameters from the CyMorph system (see Section 4.2);
- c) *Error* contains the Error flag after processing CyMorph;
- d) *ML2classes* is the classification obtained with the TML approach, using CyMorph and Decision Tree to separate galaxies into elliptical and spiral galaxies. Here, we maintain the restriction about K : galaxies with $5 \leq K < 10$ are classified by the model built up with the sample with $K \geq 5$ restriction; galaxies with $10 \leq K < 20$ are classified by the model built up with the sample with $K \geq 10$ restriction; galaxies with $K \geq 20$ are classified by the model built up with the sample with $K \geq 20$ restriction; galaxies with $K < 5$ are not classified.
- e) *CNN2classes1stClass* is the class with the highest probability considering the two classes problem. Analogously for *CNN2classes2ndClass*, and for three classes classification with *CNN3classes1stClass*, *CNN3classes2ndClass* and *CNN3classes3rdClass*.
- f) *CNN2classes1stClassPerc* is the probability percentage of the 1st class in the two classes problem. Analogously for *CNN2classes2ndClassPerc*, and for three classes classification with *CNN3classes1stClassPerc*, *CNN3classes2ndClassPerc* and *CNN3classes3rdClassPerc*.

Both classifications for two classes problems are performed by models trained with imbalanced datasets. For the 3 classes separation, we use the model trained with the SMOTE dataset. For classifications provided by CNN, we use models trained with $K \geq 5$ dataset. On the one hand, it is important to remember that TML does not classify galaxies with some problem detected by CyMorph. On the other hand,

DL acts directly upon images and has no error detection. This catalog represents a significant improvement for extragalactic studies related to galaxy morphologies. The Galaxy Zoo project (LINTOTT et al., 2008; LINTOTT et al., 2011; WILLETT et al., 2013) is a great success in offering large numbers of galaxies with reliable morphological classification. Nevertheless, GZ does not provide morphology information for a significant fraction of galaxies. Our catalog complement such effort. We show in Subsection 4.4.6 that such classification is especially relevant in sparsely inhabited areas of the colour-magnitude diagram, i.e. red spirals and blue ellipticals, for which the lack of objects renders the measurement imprecise. In Figure 4.14 we exhibit some typical examples of each class, where we can see the high quality of the classification for large objects ($K \geq 10$).

Figure 4.14 - Sample of galaxies classified by our deep learning approach.



A sample of galaxies classified here in this work using Deep Learning, Galaxy Zoo 2 as supervision (3 classes problem) from $K \geq 10$ dataset. Elliptical galaxies (two top rows), unbarred spirals (two middle rows), and barred spirals (two bottom rows).

Source: Barchi et al. (2020).

5 GENERAL DISCUSSION

The discussion for machine and deep learning applied to galaxy morphology is extensively performed in Subsection 4.4 and summarized in Subsection 6.1. The following Subsection focuses on computational aspects of this specific work.

5.1 Computational aspects

From a practical perspective, this section provides relevant information on hardware, software and processing time for future studies with similar approach. We address technical details on CyMorph, TML, DL and data.

In Section 4.2, we present CyMorph, a non-parametric galaxy morphological system written in `Cython`. `Cython` is an extension from `python` which allows explicit type declarations and its code is directly compiled to `C`. `Cython` has the high-level aspect from `python` and treats the large overhead for numerical loops from `python` by interacting natively with `C` (BEHNEL et al., 2011).

CyMorph has a configuration file to specify all details before run. In this file, it is possible to edit the path to the images of fields of view of the sky and stamps; whether to delete image files after processing or not; whether the image to be processed is already cut in the form of stamp or not; whether all metrics except concentration are supposed to act upon the segmented image or not; dimensions of the stamps in function of the Petrosian Radius (R_p); whether to save or not the images generated in intermediate steps (clean, smoothed and rotated images, for example); which metrics to process; and the desired configuration for each metric – the optimal configuration described in this work is set to be the default. If the field of view of a desired galaxy is not saved in disk, CyMorph also downloads the field of view from Sloan Digital Sky Survey Data Release 7 (SDSS-DR7) database before preprocessing.

When working with galaxy morphology on astrophysical images, we deal with fields of view of the sky and galaxy stamps. A typical field of view from SDSS DR7 has 5.9MB. One average galaxy stamp has approximately 21KB. Considering a common CyMorph run which saves the original stamp (raw cut from the field of view), clean stamp (after cleaning other objects) and a segmented stamp, we need $\sim 63\text{KB}$ of storage for each galaxy. For the final catalog of this work, for example, we have $670.729 \times 3 \times 21\text{KB} \sim 39.7\text{GB}$ just for the stamps.

There is a branch in Computer Science for studying the performance of algorithms by mathematical analysis called Analysis of Algorithms. The most common notation

for measuring the execution of an algorithm is the Big-O (\mathcal{O}). Any function whose magnitude is upper-bounded by $c \times f(n)$ (where c is a constant and n is the input size) for all sufficiently large n is $\mathcal{O}(f(n))$ (KNUTH; GREENE, 1999). Big-O is a convenient way to express the worst-case scenario for a given algorithm.

All tasks with high computational cost of CyMorph are $\mathcal{O}(n^2)$ because they operate upon an image. Preprocessing, each step to calculate the concentration, asymmetry, smoothness, entropy and Gradient Pattern Analysis (G2) are $\mathcal{O}(n^2)$. We do not discuss the Big-O for SExtractor, Machine and Deep Learning algorithms in details since we are using well-established packages and frameworks for such methodologies. For all approaches presented in this work, we discuss processing time and computer architecture required following in this section.

Since the goal is to compute the same metrics upon all galaxies from the sample, the correct approach is to parallelize the execution with regard to data. CyMorph uses `Open MPI` (GABRIEL et al., 2004) and `mpi4py` (DALCÍN et al., 2005) to perform single program, multiple data (SPMD) parallelism. The list of galaxies is distributed equally among all processes. Every process has its own output file. When all galaxies are processed, the main process concatenates the output files from all the children processes to produce the final result.

We describe here the hardware used to process CyMorph on different samples. We use DELL Precision Tower with 64 GB of RAM and 8 CPU cores for smaller samples (< 200.000 objects). For a sample with ~ 100.000 objects, CyMorph takes ~ 1 day, 7 hours and 20 minutes to process in this machine with 16 parallel processes. We use Xeon Phi Knights Landing with 68 available CPU cores and 512 GB of RAM for medium samples (~ 240.000 objects). With 40 parallel processes, CyMorph takes ~ 3 days and 6 hours to process this sample in this machine. For the biggest sample we analyse in this work, we use the Helios Cluster (located at the building of Atmospheric and Space Sciences, in National Institute for Space Research — INPE) which have 40 CPU units and 64 GB of RAM. CyMorph takes ~ 9 days and 1 hour to process this sample with 20 parallel processes in this machine.

We perform the TML experiments with `scikit-learn` (PEDREGOSA et al., 2011) `python` library using the DELL Precision Tower machine described above. Generally, it takes seconds to build up the classification model and classify galaxies. Considering the ~ 240.000 objects sample, it takes ~ 3 seconds to build up the model and ~ 2 seconds to classify.

For DL experiments, we use `NVIDIA DIGITS 6.1.1` (YEAGER et al., 2015) and `Caffe 0.17` (JIA et al., 2014) as frameworks to build up the dataset, train the networks and classify. Our input images have 256x256 (applying stretch, if needed). DL run at an NVIDIA DGX-1 machine with 40 CPU units, 512 GB of RAM and 8 Graphical Processing Units (GPU) TESLA P100 with 16 GB of memory (located at Federal Fluminense University, Brazil). Each training process uses 2 TESLA P100 GPU. For the sample with ~ 240.000 , it takes ~ 2 hours and 44 minutes to build up the classification model and ~ 24 minutes to classify.

6 CONCLUSION

6.1 Summary

With the new photometric surveys coming up, in several bands and with varying depth, it is of paramount importance to have the proper machinery for morphological classification, which is one of the first elements to create a reliable galaxy catalog, from which we can select clusters of galaxies and study the large scale structure of the universe. Here, we present models and methodologies to achieve these goals. We investigate the limits of applicability of TML and DL, in the supervised mode and compare their performances. We revisit the non-parametric methodology using C, A, S, H and G_2 and study some details ignored in previous works. Also, we examine how different methods are sensitive to the size of the galaxies, here identified by the ratio between the object's area and the PSF area. Finally, we remind the readers again of the importance of comparing TML with DL, since they are radically different approaches that in principle should result in similar classes. In the following, we summarize the main contributions of this paper:

- a) We investigated how parameters involved in the TML (S,A,H, and G_2) depend on the threshold used to obtain the segmented image. Although this seems a minor detail, it has proven to be an important ingredient in improving the TML performance, since the separation between ellipticals and spirals, δ_{GHS} , is maximized according to the threshold. Comparison with the traditional CAS system shows a considerable improvement in distinguishing ellipticals from spirals, namely, for CAS ($\delta_{GHS} = 0.79, 0.50, 0.21$ for C, A and S respectively), while in our modified CAS we have $\delta_{GHS} = 0.84, 0.83, 0.92$. Besides, the new parameters H and G_2 have their δ_{GHS} values very high (0.87 and 0.90, respectively) attesting their usefulness in galaxy morphology analysis. We list all these parameters in our main catalog with 670,560 galaxies.
- b) One way of testing the quality of our morphological classification (based on photometric data) is to compare with independent classes established with different data (spectroscopy). We use only galaxies from Galaxy Zoo 1 classified as 'Undefined' and applied our Decision Tree with the traditional machine learning approach. The result is presented in Figure 4.9 where we see an excellent performance of our classes in distinguishing the stellar population properties of ellipticals and spirals. Ellipticals have ages peaked around 9 Gyr while spirals are younger and the distribution is more

spread. Ellipticals have larger $M_{stellar}$ compared to spirals with a difference of ~ 0.9 dex. The difference in metallicity (~ 0.4 dex) between ellipticals and spirals is noticeable, specially for larger systems. Also, ellipticals show larger values of central velocity dispersion compared to spirals, for which we even see a bimodality reflecting the disk to bulge ratio variation in this morphological type.

- c) We present a preliminary result on SFA which study was always hampered by the lack of reliable morphological classification for a sizeable sample. Our catalog provides the necessary input data for such analysis. We show that the bluest galaxies are currently experiencing strong bursts while red galaxies are quenching. Also, we present for the first time a significant distinction between SFA values for spirals and ellipticals in the green valley. We find that the star formation histories of spirals and ellipticals are only significantly different during their transition to the red sequence. A full analysis of this topic and consequences for galaxy evolution is presented in Sá-Freitas et al. (in prep).
- d) We use a deep convolutional neural network (CNN) - GoogLeNet Inception, to obtain morphological classifications for galaxies for all galaxies in the main catalog under study here. With the twenty-two layer network and imbalanced datasets, the results obtained considering two classes are very consistent ($OA \geq 98.7\%$) and for the three classes problem they are still good, considering the quality of the data ($OA \sim 82\%$). Also, in comparison with TML, DL outperforms by $\Delta OA \sim 4\%$ and $\Delta AUC \sim 0.07$ for galaxies with $K \geq 5$.
- e) We make public a complete catalog for 670,560 galaxies, in the redshift range $0.03 < z < 0.1$, Petrosian magnitude in r-band brighter than 17.78, and $|b| \geq 30^\circ$. The input data comes from SDSS-DR7. We provide morphological classification using TML and DL, together with all parameters measured with our new non-parametric method (see 4.6 for catalog details). We append classifications (T-Type) from Nair and Abraham (2010) and Sánchez et al. (2018) whenever available.

6.2 Concluding remarks

In the big data era, it is of paramount importance to have the expertise to explore different computational approaches and systems to attack each specific problem.

The main paper of this research presented in Chapter 4 provides the best results we obtained in a catalog with morphological information for almost 700.000 galaxies. Here, I present concluding remarks on the hypotheses presented in Introduction (Chapter 1):

- a) **Hypothesis 1:** *It is hypothesized that TML and DL approaches can reliably perform galaxy morphology classification, considering the separation between ETGs and LTGs.* In all of our samples, both strategies have over 94.5% Overall Accuracy (OA) when distinguishing elliptical from spiral galaxies — 99% OA in average when using DL models, i.e., we can imitate classification provided by the human eye. Thus, our data, methodologies and experiments support this hypothesis.
- b) **Hypothesis 2:** *DL achieves higher standards of performance than TML for visual classification, however, TML has a similar performance (considering two classes) while preserving meaningful features.* DL’s higher standards of performance over ML are cited in the previous item. Given the two classes problem, TML and DL have $\Delta OA \approx 4.0\%$, $\Delta OA \approx 3.4\%$, $\Delta OA \approx 0.9\%$ for $K \geq 5$, $K \geq 10$, $K \geq 20$, respectively. We can safely assert that the performance for TML is not far behind. And, as shown throughout this document, TML features, and, consequently, classification, are more concretely and directly understandable for humans than DL’s — highlight to Figures 3.4 and 4.4 with plots for DL features and TML (CyMorph) features. Explaining further, it is possible to define and understand galaxy morphology types by analyzing non-parametric morphology features. Although one can state that we understand deep neural networks are extracting texture, rough and thin patterns from the images, it is not easy to have a rational grasp of these sets of features.

Chapter 5 presents a discussion on computational aspects. Annex chapters present other important publications I was involved in the course of this PhD.

It is worth mentioning other projects that I have collaborated with Dra. Marcelle Soares-Santos and her research group at Brandeis University in my one-year abroad (CAPES Sandwich Program)¹:

- a) Code improvement for galaxy cluster finder and characterization;

¹My academic website can be accessed through this link: <https://paulobarchi.github.io/>.

- b) Design, prototype and perform preliminary tests on detecting the EM counterpart of Gravitational Waves with Deep Learning;
- c) System for preprocessing stamps of galaxies for Dark Energy Survey database.²

²The repositories for such developments can be accessed either on my github (<https://github.com/paulobarchi>) or on the github of Dra. Marcelle Soares-Santos research group (<https://github.com/SSantosLab>).

REFERENCES

- ABBOTT, T. et al. The dark energy survey: more than dark energy - an overview. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 460, p. 1270–1299, 2016. [19](#)
- ABRAHAM, R. G. et al. The morphologies of distant galaxies. II. classifications from the hubble space telescope medium deep survey. **Astrophysical Journal Supplement Series**, v. 107, p. 1, nov. 1996. [1](#), [24](#), [25](#), [34](#), [36](#), [37](#)
- AL-JARRAH, O. Y.; YOO, P. D.; MUHAIDAT, S.; KARAGIANNIDIS, G. K.; TAHA, K. Efficient machine learning for big data: a review. **Big Data Research**, v. 2, n. 3, p. 87–93, 2015. [1](#)
- BAILLARD, A.; BERTIN, E.; DE LAPPARENT, V.; FOUQUÉ, P.; ARNOUITS, S.; MELLIER, Y.; PELLÓ, R.; LEBORGNE, J.-F.; PRUGNIEL, P.; MAKAROV, D.; MAKAROVA, L.; MCCRACKEN, H. J.; BIJAOU, A.; TASCA, L. The FIGI catalogue of 4458 nearby galaxies with detailed morphology. **Astronomy and Astrophysics**, v. 532, p. A74, aug. 2011. [16](#)
- BALDRY, I. K. et al. Quantifying the bimodal color-magnitude distribution of galaxies. **Astrophysical Journal**, v. 600, n. 2, p. 681–694, jan 2004. ISSN 0004-637X. [56](#)
- BALL, N.; BRUNNER, R. Data mining and machine learning in astronomy. **International Journal of Modern Physics D**, v. 19, n. 7, p. 1049–1106, 7 2010. ISSN 0218-2718. [14](#)
- BARCHI, P.; CARVALHO, R. de; ROSA, R.; SAUTTER, R.; SOARES-SANTOS, M.; MARQUES, B.; CLUA, E.; GONÇALVES, T.; SÁ-FREITAS, C. de; MOURA, T. Machine and deep learning applied to galaxy morphology-a comparative study. **Astronomy and Computing**, v. 30, p. 100334, 2020. [2](#), [19](#), [33](#), [35](#), [39](#), [40](#), [42](#), [43](#), [46](#), [48](#), [50](#), [53](#), [55](#), [57](#), [59](#), [60](#), [61](#), [64](#)
- BARCHI, P. H.; SAUTTER, R.; DA COSTA, F. G.; MOURA, T. C.; STALDER, D. H.; ROSA, R. R.; CARVALHO, R. R. de. Improving galaxy morphology with machine learning. **Journal of Computacional Interdisciplinary Sciences**, v. 7, n. 3, p. 114–120, 2016. [1](#), [16](#), [35](#)
- BEHNEL, S.; BRADSHAW, R.; CITRO, C.; DALCIN, L.; SELJEBOTN, D.; SMITH, K. Cython: the best of both worlds. **Computing in Science Engineering**, v. 13, n. 2, p. 31–39, 2011. ISSN 1521-9615. [65](#)

- BERTIN, E.; ARNOUITS, S. Sextractor: software for source extraction. **Astronomy and Astrophysics, Supplement**, v. 117, p. 393–404, jun. 1996. 20, 38
- BISHOP, C. M. **Pattern recognition and machine learning (information science and statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. 37
- BLANTON, M. R.; MOUSTAKAS, J. Physical properties and environments of nearby galaxies. **Annual Review of Astronomy and Astrophysics**, v. 47, n. 1, p. 159–210, 2009. 33
- BLUMER, A.; EHRENFEUCHT, A.; HAUSSLER, D.; WARMUTH, M. K. Occam’s razor. **Information Processing Letters**, v. 24, n. 6, p. 377–380, 1987. 49
- BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. **Pattern Recognition**, v. 30, n. 7, p. 1145–1159, 1997. 8, 45, 49, 50
- CONSELICE, C. J. The relationship between stellar light distributions of galaxies and their formation histories. **Astrophysical Journal Supplement Series**, v. 147, p. 1–28, jul. 2003. xi, 1, 22, 24, 25, 34, 36, 37, 41, 42
- CONSELICE, C. J.; BERSHADY, M. A.; JANGREN, A. The asymmetry of galaxies: physical morphology for nearby and high-redshift galaxies. **Astrophysical Journal**, v. 529, p. 886–910, feb. 2000. 34, 36
- CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. ISSN 0885-6125. 10
- DALCÍN, L.; PAZ, R.; STORTI, M. Mpi for python. **Journal of Parallel and Distributed Computing**, v. 65, n. 9, p. 1108–1115, 2005. ISSN 0743-7315. 66
- DE CARVALHO, R. R.; RIBEIRO, A. L. B.; STALDER, D. H.; ROSA, R. R.; COSTA, A. P.; MOURA, T. C. Investigating the relation between galaxy properties and the gaussianity of the velocity distribution of groups and clusters. **Astronomical Journal**, v. 154, p. 96, sep. 2017. 54
- DIELEMAN, S.; WILLETT, K. W.; DAMBRE, J. Rotation-invariant convolutional neural networks for galaxy morphology prediction. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 450, n. 2, p. 1441–1459, 2015. 16, 35

EISENSTEIN, D. J. et al. SDSS-III: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems. **Astrophysical Journal**, v. 142, n. 3, p. 72, aug 2011. [17](#), [18](#), [23](#), [34](#)

FEIGELSON, E. D.; BABU, J. **Statistical challenges in astronomy**. [S.l.]: Springer Science & Business Media, 2006. [1](#), [17](#), [34](#)

FERRARI, F.; DE CARVALHO, R. R.; TREVISAN, M. Morfometryka - a new way of establishing morphological classification of galaxies. **Astrophysical Journal**, v. 814, p. 55, Nov 2015. [1](#), [22](#), [23](#), [24](#), [25](#), [27](#), [34](#), [36](#), [37](#)

FREEMAN, P. E. et al. New image statistics for detecting disturbed galaxy morphologies at high redshift. **Monthly Notices of the Royal Astronomical Society**, v. 434, n. 1, p. 282–295, 06 2013. ISSN 0035-8711. [35](#)

GABRIEL, E.; FAGG, G. E.; BOSILCA, G.; ANGSKUN, T.; DONGARRA, J. J.; SQUYRES, J. M.; SAHAY, V.; KAMBADUR, P.; BARRETT, B.; LUMSDAINE, A.; CASTAIN, R. H.; DANIEL, D. J.; GRAHAM, R. L.; WOODALL, T. S. Open MPI: goals, concept, and design of a next generation mpi implementation. In: KRANZLMÜLLER, D.; KACSUK, P.; DONGARRA, J. (Ed.). **Recent advances in parallel virtual machine and message passing interface**. Berlin, Heidelberg: Springer, 2004. p. 97–104. [66](#)

GÉRON, A. **Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems**. [S.l.]: O'Reilly Multimedia, 2019. ISBN 9781492032649. [10](#), [11](#), [13](#), [14](#), [27](#), [28](#), [44](#), [45](#), [50](#)

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT Press, 2016. [11](#), [12](#), [13](#), [27](#), [28](#), [35](#), [36](#), [44](#), [45](#), [47](#)

GRAHAM, A. W.; ERWIN, P.; CAON, N.; TRUJILLO, I. A correlation between galaxy light concentration and supermassive black hole mass. **The Astrophysical Journal**, v. 563, n. 1, p. L11–L14, dec 2001. [22](#), [23](#)

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** Las Vegas, NV, USA: IEEE, 2016. [29](#), [30](#)

HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18–28, 1998. ISSN 1094-7167. 10

HUBBLE, E. **Realm of the nebulae**. [S.l.]: New Haven: Yale University Press, 1936. ISBN 9780300025002. 1, 5, 33

HUBBLE, E. P. Extragalactic nebulae. **Astrophysical Journal**, v. 64, p. 321–369, 1926. 1, 5, 33

HUERTAS-COMPANY, M. et al. A catalog of visual-like morphologies in the 5 candels fields using deep learning. **Astrophysical Journal Supplement Series**, v. 221, p. 8, 2015. 16, 35

HUERTAS-COMPANY, M. et al. Deep learning identifies high-z galaxies in a central blue nugget phase in a characteristic mass range. **Astrophysical Journal**, American Astronomical Society, v. 858, n. 2, p. 114, may 2018. 16, 35

IVEZIĆ, Ž.; CONNOLLY, A. J.; VANDERPLAS, J. T.; GRAY, A. **Statistics, data mining, and machine learning in astronomy: a practical python guide for the analysis of data**. Princeton, NJ, USA: Princeton University Press, 2014. ISBN 9780691151687. 1, 14, 17, 34

JIA, Y.; SHELFHAMER, E.; DONAHUE, J.; KARAYEV, S.; LONG, J.; GIRSHICK, R.; GUADARRAMA, S.; DARRELL, T. Caffe: convolutional architecture for fast feature embedding. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 24., 2014. **Proceedings...** New York, NY, USA: ACM, 2014. (MM '14), p. 675–678. ISBN 978-1-4503-3063-3. 67

KASZYNSKI, R.; PISKOROWSKI, J. New concept of delay equalized low-pass butterworth filters. In: IEEE INTERNATIONAL SYMPOSIUM ON INDUSTRIAL ELECTRONICS, 2006. **Proceedings...** Montreal, Quebec, Canada, 2006. v. 1, p. 171–175. ISSN 2163-5137. 37

KENT, S. M. Ccd surface photometry of field galaxies. ii - bulge/disk decompositions. **Astrophysical Journal Supplement Series**, v. 59, p. 115–159, 1985. 22, 36

KHALIFA, N. E. M.; TAHA, M. H. N.; HASSANIEN, A. E.; SELIM, I. M. Deep galaxy: classification of galaxies based on deep convolutional neural networks. **arXiv e-prints**, Sep 2017. 2, 16, 35

KHAN, A.; HUERTA, E. A.; WANG, S.; GRUENDL, R. Unsupervised learning and data clustering for the construction of galaxy catalogs in the dark energy survey. **arXiv e-prints**, Dec 2018. 35

KIM, I.; LEE, A. B.; LEI, J. Global and local two-sample tests via regression. **arXiv e-prints**, May 2019. 38

KNUTH, D. E.; GREENE, D. H. **Mathematics for the analysis of algorithms**. 3. ed. USA: Birkhauser Boston, 1999. ISBN 0817635157. 66

LAW, D. R. et al. The physical nature of rest-uv galaxy morphology during the peak epoch of galaxy formation. **Astrophysical Journal**, v. 656, p. 1–26, 02 2007. 35

LECUN, Y. Generalization and network design strategies. In: PFEIFER, R.; SCHRETER, Z.; FOGELMAN, F.; STEELS, L. (Ed.). **Connectionism in perspective**. New York: Elsevier, 1989. 12

LINTOTT, C. J. et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 389, n. 3, p. 1179–1189, 09 2008. ISSN 0035-8711. 7, 18, 29, 34, 35, 47, 63

LINTOTT, C. J. et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 410, n. 1, p. 166–178, 2011. 7, 18, 29, 34, 35, 47, 63

LOTZ, J. M.; PRIMACK, J.; MADAU, P. A new nonparametric approach to galaxy morphological classification. **Astronomical Journal**, v. 128, p. 163–182, jul. 2004. xi, 1, 25, 34, 36, 37, 41, 42

MARTIN, D.; GONCALVES, T.; DARVISH, B.; SEIBERT, M.; SCHIMINOVICH, D. Quenching or bursting: Star formation acceleration - a new methodology for tracing galaxy evolution. **Astrophysical Journal**, v. 842, n. 1, 2017. ISSN 15384357. 56

MITCHELL, T. M. **Machine learning**. New York, NY, USA: McGraw-Hill, 1997. ISBN 0070428077, 9780070428072. 8, 11, 27, 28, 44, 45

MO, H.; BOSCH, F. Van den; WHITE, S. **Galaxy formation and evolution**. [S.l.]: Cambridge University Press, 2010. ISBN 9780521857932. 1, 5

- MORGAN, W. W.; MAYALL, N. U. A spectral classification of galaxies. **Publications of the Astronomical Society of the Pacific**, v. 69, p. 291, 1957. 22, 36
- NAIR, P. B.; ABRAHAM, R. G. A catalog of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey. **Astrophysical Journal Supplement Series**, v. 186, n. 2, p. 427–456, feb 2010. xi, 58, 59, 60, 70
- NOGUEIRA-CAVALCANTE, J. P.; GONÇALVES, T. S.; MENÉNDEZ-DELMESTRE, K.; SHETH, K. Star formation quenching in green valley galaxies at $0.5 < z < 1.0$ and constraints with galaxy morphologies. **Monthly Notices of the Royal Astronomical Society**, v. 473, n. 1, p. 1346–1358, jan 2018. ISSN 0035-8711. 56
- PEDREGOSA, F. et al. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. ISSN 1532-4435. 28, 44, 47, 66
- PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. [S.l.]: Thomson Learning, 2007. 528 p. ISBN 978-85-221-0595-3. 37
- PETROSIAN, V. Surface brightness and evolution of galaxies. **Astrophysical Journal**, v. 209, p. L1–L5, 1976. 18, 23
- POZZETTI, L. et al. Zcosmos - 10k-bright spectroscopic sample - the bimodality in the galaxy stellar mass function: exploring its evolution with redshift. **Astronomy and Astrophysics**, v. 523, p. A13, 2010. 33
- PRESS, S. J. **Applied multivariate analysis: using Bayesian and frequentist methods of inference**. [S.l.]: Dover Publications, 2005. ISBN 9780486442365 / 0486442365. 25, 37
- QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, p. 81–106, 1986. 10
- RAMOS, F. M.; ROSA, R. R.; NETO, C. R.; ZANANDREA, A. Generalized complex entropic form for gradient pattern analysis of spatio-temporal dynamics. **Physica A: Statistical Mechanics and its Applications**, v. 283, n. 1, p. 171 – 174, 2000. ISSN 0378-4371. 37

- ROBERTS, M. S.; HAYNES, M. P. Physical parameters along the hubble sequence. **Annual Review of Astronomy and Astrophysics**, v. 32, n. 1, p. 115–152, 1994. 33
- ROSA, R. R.; CAMPOS, M. R.; RAMOS, F. M.; VIJAYKUMAR, N. L.; FUJIWARA, S.; SATO, T. Gradient pattern analysis of structural dynamics: application to molecular system relaxation. **Brazilian Journal of Physics**, v. 33, n. 3, p. 605–610, 2003. 37
- ROSA, R. R.; DE CARVALHO, R. R.; SAUTTER, R. A.; BARCHI, P. H.; STALDER, D. H.; MOURA, T. C.; REMBOLD, S. B.; MORELL, D. R. F.; FERREIRA, N. C. Gradient pattern analysis applied to galaxy morphology. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 477, n. 1, p. L101–L105, 2018. 1, 26, 27, 34, 35, 36, 37, 38
- ROSA, R. R.; SHARMA, A. S.; VALDIVIA, J. A. Characterization of asymmetric fragmentation patterns in spatially extended systems. **International Journal of Modern Physics C (IJMPC)**, v. 10, n. 01, p. 147–163, 1999. 37
- RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 2015. 36
- SÁNCHEZ, H. D.; HUERTAS-COMPANY, M.; BERNARDI, M.; TUCCILLO, D.; FISCHER, J. L. Improving galaxy morphologies for sdss with deep learning. **Monthly Notices of the Royal Astronomical Society (MNRAS)**, v. 476, n. 3, p. 3661–3676, 2018. xi, 2, 16, 35, 36, 58, 60, 61, 70
- SAUTTER, R. A. **Gradient pattern analysis: new methodological and computational features and application**. Dissertation (Master in Applied Computing) — National Institute for Space Research (INPE), São José dos Campos, Brasil, 2018. 37, 38
- SAUTTER, R. A.; BARCHI, P. H. pyghs: computing geometric histogram separation for binomial proportion patterns. **Journal of Computational Interdisciplinary Sciences**, v. 8, n. 1, p. 121, 2017. 35, 38
- SCHAWINSKI, K. et al. The green valley is a red herring: galaxy zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies. **eprint arXiv:1402.4814**, 2014. 56
- SCHAWINSKI, K.; ZHANG, C.; ZHANG, H.; FOWLER, L.; SANTHANAM, G. K. Generative adversarial networks recover features in astrophysical images of

- galaxies beyond the deconvolution limit. **Monthly Notices of the Royal Astronomical Society: Letters**, v. 467, n. 1, p. L110–L114, 2017. 16
- SZEGEDY, C. et al. Going deeper with convolutions. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2015. **Proceedings...** Boston, MA, USA: IEEE, 2015. 14, 15, 29, 30
- TAKAMIYA, M. Galaxy structural parameters: star formation rate and evolution with redshift. **Astrophysical Journal Supplement Series**, v. 122, n. 1, p. 109–150, may 1999. 36
- TAN, P.-N.; STENBACH, M.; KUMAR, V. **Introduction to data mining**. Boston, MA, USA: Addison-Wesley Longman Publishing, 2005. ISBN 0321321367. 1, 17
- VASCONCELLOS, E.; CARVALHO, R. de; GAL, R.; LABARBERA, F.; CAPELATO, H.; VELHO, H. F. C.; TREVISAN, M.; RUIZ, R. Decision tree classifiers for star/galaxy separation. **The Astronomical Journal**, v. 141, n. 6, p. 189, 2011. 14
- VAUCOULEURS, G. de. Revised classification of 1500 bright galaxies. **Astrophysical Journal Supplement Series**, v. 8, p. 31, apr. 1963. 16, 33
- WAY, M. J.; SCARGLE, J. D.; ALI, K. M.; SRIVASTAVA, A. N. **Advances in machine learning and data mining for astronomy**. New York, NY, USA: Chapman & Hall/CRC, 2012. ISBN 143984173X, 9781439841730. 1, 34
- WILLETT, K. W. 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 (MNRAS)**, v. 435, n. 4, p. 2835–2860, 2013. 8, 18, 29, 34, 35, 63
- WYDER, T. K. et al. The UV-optical galaxy color-magnitude diagram. i. basic properties. **Astrophysical Journal Supplement Series**, v. 173, n. 2, p. 293–314, dec 2007. ISSN 0067-0049. 56
- YANG, Q.; WU, X. 10 challenging problems in data mining research. **International Journal of Information Technology and Decision Making**, v. 5, n. 4, p. 597–604, 2006. 47
- YEAGER, L.; BERNAUER, J.; GRAY, A.; HOUSTON, M. Digits: the deep learning gpu training system. In: ICML AUTO ML WORKSHOP, 2015. **Proceedings...** Paris, France, 2015. 67

ANNEX A - PUBLICATIONS

A.1 Publication 1

Title: *Machine and Deep Learning Applied to Galaxy Morphology - A comparative Study.*

Authors: P.H.Barchi, R.R.de Carvalho, R.R.Rosa, R.A.Sautter, M.Souares-Santos, B.A.D.Marques, E.Clua, T.S.Gonçalves, C.de Sá-Freitas, T.C. Moura

Journal: Astronomy and Computing, Elsevier.

Volume: 30.

Number: 100334.

Year: 2020.

DOI: <https://doi.org/10.1016/j.ascom.2019.100334>.

A.2 Publication 2

Title: *Gradient Pattern Analysis Applied to Galaxy Morphology.*

Authors: R. R. Rosa, R. R. de Carvalho, R. Sautter, P. H. Barchi, D. H. Stalder, T. C. Moura & N. C. Ferreira.

Journal: Monthly Notices of the Royal Astronomical Society.

Volume: 477.

Issue: 1.

Pages: L101–L105.

Year: 2018.

DOI: <https://doi.org/10.1093/mnrasl/sly054>.

A.3 Publication 3

Title: *Modeling Social and Geopolitical Disasters as Extreme Events: A Case Study Considering the Complex Dynamics of International Armed Conflicts.*

Authors: R. R. Rosa, J. Neelakshi, G. A. L. L. Pinheiro, P. H. Barchi, E. H. Shigue-mori.

Book title: Towards Mathematics, Computers and Environment: A Disasters Per-spective.

Chapter: 12.

Pages: 233–254.

Publisher: Springer International Publishing.

Year: 2019.

DOI: https://doi.org/10.1007/978-3-030-21205-6_12.

A.4 Publication 4

Title: *pyGHS: Computing Geometric Histogram Separation in Binomial Proportion Patterns.*

Authors: R. Sautter, P. H. Barchi.

Journal: Journal of Computational Interdisciplinary Sciences.

Volume: 8.

Issue: 1.

Paper: 121.

Year: 2017

DOI: <https://doi.org/10.6062/jcis.2017.08.01.0121>

A.5 Publication 5

Title: *Improving Galaxy Morphology with Machine Learning.*

Authors: P. H. Barchi, R. Sautter, F. G. da Costa, T. C. Moura, D. H. Stalder, R. R. Rosa & R. R. de Carvalho.

Journal: Journal of Computational Interdisciplinary Sciences.

Volume: 7.

Issue: 3.

Pages: 114–120.

Year: 2016.

DOI: <https://doi.org/10.6062/jcis.2016.07.03.0114>

ANNEX B - EXTRACTING THE SLIME MOLD GRAPH FROM THE COSMIC WEB

I have participated in the *Kavli Summer School in Astrophysics 2019: Machine Learning in the era of large astronomical surveys* at University of California, Santa Cruz (UCSC). The full version of the report delivered to the program can be accessed through this link as well: <https://kspa.soe.ucsc.edu/sites/default/files/Barchi.pdf>.

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.