UNIVERSITY OF NOTTINGHAM

SCHOOL OF COMPUTER SCIENCE

Novel Automated Classification Approaches for Citizen Science

Author:
Manuel Alejandro
Jiménez Morales

Supervisors:
Dr. Isaac Triguero
Dr. Mercedes Torres Torres
Dr. Robert John



Doctoral dissertation submitted in partial fulfillment
of the requirements for the degree of Doctor of Philosophy

September 15, 2020

Declaration

I confirm that this thesis dissertation presented for the degree of Doctor of Philosophy at the School of Computer Science, University of Nottingham, has

- been composed entirely by myself,
- been solely the result of my own work except stated otherwise,
- not been submitted in whole or in part for any other degree or qualification in any other academic institution.

Manuel Alejandro Jiménez Morales
September 2020

"... if we stand on the shore and look at the sea, we see the water, the waves breaking, the foam, the sloshing motion of the water, the sound, the air, the winds and the clouds, the sun and the blue sky, and light; there are sand and there are rocks of various hardness and permanence, color and texture. There are animals and seaweed, hunger and disease, and the observer on the beach; there may be even happiness and thought. [...] *Curiosity demands that we ask questions, that we try to put things together and try to understand this multitude of aspects as perhaps resulting from the action of a relatively small number of elemental things and forces acting in an infinite variety of combinations."*

Richard P. Feynman
Six Easy Pieces, Chpt. 2

Abstract

Citizen science, traditionally known as the engagement of amateur participants in research, is showing a great potential for large-scale processing of data. In areas such as astronomy, ecology, or geo-sciences, where emerging technologies generate huge volumes of data, citizen science projects enable image classification at a rate not possible to accomplish by experts alone. Using the power of the web, virtual communities of volunteers sharing a common goal are able to coordinate the classification of hundreds of thousands of images in a reasonable amount of time. However, expert evaluations usually reveal biases and uncertainty in the results, since the participants involved are typically inexperienced in the task and hold variable skills and backgrounds. Consequently, the research community tends to distrust citizen science outcomes, claiming a generalised lack of accuracy and validation, and leaving the major part of the resulting data unemployed after the finalisation of the projects.

Citizen science also offers a great amount of labelled data at a reduced cost for the training of machine learning classifiers. Nonetheless, current efforts attempting the exploitation of citizen science outcomes with machine learning tools have ignored the inherent uncertainty in results as well as the potential of expert classifications to ameliorate this issue. The ultimate goal has mainly been to replicate the amateur endeavours, thus propagating their biases and limitations in the automated classification. Similarly, the potential behind the learning from unlabelled data to alleviate this uncertainty has also been disregarded. This framework claims for a solution that can take advantage of all levels of knowledge: expert classifications, citizen science data, and unlabelled data. However, the synergy between these sources of data remains unexplored, waiting for the development of new methodologies that may lead to an enhanced automated classification.

This thesis focuses on the development of automated approaches for classification problems aided by citizen science projects on the web, aiming to leverage the inherent uncertainty in the results and all levels of knowledge available about the problem. As a case study, we select the longest running implementation of a scientific problem aided by modern citizen science: the classification of galaxies from images. We exploit the results of the first edition of the *Galaxy Zoo*, a citizen science project that nowadays represents the largest galaxy image database manually annotated. The research is completed through three progressive stages. First, we introduce a novel multi-stage approach to handle the uncertainty within data labelled in the course of citizen science projects. Our method proposes a set of transformations that leverage the uncertainty in amateur classifications in conjunction with a hybridisation strategy that provides the best aggregation of the transformed data for improving the quality and confidence in the results. The second stage comprises a thorough study of machine learning methods for image classification, introducing the use of autoencoders to learn from unlabelled data, and exploring the learning from amateur and expert classifications by the exploitation of pre-training and fine-tuning of convolutional neural networks. Finally, in the third stage of the research, the previous findings are combined to propose a solution to the novel learning paradigm defined that is able to exploit data either labelled by experts and amateurs in the course of citizen science projects, and unlabelled data.

In summary, the research conducted here introduces a set of novel mechanisms towards an improved automated classification based on citizen science data, expert classifications, and raw data. As a result, the proposed method for handling the uncertainty boosts the accuracy and is able to classify a higher number of images in comparison with previous approaches. This is accomplished by taking advantage of the uncertainty measured by participants themselves. The use of autoencoders greatly speeds up feature extraction with respect to state-of-the-art methods, also revealing the potential behind the exploitation of amateur and expert classifications by deep learning-based classifiers. In last place, a novel approach leverages all insights previously found and presents an innovative setting to learn from expert and amateur classifications and unlabelled data that surpasses the performance obtained using such label sets separately or joint. These results have also signified a global study of the automated classification of galaxy images problem that, from state-of-the-art approaches, have contributed new methods built on the boundary amongst citizen science, astroinformatics, and machine learning fields of study.

Acknowledgements

This place for the acknowledgements could not begin without a warm and grateful memory of my supervisor the Professor Robert John, who sadly will not see the completion of this academic journey. *Bob*, as was widely known within our research group, played a key role for my acceptance and subsequent eligibility for the Ph.D. scholarship at the University of Nottingham that has resulted in the work presented here. I will always keep a fond memory of his bonhomie and kindness with every one in the school. Together with Isaac Triguero and Mercedes Torres Torres, my supervisors have meant to me an example of close support and dedication to their Ph.D. students in every single step of our trip, which can only arise from a true vocation to academia. My most sincere gratitude to the three of them.

Along with my supervisors, I would also like to acknowledge the support of my family and friends during these three years of Ph.D. studies. It has been an honour to be part of the Computational Optimisation and Learning (COL) group, the former ASAP, where I have shared really enjoyable moments with colleagues and staff. My grateful thanks to my friends in Nottingham, who have made of this journey something much more pleasant and easier to bear with. And last but not least, thanks to my family for their close support and attention during this time away from home.

Contents

Abstract **vii**

Acknowledgements **ix**

Contents **xi**

List of Figures **xv**

List of Tables **xvii**

List of Acronyms **xxi**

1 Introduction **1**
1.1 Citizen Science for Classification 2
1.2 Case Study: Classification of Galaxy Images 4
1.3 Motivation and Objectives 6
1.4 Contributions 9
1.5 Structure of the Thesis 11

2 Background **13**
2.1 Introduction 13
2.2 Citizen Science: An Overview 14
2.3 Astroinformatics and Galaxy Classification 17
2.4 Manual Classification: The *Galaxy Zoo* Project 19
2.4.1 GZ1 Data 20
2.5 Machine Learning Fundamentals 21
2.5.1 Classic Machine Learning 22
2.5.2 Deep Learning Approaches 24
Convolutional Neural Networks 25
Autoencoders 27
2.5.3 Performance Metrics 28
2.6 Machine Learning for Galaxy Classification 29
2.6.1 Feature Extraction *plus* Classifier 30

2.6.2 Convolutional Neural Networks 31
2.6.3 Autoencoders 32
2.7 Summary 32

3 Handling Uncertainty in Citizen Science Data 35
3.1 Introduction 35
3.2 Preliminaries 36
3.3 Methodology 37
3.3.1 Motivation: Three Sources of Uncertainty within CzS Data 38
3.3.2 Three Transformations for Data Refinement 40
Normalisation: Reinforcement of Main Classes 41
DK Votes Shift: Evaluation of Measured Uncertainty . . 42
Votes Boost: Addition of Confidence to Highly-voted Examples 44
3.3.3 Hybridisation Strategy 45
3.4 Case Study: Improving Galaxy Classification with CzS Data . . 47
3.4.1 Galaxy Zoo Classifications 48
3.4.2 Experiments Evaluation 49
3.4.3 Experimental Setting 50
Single Transformations Test 52
Hybridisation of Transformations 55
3.5 Discussion of Results 56
3.6 Summary 60

4 Galaxy Image Classification Based on Citizen Science Data 63
4.1 Introduction 63
4.2 Methodology: Galaxy Image Classification with Deep Learning 64
4.2.1 Autoencoders for Feature Extraction on Galaxy Images . 64
4.2.2 Convolutional Neural Network for Galaxy Image Classification 67
4.2.3 Pre-training of CNNs with CzS Data 68
4.3 Experimental Setup 69
4.3.1 Standard Classification Algorithms 69
4.3.2 Feature Selection Methods 70
4.3.3 DL-based Models 71
4.3.4 Image Data and Evaluation of Experiments 72
4.4 Results and Analysis 74
4.4.1 GZ1-E: Expert Classifications 74
4.4.2 GZ1-A: Amateur Classifications 78

Scalability of Methods 78
Pre-training with Amateur and Expert Labels 81
4.4.3 Analysis of Results 82
4.5 Summary 84

5 CzSL: A Novel Learning Paradigm for Citizen Science 87
5.1 Introduction 87
5.2 CzSL: Learning from CzS Data, Experts and Unlabelled Data . . 89
5.2.1 Motivation: Levels of Knowledge 89
5.2.2 CzSL: A Unified Deep Learning Approach for Citizen Science 90
5.2.3 Implementation 93
CAE-CNN Pre-training and CNN Fine-tuning 93
MLP for Expert Knowledge Extension 95
5.3 Experimental Framework 96
5.3.1 Datasets 96
5.3.2 Experimental Evaluation 98
5.3.3 Comparative Approaches and Parameters 100
5.4 Results and Analysis 103
5.4.1 Oracles 103
5.4.2 Binary and Multi-class Classification 105
5.4.3 Imbalanced Classification 108
5.5 Summary 110

6 Conclusions 113
6.1 Summary of Contributions 113
6.1.1 Adaptation of CzSL to Other Problems 116
6.2 Limitations 119
6.3 Future Work 122

A Galaxy Zoo 1: Data Exploratory Analysis 125
A.1 Introduction 125
A.2 Distribution of the data 126
A.2.1 Original and Debiased Scores 127
A.2.2 Total Number of Votes 129
A.3 GZ1 Classifications 132
A.4 Expert Classifications 133
A.5 Evaluation of Different Thresholds 134
A.6 Summary 136

B Fast and Robust Fuzzy-kNN Classification with CzS Data **137**
B.1 Introduction 137
B.2 Preliminaries 138
B.2.1 Uncertainty in CzS Data 138
B.2.2 Fuzzy k-Nearest Neighbour Classification 139
B.3 Methodology 140
B.3.1 Human-based Fuzzification 140
B.3.2 Instance Selection Based on Confidence 141
B.4 Experiments and Analysis 142
B.4.1 Experimental Setup 142
B.4.2 Results and Analysis 145
B.5 Summary 147

Bibliography **149**

List of Figures

1.1 Two gaps generated at the end of CzS projects. 7
1.2 Levels of knowledge within CzS data. 8

2.1 Potential uses of ML within the CzS workflow. 16
2.2 The Hubble tuning fork. 18
2.3 Selection of original images from the GZ1 project. 21
2.4 Schematic representation of a feed-forward ANN. 25
2.5 Representation of a convolution layer followed by a pooling layer. 26
2.6 Schematic view of the general structure of an AE. 27
2.7 Confusion matrix diagram (2-class problem). 28
2.8 Example of CNN employed in galaxy classification. 32

3.1 Three sources of uncertainty present in CzS data. 46
3.2 General workflow of the proposed approach for handling the uncertainty y CzS data. 48
3.3 Test of transformation {1} over raw and debiased scores. 53
3.4 Test of transformation {2} over raw and debiased scores. 54
3.5 Distribution of DK votes and total number of votes across the GZ1 subset in logarithmic scale. 55
3.6 Test of sequence {13} over raw and debiased scores. 55
3.7 Ranking of hybrid sequences of transformations with fixed values of parameters. 57
3.8 Ranking of hybrid sequences of transformations implementing parameter optimisation and 70/30 validation of the data. 58

4.1 Architecture of the proposed deep autoencoder (DAE). 66
4.2 Architecture of the proposed convolutional autoencoder (CAE). 66
4.3 Architecture of the proposed CNN. 67
4.4 FE runtime in logarithmic scale for GZ1-E sample. 75
4.5 Sample of galaxy images reconstructed by the proposed AE models. 75

4.6 Classification runtimes in logarithmic scale for 64x images of the GZ1-E sample. 77
4.7 Classification runtimes in logarithmic scale for 128x images of the GZ1-E sample. 78
4.8 Total classification runtime in logarithmic scale for both strategies studied with the GZ1-E sample. 79
4.9 Sample of 128x colour images from the GZ1-E subset. 80
4.10 FE and classification runtimes in logarithmic scale for the GZ1-A sample. 81
4.11 Total classification runtime in logarithmic scale for both strategies studied with the GZ1-A sample. 81

5.1 The CzSL learning workflow. 94
5.2 Architecture of the MLP implemented in the CzSL approach. . . 96
5.3 Execution times in logarithmic scale for oracles. 104
5.4 Execution times in logarithmic scale for GZ1-(SE) and GZ1-(RLE) datasets. 106
5.5 Graphical representation of experiments' runtime in logarithmic scale for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques. 109

A.1 Number of galaxies across original and debiased score types. . . 127
A.2 Number of galaxies across original and debiased score types (cumulative). 128
A.3 Number of galaxies across the total number of votes received. . 130
A.4 Number of galaxies across the total number of votes received in logarithmic scale. 130
A.5 Number of galaxies across the total number of votes received (cumulative). 131
A.6 Number of galaxies across total number of votes received with a Normal distribution overlapped. 131
A.7 Comparative representation of Acc and RR achieved by each score type across a selected set of thresholds. 136

B.1 Workflow of the proposed approach with the Fuzzy-kNN classifier. 143
B.2 Number of instances removed by the different thresholds (Th). . 146
B.3 Comparison of algorithms' runtime. 146

List of Tables

2.1 Typical look of the data recorded in the course of a CzS project involving image classification. 15

3.1 Normalised scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1. 42
3.2 Shifted scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1. 44
3.3 Modified scores obtained from Equation 3.5 for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1. . . 45
3.4 Evaluation of GZ1 flags using expert catalogues separately. . . . 50
3.5 Evaluation of GZ1 flags using the joint expert catalogue. 50

4.1 Topology of the proposed DAE. 65
4.2 Topology of the proposed CAE. 67
4.3 Topology of the proposed CNN. 68
4.4 Parameters for the selected standard classifiers. 70
4.5 Data samples taken from the GZ1-T2 dataset used in the experiments. 72
4.6 Number of extracted features in GZ1-E subset. 74
4.7 Results for 64x and 128x images of GZ1-E sample. 76
4.8 Results of proposed CNN and ResNet for 64x and 128x images of the GZ1-E sample. 76
4.9 Results for the GZ1-A sample. 79
4.10 Results of proposed CNN and ResNet for the GZ1-A sample. . . 81
4.11 Results of proposed CNN and ResNet, implementing the pre-training and fine-tuning with the GZ1-A* and GZ1-E subsets, respectively, and with no pre-training process. 82

5.1 Datasets used in the experiments with their distribution of classes in percentage. 98
5.2 Comparative approaches established for the evaluation of the CzSL model. 102

5.3 Results for oracles in the three datasets investigated. 104
5.4 Results of all approaches investigated with GZ1-(SE) and GZ1-(RLE) datasets. 106
5.5 Breakdown of CzSL model runtime for GZ1-(SE) and GZ1-(RLE) datasets. 106
5.6 Breakdown MLP-SL model runtime for GZ1-(SE) and GZ1-(RLE) datasets. 106
5.7 Results of all approaches investigated for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques. 108
5.8 Breakdown of CzSL model runtime for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques. 108
5.9 Breakdown of MLP-SL model runtime for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques. 109

A.1 Basic statistics for original and debiased score types. 128
A.2 Basic statistics for total number of votes received by each galaxy image. 129
A.3 Disagreement between GZ1 flags and a 0.8 threshold over debiased scores. 132
A.4 Count of galaxies as classified by the GZ1 flags with respect to the application of different thresholds over the debiased *Elliptical* score. 133
A.5 Count of galaxies as classified by the GZ1 flags with respect to the application of different thresholds over the debiased *Spiral* score. 133
A.6 Comparison between GZ1 flags and expert classifications from MOSES and Longo catalogues. 134
A.7 Evaluation of GZ1 flags considering the joint catalogue from MOSES and Longo as ground truth. 134
A.8 Acc and RR values for GZ1 classifications (GZ1 flags) considering the joint catalogue from MOSES and Longo as ground truth. 135
A.9 Number of prototypes. 135

B.1 Typical look of the data recorded in the course of a CzS project involving image classification. 138
B.2 Parameter settings of the algorithms involved. 144
B.3 Averaged Acc and runtime values for the comparative study of algorithms. 145

B.4 Averaged Acc and runtime values of CSF-kNN with instance selection. 146
B.5 Results of CSF-kNN with 1.0 threshold over the whole GZ1 dataset. 147

List of Acronyms

Acc	**Acc**uracy
AE	**A**uto**E**ncoder
ANN	**A**rtificial **N**eural **N**etwork
CAE	**C**onvolutional **A**uto**E**ncoder
CNN	**C**onvolutional **N**eural **N**etwork
CzS	**C**iti**z**en **S**cience
CzSL	**C**iti**z**en **S**cience **L**earning
DAE	**D**eep **A**uto**E**ncoder
DK	**D**on't **K**now
DL	**D**eep **L**earning
DM	**D**ata **M**ining
FE	**F**eature **E**xtraction
FS	**F**eature **S**election
GAN	**G**enerative **A**dversarial **N**etwork
GZ1	**G**alaxy **Z**oo **1**
GZ1-A	**GZ1 A**mateur
GZ1-E	**GZ1 E**xpert
GZ1-T2	**GZ1 T**able **2**
HV	**H**yper**V**olume
IR	**I**mbalance **R**atio
IU	**I**nherent **U**ncertainty
kNN	**k**-**N**earest **N**eighbour
LC	**L**evel of **C**onfidence
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
MU	**M**easured **U**ncertainty
RF	**R**andom **F**orests
ROS	**R**andom **U**nder-**S**ampling
RR	**R**ejection **R**ate
RUS	**R**andom **O**ver-**S**ampling
SDSS	**S**loan **D**igital **S**ky **S**erver
SSL	**S**emi-**S**upervised **L**earning
SVM	**S**upport **V**ector **M**achine

Dedicated to all my teachers:
those present, those who were,
and those who are yet to come.

Chapter 1

Introduction

Nowadays, an increasing number of research fields are faced with the (automated) generation of huge volumes of data. Taking the form of telescopes, geosensors or medical equipment, these modern facilities are providing professionals and researchers with tons of hardly manageable images, which remain unlabelled and awaiting experts to manually label it. As a partial solution, this environment has promoted the approach of citizens to science, greatly boosted by the widespread use of the Internet. Myriads of efforts have been joined together in the so-called *citizen science*, producing significant contributions to science and research through the implementation of projects on the web (Cohn, 2008). However, this approach still constitutes a temporary solution, since upcoming technologies announce the generation of data at rates not even possible to be processed with the aid of citizen science. The development of automated methods to deal with innate uncertainty in amateur classifications and leverage expert knowledge will be crucial to obtain accurate models. These will derive from an improved learning framework able to exploit all available resources: experts, amateurs, and unlabelled data.

This thesis deals with classification problems aided by citizen science. As a case study, the classification of galaxy images is employed, which highlights a scientific problem that has received an extensive set of results from the development of citizen science projects through the last years. In this chapter, the basic concepts and structure of the conducted research are presented. The fundamentals of citizen science and the case study are given in Sections 1.1 and 1.2, respectively. After this, the motivation for this research along with the research questions that naturally arise are discussed in Section 1.3. The contributions of the thesis are outlined in Section 1.4. Finally, the structure of the whole document is summarised in Section 1.5.

1.1 Citizen Science for Classification

Connectivity is promoting the emergence of a great potential amongst virtual communities of people that share a common goal. In some cases, this goal may consist in making a significant contribution towards the solution of a complex scientific problem. Whereas in the past the analysis of these problems used to be restricted to a group of experts in the subject, today this research has become infeasible when the processing of large amounts of data is required. In this context, citizen science (CzS) refers to the development of scientific research conducted by amateur volunteers from the general public (Cohn, 2008). As a form of crowdsourcing (Brabham, 2008), this practice is re-emerging and encouraging the crowd to help researchers to complete high time-consuming tasks for which no reliable automatic procedures are available yet, such as, labelling of images (Wright et al., 2017), detection of patterns in graphic data (Zevin et al., 2017), or transcription of handwritten texts (Grayson, 2016).

What is now widely known under the term *citizen science* has a long history, rooted in the earliest stages of modern science (Silvertown, 2009). The origin of CzS goes back to the origin of science itself, given that the very first scientists used to participate in research-related tasks as volunteers. However, the emergence of information technologies (e.g. mobile applications, cloud computing, wireless and sensor technologies) and the Internet have fostered the exploitation of these facilities for the development of CzS projects on the web (Bonney et al., 2014), greatly expanding the potential of this concept under the umbrella of multiple initiatives. CzS projects nowadays cover a wide range of disciplines such as astronomy, ecology, environmental sciences, medicine, or the study of historical legacies, often encompassed in on-line platforms that host the projects and make possible an easy development by researchers.

These modern implementations of CzS have turned into a method of distributed data analysis by simply taking advantage of the innate human capacity for pattern recognition and anomaly detection (Trouille, Lintott, and Fortson, 2019). For instance, the *EteRNA* (Lee et al., 2014) on-line gaming environment challenges players to design new ways to fold RNA molecules to find solutions for diseases such as tuberculosis (Cooper et al., 2010). These new molecular structures are then synthesised in Stanford's medical laboratories. Other examples of on-line CzS projects include *Eyewire* (Kim et al., 2014), which aims to shed light over the mystery around mammalians' retina ability to detect motion by

asking volunteers to colour serial electron microscopy images; or *Cosmoquest*[1], where people from the general public can take part in actual NASA science projects. There are also a growing number of projects devoted to the transcription of ancient documents across humanities, such as *From the Page*[2] or the Smithsonian Transcript Center[3]. For a comprehensive list of up-to-date listing of CzS projects taking place, we refer to the CitizenScience[4] and the SciStarter[5] websites, two of the most widely extended communities for the development of crowdsourcing and CzS projects worldwide.

This thesis focuses on classification problems in CzS, which generally aim at the classification of huge collections of images according to a number of classes. These classes capture the interest of a specific research field, and identifying them correctly is the target of the participants. This sort of project involves, for instance, the recognition of structures in cell images (Reis et al., 2015), animal species in images taken in the Savannah (Anderson et al., 2016), or types of storms in actual data taken from meteorological satellites (Hennon et al., 2015). Amongst others, the emergent discipline of *astroinformatics* (Ball and Brunner, 2010) has greatly benefited from the analysis of astronomical data in multiple projects, providing data analysis at a scale never reached in the past (Lintott et al., 2011; Schwamb et al., 2012; Barnard et al., 2014). Nevertheless, many challenges arise when trying to maximise the benefits from this kind of large-scale analysis, including the best use of expert classifications (Wardlaw et al., 2018) or participants' engagement in this type of voluntary scientific contribution (Sprinks et al., 2017).

Final results of CzS projects, referred from now on as CzS data, have also attracted the attention of data scientists. On the one hand, this research has focused on the mining of CzS data using an *off-line* approach, that is, the study of results once the project has been concluded. Data scientists have tested the capabilities of data mining (DM) and machine learning (ML) techniques (Witten and Frank, 2005), aiming to replicate amateurs' performance (Banerji et al., 2010; Dieleman, Willett, and Dambre, 2015) or training ML algorithms for a certain problem (Shamir et al., 2014; Beaumont et al., 2014). On the other hand, ML resources are also being used for optimising amateurs' endeavours

[1] https://cosmoquest.org
[2] https://fromthepage.com
[3] https://siarchives.si.edu/collections/sirissic14645
[4] https://www.citizenscience.gov/
[5] https://scistarter.org/

through the course of a project, following an *on-line* approach. This framework encompasses progressive training of new participants as they acquire experience in the problem, or the interaction between a ML classifier and new labelled data as it is generated during the project's run (Zevin et al., 2017).

Despite this, CzS still causes scepticism within the research community (Show, 2015). Even though it offers opportunities for research not possible to accomplish by experts alone, it is not universally accepted as a valid research method (Bonney et al., 2014). The reasons lie in the quality of results, which are often questioned due to several drawbacks involving the prevalence of biases, lack of accuracy and expert validation (Kosmala et al., 2016). Amateurs participating in CzS projects exhibit a wide range of skills, but it is not guaranteed that they hold any background in science or research. Moreover, there is always some degree of uncertainty in classification problems, which usually tends to bring additional vagueness in the definition of the classes (type of birds, shape of galaxies, patterns in a graph, etc.). Consequently, depending on the problem and the participants' expertise with the classification task, the confidence on classifications varies and CzS data thus hold an intrinsic uncertainty.

The research presented in this dissertation focuses on the development of novel off-line approaches for CzS data. We first target the prevalent uncertainty in such data to mitigate the lack of confidence in CzS results and maximise the exploitation and outreach of CzS projects. First, a methodology is developed to enhance confidence in amateur classifications. Expert labels are taken as ground truth and leveraged to refine original CzS data. This has not been attempted before within the specialised literature. After this, the transformed data will serve as input to ML classifiers in subsequent stages of the research, implementing a novel approach that learns from expert and amateur classifications in conjunction with unlabelled data.

1.2 Case Study: Classification of Galaxy Images

Classification is one of the core tasks addressed by ML algorithms (Kotsiantis, 2007; Witten and Frank, 2005). A classifier is usually trained to learn patterns from input data and aims to predict the label to be assigned to unseen data instances (Duda, Hart, and Stork, 2001). In image classification, this entails the categorisation of images into two or more classes, be it mutually exclusive classes (multi-class classification, in which only one single label is assigned) or not (multi-label classification, where different classes coexist). These

paradigms are widely implemented in multiple real-world applications such as fingerprint identification (Peralta et al., 2018) or the recognition of facial emotions (Jain, Shamsolmoali, and Sehdev, 2019), and they have also become a useful tool in science and engineering (Lu et al., 2019; Umehara et al., 2019).

The traditional ML approach to image classification requires the extraction of features from an image. Classical learning algorithms (e.g. decision trees, k-nearest neighbours) do not cope well with images directly, which is typically solved by transforming the image pixels into a new feature space by means of feature extraction (FE) techniques (Nixon and Aguado, 2008). In contrast, deep learning (DL) based approaches (Goodfellow, Bengio, and Courville, 2016) need minor or no data preprocessing for the classification of images. More specifically, convolutional neural networks (CNNs) (Bengio, 2009; Gu et al., 2018) provide excellent solutions (Rawat and Wang, 2017), capable of taking a raw image as input and performing an implicit FE process along with the classification in one single step. Recent state-of-the-art CNNs are usually composed of a very large number of layers (Simonyan and Zisserman, 2015) when dealing with challenging image classification problems(He et al., 2016). Alternatively, DL can also be used to extract features of an image by means of autoencoders (AEs) (Charte et al., 2018), which have also been proposed to ease the learning of standard classifiers (Pulgar et al., 2018). Whereas CNNs often need to learn the image features from scratch using a large amount of labelled data, AEs enable the encapsulation of the FE process for a particular problem without any need of labels, which can be advantageous for the classification of big collections of images and the use of other kind of classifiers (Pulgar et al., 2018).

Classification is also one of the core tasks in astronomical research. Expert astronomers have seen their data processing capabilities greatly exceeded with the advent of modern instrumentation (Tyson, 2010), which has led to the rapid development of the novel astroinformatics discipline to help analyse the data provided (Borne, 2010). In most cases, this entails the classification of large collections of astronomical images (Ackermann et al., 2018; González, Muñoz, and Hernández, 2018; Petrillo et al., 2019). Particularly, the morphological classification of galaxy images aims at the categorisation of these objects into two main classes (morphologies), *elliptical* and *spiral* (Hubble, 1926). The morphology is a key indicator for understanding the galaxy inner structure and physical processes, also revealing aspects about the formation and evolution of the universe (Sandage, 2005). Due to the huge amounts of images produced

in modern telescopes (Borne, 2013), astronomers have been drawing upon the general public for this task using the Internet, and taking advantage of the CzS movement (Cohn, 2008; Simpson, Page, and De Roure, 2014). This was first materialised with the release of the *Galaxy Zoo* project (Lintott et al., 2008), which generated the largest manually annotated catalogue of galaxy images to date (Lintott et al., 2011). In its first edition, the Galaxy Zoo 1 (GZ1) project finished classifying nearly one million galaxy images with the help of more than 200 thousand volunteers, and has continued expanding since then in multiple follow-up releases of the project. This project was first focused on the distinction between spiral and elliptical classes, but set a baseline for other classification problems such as the search of galaxy mergers (Darg et al., 2010) or the identification of the handedness of the arms in spiral examples (Longo, 2011).

Nonetheless, the next generation of astronomical surveys that will produce billions of galaxy images is announcing the limitations of the CzS approach (Ivezić et al., 2019). ML methods are needed, pursuing a robust automation of the classification task, and multiple efforts are being developed in this direction (Banerji et al., 2010; Shamir, 2012; Dieleman, Willett, and Dambre, 2015; Zhu et al., 2019). The research presented here ultimately aims to maximise the potential usage of CzS data in conjunction with expert classifications, in the search of an improved automated classification accuracy for the problem of interest.

1.3 Motivation and Objectives

After the success of the GZ1 edition nearly a decade ago, this modern implementation of the CzS was extended to create *The Zooniverse*[6], a platform where amateurs can take part in real research problems across science and humanities. This generalisation of the GZ1 workflow has turned CzS into a data processing method suitable for data-intensive research, able to provide valuable outcomes in fields in which the quantity of data exceeds the analysis capabilities of expert researchers. Despite this, the interest of the CzS paradigm from a DM perspective is twofold. In first place, CzS projects do not provide final classifications for the data but a record of amateur votes between the options offered on the web during the course of the project. Hence, turning CzS data into a reliable dataset is not straightforward, and DM techniques can assist this

[6] http://www.zooniverse.org

process by supplying automated strategies that take advantage of the entire information present in CzS data, including the prevalent uncertainty. In second place, these amateur-based classifications represent a source of labelled examples for the training of ML algorithms that greatly exceeds the existing expert catalogues in several orders of magnitude. Such amount of labelled data suggests a favourable framework for the learning of models whose performance highly relies on the volume of data needed for their training. However, the best approach would take advantage of both amateur and expert classifications in order to exploit the uncertainty in CzS data to leverage as much data as possible. These two processes are represented in Figure 1.1. **Step 1** covers the generation of true labels from amateur votes recorded within CzS data, whereas **Step 2** comprises the use of this data to train a ML classifier. In this framework, the process of going from amateur clicks in CzS projects to a proper automated classification with ML techniques shows a wide margin of improvement that remains unexplored to date. The research presented here addresses both gaps.

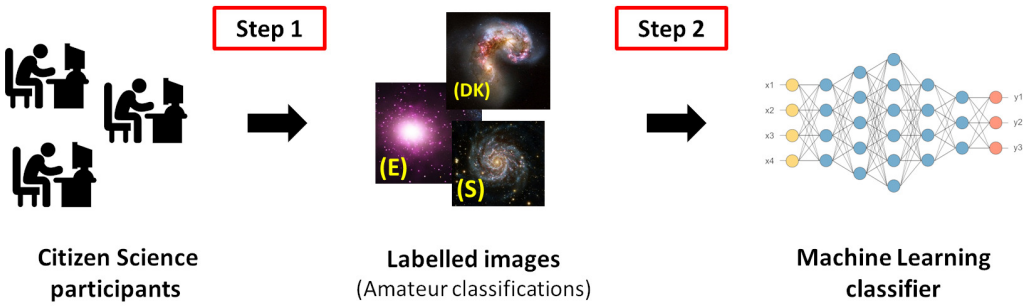


FIGURE 1.1: Two gaps generated at the end of CzS projects.

In galaxy classification, it is a fact that the morphology of a galaxy is a function of the physical size, brightness and colour of the image. There are two main morphological types based on the presence or not of a disk, spiral and elliptical, respectively. However, the multiplicity of hybrid types in between and the wide range of image conditions turn the classification of this sort of images into a very complex task. Certain level of uncertainty is thus unavoidable: there is no universal classification (Lahav et al., 1995). However, this intrinsic limitation of the problem does not prevent a professional astronomer from developing more accurate and reliable classifications. This can be due either to their expertise and knowledge of the subject or the use of complementary sources of relevant information whose analysis is restricted to amateurs. Therefore, besides the evaluation of amateur performance that has been completed to date (Lintott et al., 2008; Kosmala et al., 2016), the consideration of expert classifications is crucial for the improvement of CzS data and the ultimate achievement

of a trustworthy automated classification.

In this framework, we propose the division of the data into different *levels of knowledge*, where the highest uncertainty is assigned to unlabelled raw data and the highest knowledge corresponds to expert labelled data. In an intermediate place between them, CzS data represents an average knowledge that holds some degree of uncertainty. This data framework leads to split the CzS data into two categories naturally overlapped: examples with expert and amateur classification, and examples with amateur classification only. In addition, new technologies allow for the generation of data at a rate never reached before, suggesting the availability of unlabelled data at a scale virtually endless. Figure 1.2 shows the levels of knowledge identified in CzS projects.

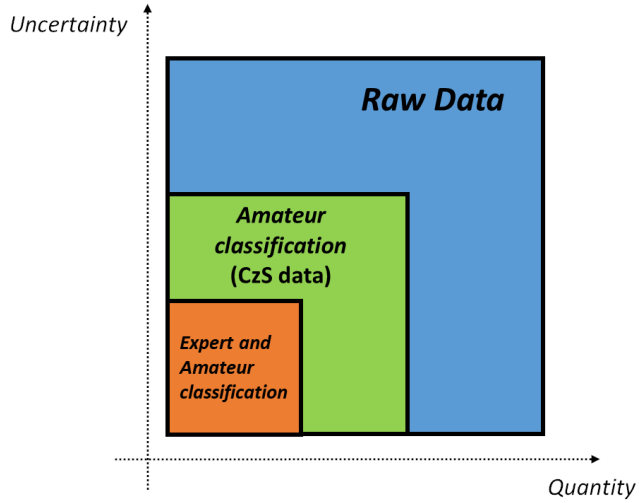


FIGURE 1.2: Levels of knowledge within CzS data.

These are motivations to explore a novel treatment of the CzS data in conjunction with ML approaches. The distinction between different levels of knowledge could help any problem in which the required task is suitable to hide an intrinsic discrepancy between expert and amateur performances. This is the case of numerous projects across the web dealing with classification problems in distinct domains such as the search of exoplanets in *Exoplanet Explorers*[7], the study of the Tanzanian wildlife in *Snapshot Serengeti*[8], or the fight against bacteria resistance to antibiotics in *Bash the Bug*[9].

[7] http://www.zooniverse.org/projects/ianc2/exoplanet-explorers
[8] http://www.snapshotserengeti.org/
[9] http://www.zooniverse.org/projects/mrniaboc/bash-the-bug

This particular framework offers the opportunity to develop a sort of *ladder learning* that takes advantage of all levels of knowledge towards a robust automated classification. Such improvements in the analysis of CzS data can suppose a positive impact in research and promote CzS as a promising resource for data-intensive research. Several research questions derive from the motivations presented above, as follows:

- Is CzS data being adequately exploited in order to obtain the desirable benefits of CzS projects?
- Although expert classifications are often employed to evaluate amateurs' performance in classification tasks, could the available expert knowledge be used to alleviate the prevalent uncertainty in CzS data?
- When CzS data is used in the training of ML classifiers, which DL-based approaches are prone to offer the best characteristics to learn from amateur- and expert-labelled data?
- Is it possible to develop a ML approach to learn from all levels of knowledge that CzS projects offer?

These research questions are summarised in the aim of this thesis: **To develop novel automated classification methods for CzS taking advantage of the three levels of knowledge**. From the motivations referred above, we derive three main objectives that will lead to the subsequent research phases deployed in the following chapters, namely:

- **Objective 1**: Improve the CzS utility with the aid of expert classifications (expert knowledge).
- **Objective 2**: Explore suitable ML approaches for image classification in the framework of CzS.
- **Objective 3**: Develop a unified solution to leverage all levels of knowledge available around CzS projects.

1.4 Contributions

The work presented in this dissertation focuses on the exploration of novel automated approaches for classification problems aided by CzS projects. The research is conducted employing the classification of galaxy images as a case study, which has produced the largest and most consolidate body of research works in the field. This is a consequence of being the pioneering and most

successful classification problem to date that has exploited the modern implementation of citizen science using the power of the Internet. After completing a whole review of the state-of-art about citizen science, the chosen case study, and machine learning approaches for this classification problem, three research stages are completed following a progressive order. The main contributions of this thesis are:

- A novel methodology for the exploitation of the prevalent uncertainty in CzS data. This approach performs an aggregation of information to deal with the uncertainty included in CzS data that increases the final number of instances classified and boosts the classification accuracy. Expert classifications are employed for the first time in order to improve the utility of CzS outcomes.
- A thorough study of ML approaches for image classification is conducted. AEs are proposed for FE of galaxy images, which results in a meaningful reduction of runtime and computational cost in comparison with state-of-the-art approaches. AEs will also signify a crucial element in the development of a novel learning paradigm that brings together all previous insights.
- An approach that combines expert and amateur classifications for the learning of a ML classifier for the first time. This is accomplished by adapting well-established pre-training and fine-tuning schemes for CNNs, showing an enhanced exploitation of these resources in comparison with the isolated use of CzS data and/or expert catalogues alone.
- A novel learning paradigm and an effective solution to learn from all levels of knowledge deployed in the course of CzS projects (i.e. amateur, expert, and unlabelled). This approach, designated as Citizen Science Learning (CzSL), demonstrates better results in comparison with a set of comparative algorithms in distinct classification settings derived from the GZ1 data, including binary, multi-class and imbalanced classification.

The aforementioned contributions are part of or included in the following list of works completed during the Ph.D. studies:

- M. Jiménez, I. Triguero and R. John, A First Approach for Handling Uncertainty in Citizen Science, in *Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.
  This paper was shortlisted for the best conference paper award. The content of this paper is covered in Chapter 3.

- M. Jiménez, I. Triguero and R. John, Handling Uncertainty in Citizen Science Data: Towards an Improved Amateur-based Large-scale Classification, *Information Sciences* 479, pp. 301–320, 2019.
  The content of this paper is covered in Chapter 2 and Chapter 3.
- M. Jiménez, M. Torres Torres, R. John and I. Triguero, A Preliminary Approach for the Exploitation of Citizen Science Data for Fast and Robust Fuzzy k-Nearest Neighbour Classification, in *Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019.
  The content of this paper is presented in Appendix B.
- M. Jiménez, M. Torres Torres, R. John and I. Triguero, Galaxy Image Classification Based on Citizen Science Data: A Comparative Study, *IEEE Access* 8, pp. 47232–47246, 2020.
  The content of this paper is included in Chapter 2 and Chapter 4.
- M. Jiménez, M. Torres Torres, and I. Triguero, CzSL: A Novel Learning Paradigm for Citizen Science, paper in preparation.
  These contents are presented in Chapter 5.

1.5 Structure of the Thesis

The thesis is structured in the following chapters and appendices:

- Chapter 1 presents a gentle introduction to the main topic of the research work as well as the case study employed. It also provides the motivations for the research together with a summary of the contributions achieved and an outline of the contents.
- Chapter 2 introduces the necessary background for the appropriate comprehension of the research conducted across the different stages of the thesis. This includes an overview to the CzS approach, an introduction to the galaxy classification problem, and a literature review of ML strategies of interest for enhancing such classification problem in the context of CzS projects.
- Chapter 3 describes the novel methodology developed for obtaining high quality amateur classifications from CzS data. This first stage of the research presents a novel use of expert knowledge, which had been restricted to the validation of CzS results to date. The refinement of CzS data obtained here will be employed in subsequent experimental stages of the thesis.

- Chapter 4 delves into ML approaches for the galaxy classification problem, deploying a comparative study of FE and DL methods and analysing the impact in the classification performance of experimental aspects such as the image size or the presence of colour in the images. AEs are proposed for the FE on this sort of images, and CNNs are proposed for the learning from expert and amateur classifications, showing a novel use of pre-training and fine-tuning techniques.
- Chapter 5 presents the proposed CzSL algorithm that is capable of learning from experts, amateurs, and unlabelled data. This novel learning paradigm is thoroughly tested against a set of comparative approaches defined. Experiments are carried out over three different classification settings derived from the GZ1 project results, including binary, multi-class and imbalanced classification. CzSL encompasses and confirms the benefits of the results achieved in the preceding stages of the research.
- Chapter 6 concludes the main body of research work with a reflection on the contributions provided and a summary of potential lines of future work.

Finally, two appendices include additional material about some research completed aside the main line of work. Appendix A presents an exploratory analysis of the GZ1 data that provides valuable insights about the running of this pioneering CzS project. Appendix B presents a supplementary use of CzS data that exploits the uncertainty in amateur classifications towards a fast automated classification with the fuzzy k-nearest neighbour algorithm.

Chapter 2

Background

2.1 Introduction

The convergence of computing and global communications has produced a society that feeds on information. However, most of this information is maintained in its raw form, *data*, waiting to be mined and leveraged either to make further predictions or comprehend the inner structure that lies behind. From this information, experts are sometimes able to extract valuable knowledge that may cause a positive impact. Therefore, if data is composed of an amount of recorded facts, then information is the set of patterns that underlie within this data, and which utility entirely relies on developing our automated capabilities of analysis.

This thesis delves into methods to disentangle that valuable knowledge from the data, bridging gaps between CzS, astroinformatics, and most suitable ML approaches for the case study covered. In this chapter, the essential background on these areas is provided. First, the CsZ approach is introduced in more depth in Section 2.2, explaining the particular traits of CzS data and the normal operation of CzS projects. After this, we place the galaxy classification problem within the astroinformatics discipline in Section 2.3, and review the primary manual approach to the problem that was greatly enlarged with the release of the GZ1 project in Section 2.4. Then, some fundamentals on ML methods that are thoroughly employed across the core research of the thesis are first presented at a conceptual level in Section 2.5, and later contextualised for the galaxy image classification problem in Section 2.6. Finally, the contents of the chapter are summarised in Section 2.7, where they are linked to the following chapters that deploy the main body of research work.

2.2 Citizen Science: An Overview

CzS has been a common practice for many years. This form of citizen support to science developed by volunteers goes back in time to the eighteenth century. In those days, a few amateurs started making small but significant contributions by reporting observations about meteorology and ornithology (Miller-Rushing, Primack, and Bonney, 2012). Nowadays, the great advances on the Internet and Information Technologies have broadened the ways volunteers can develop these research-related activities, to the point that CzS is being re-discovered by the scientific community as a valuable resource (Silvertown, 2009). An increasing number of projects engage day by day significant numbers of individuals through the Internet in collecting and/or analysing data, with the support of many institutions from research and academia. The Zooniverse[1] initiative is one of the main platforms for CzS project development and management (Simpson, Page, and De Roure, 2014). Currently, Zooniverse hosts more than 80 projects in topics such as space sciences, ecology, medicine and humanities, directing the joint effort of more than a million participants (Fortson et al., 2012). This has led to the publication of more than 250 scholarly articles[2], validating the utility of CzS for today's research.

There is a solid body of works devoted to the study of CzS as a social phenomenon, emphasising different aspects such as motivation of volunteers, challenges towards acquiring a real research status, or its future prospects (Cohn, 2008; Bonney et al., 2009; Silvertown, 2009; Dickinson, Zuckerberg, and Bonter, 2010; Bonney et al., 2014). A shared claim within these works is the latent potential in the crowd as a valuable resource that should not be neglected by the scientific community. Despite this, professional researchers tend to point out two main issues: a generalised lack of accuracy and the proliferation of biases within the data coming from CzS projects (Kosmala et al., 2016). To overcome this, it has become crucial the development of suitable tools for improving data accuracy, control and minimise the impact of biases, and validate final results. These issues have been addressed from several approaches (Alabri and Hunter, 2010; Yu, Wong, and Hutchinson, 2010; Bonter and Cooper, 2012; Bird et al., 2014). However, these works focus on CzS projects in the context of ecological sciences, which aim for data collection from natural environments at a massive scale. They ignore the difficulties encountered when the target is the

[1] http://www.zooniverse.org

[2] A complete list of references can be found at http://www.zooniverse.org/publications.

Image ID	*Votes*	*Class 1*	*Class 2*	$\cdots$	*Don't Know*
0152948451	58	0.310	0.414	$\cdots$	0.052
0152863349	14	0.643	0.214	$\cdots$	0.071
0152878152	33	0.000	1.000	$\cdots$	0.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0152721030	19	0.316	0.263	$\cdots$	0.263

TABLE 2.1: Typical look of the data recorded in the course of a CzS project involving image classification.

processing of data by large amounts of people. This is the case, for example, in projects coping with classification of images.

CzS projects usually involve one particular task around the processing of some sort of raw data. Once the project is released, participants interested in taking part are invited to complete the task, developing genuine data analysis. For a great number of projects, this has consisted of the classification of large collections of images. After some training is provided, amateurs are asked to classify the images displayed in the project website by choosing amongst a set of categories. These categories often hold a set of main classes, which get the major part of the votes and comprise the target of the classification problem. In addition, it is commonly offered a *Don't Know* (DK) category, useful in case no class is clearly distinguishable and that ensures any image gets a vote every time it is shown. When the project is closed down, all clicks conveniently recorded in a database are made available to a team of experts in the problem for their follow-up study. This data normally includes the count of votes for each of the classes offered to participants, and not a final label for each of the objects in the original dataset. Therefore, a suitable analysis of this data is key at this point to extracting good results from the project. However, a thorough study of this problem from the data science perspective remains unexplored. Table 2.1 presents a typical look of a dataset recorded in the course of a CzS project that involves image classification. Rows contain the information for each of the examples: the image ID, total number of votes received by the project participants, and the proportion of votes corresponding to the set of classes in the problem, including the DK votes.

Data science research applied to CzS has mainly been dedicated to the exploitation of the data after the projects are concluded. This kind of off-line approaches enable the mining of the resulting data along with additional information related to amateurs' performance, available expert knowledge on the problem, and other statistics about the running of the project. To date,

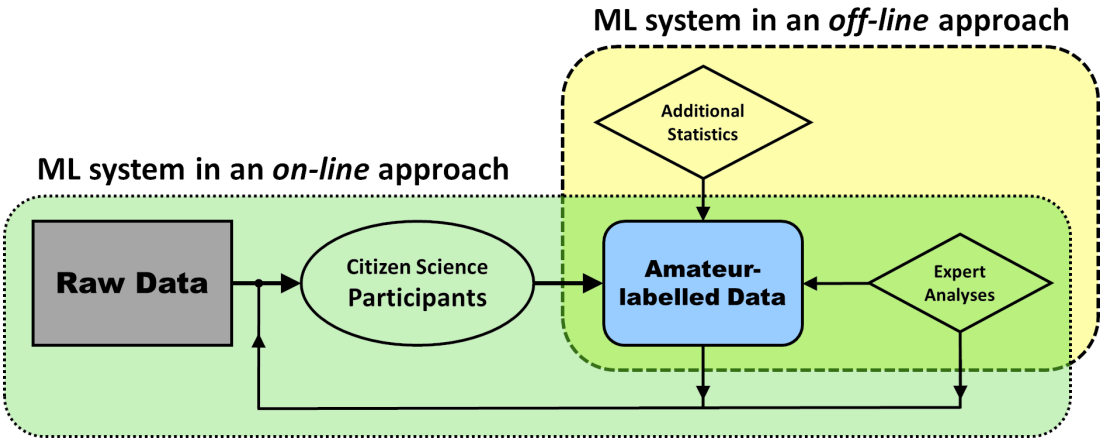


FIGURE 2.1: Potential uses of ML within the CzS workflow.

several works have been focused on emulating amateur classification skills by using ML algorithms, aiming to facilitate an automated analysis of images in diverse contexts. This goal has been achieved by taking as input either a set of features extracted from the image (Banerji et al., 2010; Kuminski et al., 2014), or the whole image within a DL approach (that performs its own feature extraction first) (Dieleman, Willett, and Dambre, 2015). It has been shown that ML classifiers can achieve similar results to those obtained by a group of amateurs, when these algorithms are trained using CzS data (Beaumont et al., 2014). However, these approaches do not address the intrinsic uncertainty in the data, and tend also to replicate the biases generated by amateurs. In addition to off-line approaches, on-line settings have recently been developed for optimising the interaction between humans and machines through the course of the projects (Kamar, Hacker, and Horvitz, 2012; Crowston, Osterlund, and Lee, 2017). These approaches involve ML systems that deal with the training of participants as they acquire expertise in the problem, the management of expert classifications, and the synergy between amateurs, experts and automated classifiers (Zevin et al., 2017). The operation of both kind of approaches is outlined in Figure 2.1, where we highlight the interrelation amongst both potential roles of ML in reinforcing CzS outcomes: on-line approaches take advantage of the synergy between experts and the training of new amateurs, learning from the amateur-labelled data. In contrast, off-line approaches aim to learn from the aggregation of amateur-labelled data, expert analyses and additional information available once the project has been concluded.

In this research, we opt for an off-line approach that targets the inherent uncertainty in projects that tackle classification problems. Our aim is to help experts increase their accuracy and confidence in amateur-labelled data in order

to improve the outcomes of this kind of projects. To this aim, the approach presented in Chapter 3 ensures an aggregation of information concerning this uncertainty that, as a form of data pre-processing, ensures a better use of this data for either research or the training of ML algorithms.

2.3 Astroinformatics and Galaxy Classification

Data is the source of astronomical discovery and knowledge. Since ancient times, astronomers have always relied on observational data for their research, with a minor dependence on experiments. Clearly, they do not build a star or a galaxy in their labs, so astronomical research entirely depends on the collection of evidence from the sky for its subsequent analysis and study. Historically, this astronomical data was obtained in observatories scattered around the world, and results were published in catalogues for the use of others. Today, in contrast, with the spectacular growth in the facilities for gathering data and storing it in large repositories, it becomes possible to perform astronomical experiments by using the data itself (Oswalt and Bond, Springer, 2013). Astronomers systematically study the space making use of large *surveys* (Borne, Taylor and Francis, 2009). These are web-based repositories readily accessible to gather information on visible objects within large regions of the sky, facilitating data access to researchers anywhere for browsing, exploration, mining and discovery (Zhang and Zhao, 2015). Nonetheless, many of the current problems posed by massive astronomical databases need to be tackled from disciplines that are far removed from astronomy, and whose practitioners do not normally interface with this science. Astronomers and astrophysicists generally are not experts in DM, data pre-processing, ML and so forth. Hence, to effectively point to the maximum scientific return from existing and future surveys, the astronomical research has experienced the emergence of a new discipline that embraces this emerging astronomical data science, named as astroinformatics (Brescia and Longo, 2013), which encompasses the organisation, access, integration and mining of data from multiple sources for discovery and decision support. This is accomplished through different areas of application such as data management, DM or data visualisation (Oswalt and Bond, Springer, 2013).

Amongst the multiple issues being faced in astroinformatics nowadays, here we focus on one of the most prominent and long-lasting problems in astronomy: the morphological classification of galaxies (Hubble, Yale University

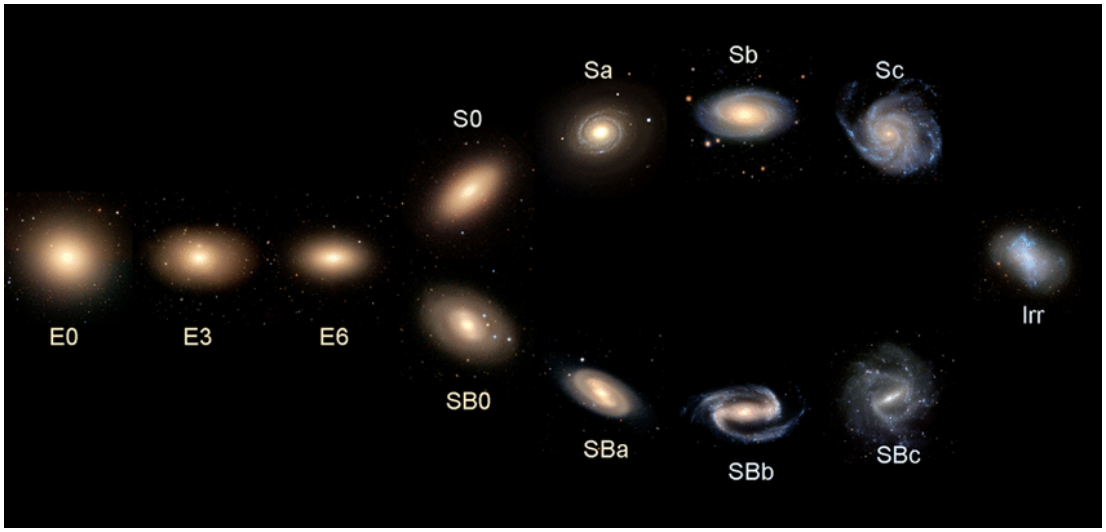


FIGURE 2.2: The Hubble tuning fork. From (Cui et al., 2014).

Press, 1936; Sandage, Carnegie Institute of Washington, 1961; Vaucouleurs et al., Springer-Verlag, 1991; Ball and Brunner, 2010). This entails the classification of these objects according to a blend of the galaxy shape, colour, and texture (Bergh, 1976). The morphology provides a first-order descriptor about the galaxy, which is key for astronomers in the study of fundamental questions about their inner physics (Hart et al., 2018), interactions (Watson et al., 2019), or evolution (González Delgado et al., 2017). There are two main morphological types based on the presence or not of a disk: spiral and elliptical, respectively. This simple distinction was first proposed by E. Hubble nearly a century ago (Hubble, 1926), and despite several reviews (Vaucouleurs, 1959; Bergh, 1976), the well-known as *Hubble tuning fork* remains as a general descriptor of the galaxy population in the near universe, going from ellipticals (E) to spirals, which are split themselves depending on the presence or not of a central bar (S and SB, respectively), as Figure 2.2 shows. Nevertheless, at larger distances, the existence of rare morphologies became important, and a huge variety of galaxy morphologies is displayed. The multiplicity of hybrid types and the limitations shown by the image resolution available at this time turn this classification problem into a very complex task. Unveiling the origins and evolution of this galaxy zoology remains as a cornerstone for understanding the past, present, and future evolution of the universe, making of this classification problem one of the most prevalent research areas in current astronomy and astrophysics.

The first approach to this problem originally consisted in the manual classification carried out by astronomers themselves (Lahav et al., 1995). However,

the increase in the volume of existing data has forced the community to find alternatives to this visual approach. In addition, facilities such as new space telescopes (Tyson, 2010) will make feasible to probe the universe at very high distances, where the discovery of new morphological classes and rare objects of high scientific interest are expected. To face this emerging scenario, the automation of these tasks constitute a huge area of today's astroinformatics research, since the next generation of surveys represent a new challenge for which no valid solution has been provided yet (Huertas-Company, 2015).

2.4 Manual Classification: The *Galaxy Zoo* Project

CzS has constituted a partial solution for the galaxy classification, with the engagement of amateur people from the general public in this kind of data analysis (Bonney et al., 2014). The first edition of the Galaxy Zoo[3] project, the GZ1, first launched in 2007, generalised the manual classification to a great number of amateur people, employing the so-called crowdsourcing and revitalising the classical concept of CzS with an implementation using the Internet. An application was made available on-line, in which members of the general public were able to view images from the Sloan Digital Sky Survey[4] (SDSS) and assign classifications according to an outlined scheme. The project was very successful, and in a period of six months over 100,000 people provided around 40 million classifications for a sample of 893,212 galaxies (Lintott et al., 2008). The final sample was also weighted, trying to discriminate between results from those who carefully think the classification, and those who take less time.

Since then, more launches have been released, involving more than 600,000 people all around the world to classify the full SDSS sample, and other surveys such as CANDELS[5], including more distant galaxies (Masters, 2019). The resulting classifications have been used in a wide range of galaxy evolution studies, giving rise to a collection of scientific studies[6]. Several of the greatest legacies have been the decoupling of colour and morphology (Skibba et al., 2009), the study of bars and bulges in spiral galaxies (Skibba et al., 2012), or the quantification and classification of merging galaxies (Darg et al., 2010).

[3] http://www.galaxyzoo.org

[4] http://www.sdss.org/

[5] http://candels.ucolick.org/.

[6] The complete list of peer-reviewed publications based on the GZ1 project results is available at http://www.zooniverse.org/about/publications.

The obvious advantage of such an approach is that detailed morphologies can be estimated without the effort of inventing new algorithms, while still taking advantage of the human brain inherent capacities to detect complex features in images, and recognise multiple shapes in a single sight. This aspect implies the recognition of different classes and subclasses in the complex morphology schema, which is greatly hard to achieve with an automated classifier. There are some problems though, such as the non-reproducibility, the necessity to train non-expert people and, above all, the evidence that with the next generation of surveys the reasonable limit of applicability of this kind of approaches will be broadly overcome (Kuminski et al., 2014). Nevertheless, the great impact of GZ1 has been the open availability of a huge amount of labelled data, as shows its multiple uses since the first data release in 2011 (Lintott et al., 2011). The fundamental value of this data relies on providing invaluable training sets for future applications of ML to astrophysical problems (peculiar morphologies, study of mergers, new morphological sequences,...), and to the galaxy classification problem as well.

2.4.1 GZ1 Data

In GZ1 project participants were asked to classify galaxy images choosing between one of six categories: *Elliptical*, *Clockwise Spiral*, *Anti-clockwise Spiral*, *Edge-on Spiral*, *Star / Don't Know*, and *Merger*. In this edition, the classification was focused on the distinction between elliptical and spiral morphologies as main classes. At the time the project was closed, each image had received an average number of $\sim$38 independent amateur classifications with a standard deviation of $\sim$14 votes, producing the amateur-labelled data employed in the experiments conducted through the following chapters. Images portrayed on the web held a common scaling of 423×423 pixels in order to provide a similar basis for all classifications (Lintott et al., 2008). One selection of such images is presented in Figure 2.3. The four in the top correspond to elliptical galaxies. The four in the bottom to spiral examples.

After the project finalisation, the GZ1 data was collected in a set of CSV files and published[7]. The so-called GZ1 Table 2 (GZ1-T2) includes the ID of the galaxy in the SDSS database, the location in the sky, total number of votes received by the galaxy, the set of original scores for all categories, and the debiased scores for elliptical and spiral classes for a total of 667,944 galaxy images.

[7] http://data.galaxyzoo.org

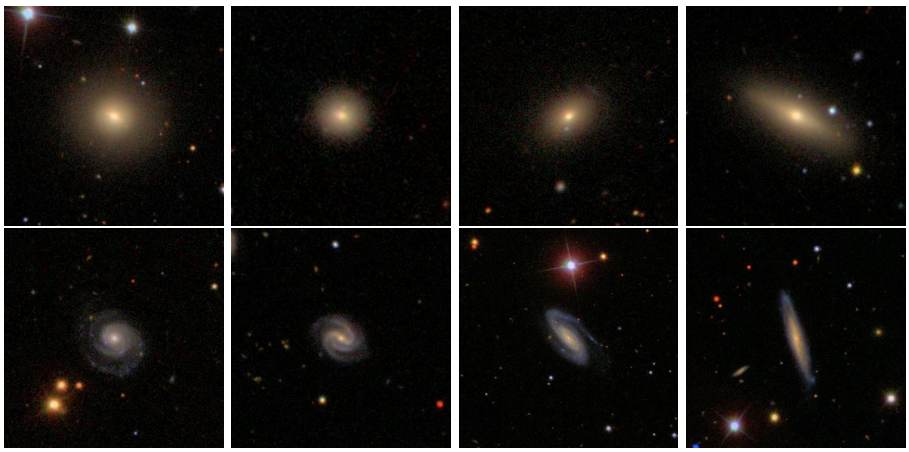


FIGURE 2.3: Selection of original images from the GZ1 project.

To validate amateurs' performance through the GZ1, the developers team originally used two expert catalogues (Lintott et al., 2008). On the one hand, the MOSES catalogue (Schawinski et al., 2007) includes 16,516 galaxies present in the GZ1 subset, all of them classified by a team of professional astronomers as *elliptical*. On the other hand, the Longo catalogue (Longo, 2011) includes 25,190 galaxies all labelled as *spiral* by another set of experts and part of the GZ1 subset as well. When both catalogues are compared, we found an overlap of 141 examples, which were removed for the consistency of results. After this adjustment is made, we take the joint expert catalogue, now composed of 41,424 galaxies from the GZ1-T2 data, which we will refer to as the GZ1 Expert subset (GZ1-E) from now on. This part of the GZ1 data have both expert and amateur classifications, and will play a fundamental role as the available expert knowledge on the problem and to assess the performance of the distinct proposed approaches through the subsequent experiments. A thorough quantitative analysis of the GZ1 data is presented in Appendix A.

2.5 Machine Learning Fundamentals

This section aims to explain, at an introductory level, the ML and DL concepts that are thoroughly employed in the experiments developed in the following chapters. For an easy-to-follow presentation, we first review the materials concerning the basics of ML and classical ML approaches in Section 2.5.1. Then, the DL domain along with the particular DL-based models later used are introduced in Section 2.5.2. Finally, we include a summary of the performance metrics utilised in Section 2.5.3, where they are briefly explained.

2.5.1 Classic Machine Learning

In broad terms, ML can be understood as the study of systems that are capable of learning from examples without having been explicitly programmed to do that (Witten and Frank, 2005). This is usually achieved by constructing a model, which is characterised by a set of parameters, and then finding an optimal set of values for those parameters by minimising an objective function that measures how well (or bad) the model has *learned* from the examples used in its training. The ultimate goal is to generate a model that is able to make predictions from new data, either classifying unseen examples, detecting anomalies, generating new instances, etc. These are known as predictive processes (Mjolsness and DeCoste, 2001), since the goal that lies behind is to disclose the knowledge that relates a set of examples with certain values of a target variable whose prediction is of interest. Depending on the information that is available to carry out the model training, we can distinguish amongst three different types of ML problems:

- **Supervised learning**: in this problem, the values of the target variable[8] are known for the set of training examples. Basically, the goal is to learn the correlation existing between a set of input features and the values taken by the target variable in order to be able to predict the output for unseen examples (for which the target variable is unknown). Depending on the nature of these values, we can differentiate two main problems within supervised ML:
  - **Classification**: the values of the target variable are discrete and a finite number of values (*classes*) are possible (Duda, Hart, and Stork, 2001). For example, the genre of a person (male or female) or the vehicles rolling on a road (lorry, car, motorbike, etc.).
  - **Regression**: the values taken by the target variable are continuous (Cherkassky and Mulier, 2006). For example, the area of a plot, a person's weight, etc.
- **Unsupervised learning**: this time, the values of the target variable are unknown for the training data. Therefore, the goal is in the description of patterns in the data, mainly in the form of relations between the input features. One of the most representative category of unsupervised ML problems is clustering (Jain, Murty, and Flynn, 1999), where the data

[8] The target variable may consist of a single quantity or a set of them, that is, a vector. In the last case, it is understood that we are referring to the target variable as a *target vector*.

is split into several groups (*clusters*) that contain examples as similar as possible in the input feature space.

- **Semi-supervised learning**: this problem lies on the intersection between the two types described above, in the sense that ML algorithms are provided with some examples in which the values of the target variable are known (the fewest) and many others in which those values are unknown. In these problems we can either consider the prediction of the target variable as a classification or a regression problem, and the description of the unlabelled data as a clustering process itself. Therefore, semi-supervised learning is usually treated as an extension of both supervised and unsupervised learning (Zhu and Goldberg, 2009; Triguero, García, and Herrera, 2015).

The model learning process is generally carried out by splitting the data into two subsets: the *training* and *test* data partitions (Witten and Frank, 2005). The former one is the part of the data used to fit the model, while the latter is the sample of data reserved to evaluate the model performance once it has been trained, using an adequate metric. In DL-based models, the training partition is divided itself for *training* and *validation*. This last sample is employed to provide an unbiased evaluation of the model fit after being trained in order to adjust several model hyper-parameters. These data partitions are usually defined in the framework of a *cross-validation* scheme consisting of a certain number of folds. In K-fold cross-validation (Nisbet, Miner, and Elder, 2009), the entire data is first partitioned into K subsets. Then, the whole learning process is repeated K times such that each time one of the K subsets is used as test set, and the other $K-1$ are joint together to form the training set. Generally, the evaluation of errors in the model predictive ability is averaged over the K trials completed.

In this thesis, we mainly focus on supervised classification, although several experiments with AEs also imply unsupervised learning (no labels are used in the training of the AE models) and a semi-supervised based approach is implemented as a comparative method in the last research stage. More specifically, we start by deploying a well-grounded study of image classification, considering the two general approaches for this contemplated in the general ML scene.

The classical ML approach to image classification requires the previous extraction of features from the image. *Feature extraction* (FE) techniques encompass the process of dimensionality reduction by which an initial set of raw data (the image pixels in this case) is reduced to a more manageable and representative

set of features, which does not hopefully require a lot of computing resources to be processed while still maintaining an accurate representation of the original data (Guyon and Elisseefl, 2006). After the FE is completed, the obtained feature set is generally classified using a standard classifier (Wu et al., 2008). However, broad advances in DL through the last decade have produced exceptional alternative solutions to this traditional approach, able to unite both FE and classification processes in a single model. In the following section, such models are examined in more detail, along with an introduction to the DL landscape and approaches employed in the experiments.

2.5.2 Deep Learning Approaches

The so-called *deep learning* is a branch of ML that has experienced an expansive evolution in the last years (Goodfellow, Bengio, and Courville, 2016). The rapid advances in ML, in conjunction with the availability of large datasets for the training of very complex models and the latest progress in hardware capabilities have promoted the development of DL approaches. In general terms, DL is similar to ML at a conceptual level. However, whilst ML covers more classical algorithms such as decision trees or k-nearest neighbours (Witten and Frank, 2005), DL focuses on artificial neural networks (ANNs) that were originally inspired by the actual operation of the human brain (Lecun, Bengio, and Hinton, 2015).

The basic idea behind DL-based approaches is to build models that represent the data at increasing levels of abstraction, by which they can autonomously discover accurate representations from the data itself (Bengio, 2009). DL models generally consist of a variable amount of layers that process the data progressively: more abstract representations of the input data are *learned* as the information gets deeper in the network. Each layer builds upon the representation obtained from the previous layer, typically by computing non-linear transformations of its input. A general example of such a model is the *feed-forward* ANN, where each layer is composed of a variable number of units called *neurons*.

A diagram of a feed-forward ANN is represented in Figure 2.4. This is a network with N layers, with input x_0 and output x_N, the topmost layer. All other layers below are *hidden layers*. Let the vector x_{n-1} be the input to layer n, W_n be the weights matrix, and b_n the vector of biases, then the output of layer n can be expressed as the vector $x_n = f(W_n x_{n-1} + b_n)$, where f is the activation function, responsible for introducing the non-linearity. Common

choices for the activation function are linear rectification [$f(x) = \max(0, x)$], which derived to ReLUs (Nair and Hinton, 2010), or the sigmoidal function [$f(x) = (1 + e^{-x})^{-1}$]. The model parameters are optimised using gradient descent (LeCun et al., 1998) and the back-propagation algorithm (Rumelhart, Hinton, and Williams, 1986), by which each of the layers' weights are iteratively modified until the ANN has been learned.

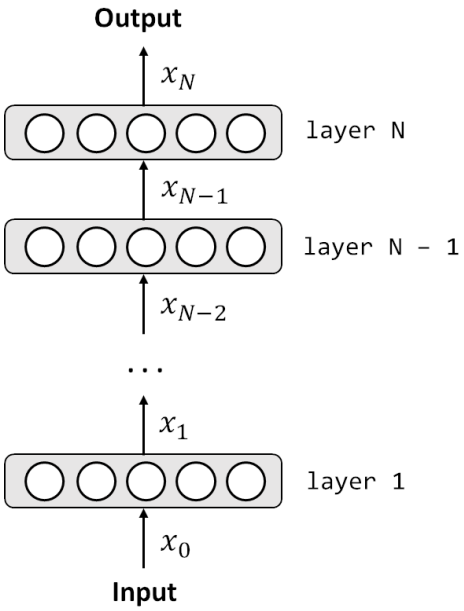


FIGURE 2.4: Schematic representation of a feed-forward ANN.

In this work we focus on the use of a particular type of feed-forward ANNs, the referred as convolutional neural networks (CNNs). They are a subclass of ANNs widely implemented when the input data presents some kind of topological structure, as it occurs in an image. In the following subsection, the special characteristics of these networks are examined.

Convolutional Neural Networks

Basically, CNNs are feed-forward ANNs with many layers that introduce constrained connectivity patterns amongst some of the hidden layers. CNNs generally take advantage of these layer types in combination: *convolution layer* and *pooling layer*. Their main features and operation are briefly explained in the following:

- **Convolution layer**: this layer type takes a stack of feature maps as input (e.g. the colour channels of an image) and convolves each of them with a set of filters to produce another stack of feature maps as output. These

filters are *learned* as a result of the process, similarly to the weights of an ANN.

- **Pooling layer**: this layer shrinks the feature maps previously extracted by a convolution layer. Hence, there is no learning in these layers (no weights or parameters are optimised). Their function is to down sample the feature maps, reducing the number of parameters by taking, for instance, the maximum value (*max pooling*) in a neighbourhood of pixels determined by a patch.

In a typical CNN dedicated to image classification, convolution and pooling layers are commonly combined forming pairs, making up the feature extraction part of the network, as it is schematised in Figure 2.5. These layers are able to progressively detect coarser representations of the input image, modelling higher levels of abstraction as more layers are stacked. When this process is completed, several fully connected (dense) layers, placed at the end of the structure, develop the classification of the features learned by convolution and pooling layers (Gu et al., 2018).

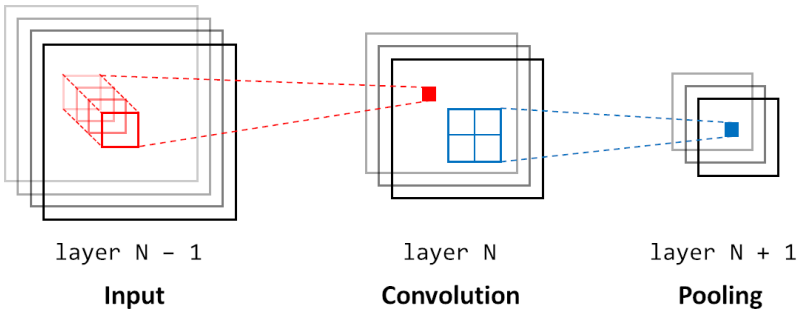


FIGURE 2.5: Representation of a convolution layer followed by a pooling layer.

In the recent literature, very deep networks with numerous layers have been investigated (Szegedy et al., 2015). The network depth is of crucial importance for challenging image classification problems (Simonyan and Zisserman, 2015). Many deep neural network architectures with hundreds of layers such as ResNet (He et al., 2016), ResNext (Xie et al., 2017) or HRnet (Wang et al., 2019a), have provided an outstanding performance in varied image datasets with multitude of different objects (Deng et al., 2009). These complex architectures are usually exploited in a pre-trained fashion (Yu and Seltzer, 2011), which saves computational efforts and allows different domains to take advantage of their prediction capabilities when the scarcity of annotated examples invalidates the training of such models from scratch (Marmanis et al., 2016;

Lopes and Valiati, 2017; Kieffer et al., 2018). This constitutes a form of *transfer learning* (Shin et al., 2016) that enables the re-use of what has been previously learned with a different model. In our research, we investigate the potential of CNNs pre-training to facilitate the learning from different levels of knowledge, which symbolises the core target of the research presented in this dissertation.

Autoencoders

AEs are a common architecture in unsupervised DL, the referred as *encoder-decoder* (Ranzato et al., 2007). In basic terms, an AE is an ANN able to build new encodings for some input by means of a symmetrical arrangement of layers that tries to resemble the input pattern to the output as closely as possible (Charte et al., 2018). The middle layer (symmetry axis of the structure) represents the *encoding*, which is computed after a training process that does not make any use of the data labelling. The general structure of an AE is deployed in Figure 2.6.

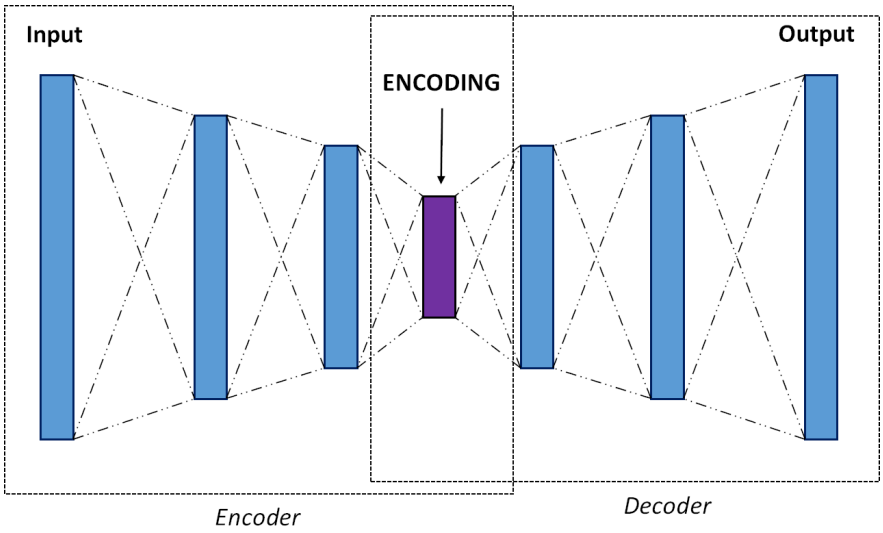


FIGURE 2.6: Schematic view of the general structure of an AE.

AEs' general structure includes an input that is mapped onto the encoding via an encoder. This encoding is, in turn, mapped to the input reconstruction (output) by means of a decoder. The set of layers (blue blocks in the figure) is set out like in a feed-forward ANN. AEs can deploy a variable number of layers, which generally implement a combination of the types introduced above. Since the target is to reproduce the input data on the output, both layers must have the same dimension. However, the hidden layers and the encoding can be higher- or lower-dimensional, depending on the properties desired. As with other ANNs, the set of weights and activation function associated with every

neuron generate the outputs layer by layer. A loss function computes the disagreement between input and reconstruction, which is commonly optimised using the stochastic gradient descent (Robbins and Monro, 1951) in conjunction with the back-propagation algorithm (Rumelhart, Hinton, and Williams, 1986).

AEs have been proposed for diverse tasks related to feature fusion, outlier detection, data compression, and dimensionality reduction to facilitate the learning of canonical classifiers (Pulgar et al., 2018). In this work, we are interested in the test of their capabilities for FE, and their potential to learn features from an image in unsupervised manner.

2.5.3 Performance Metrics

A simple way to evaluate the performance of a classifier is to build the so-called *confusion matrix* (Witten and Frank, 2005). This is defined as a 2×2 matrix that quantifies the correct and wrong predictions with respect to the actual class of the examples evaluated. Each row computes the number of examples in a predicted class, while each column measures the examples in the actual class. In the framework of binary classification (two classes), this is schematised as Figure 2.7 shows. Thus, we can determine the following integers:

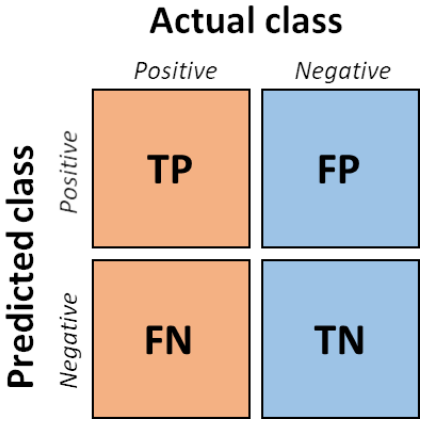


FIGURE 2.7: Confusion matrix diagram (2-class problem).

- True positive (**TP**): Number of examples predicted as positive that are actually positive.
- True negative (**TN**): Number of examples predicted as negative that are actually negative.
- False positive (**FP**): Number of examples predicted as positive that are actually negative.

- False negative (**FN**): Number of examples predicted as negative that are actually positive.

Based on these definitions, we can formally describe the following performance metrics, which are thoroughly used in the experiments:

- **Accuracy**: The accuracy (Acc) quantifies the percentage of examples correctly classified over the total number of classified examples (Witten and Frank, 2005). In mathematical terms:

$$\textbf{Acc}\,(\%) = \frac{\textit{Correct predictions}}{\textit{All predictions}}\,,$$

or equivalently:

$$\textbf{Acc}\,(\%) = \frac{TP + TN}{TP + TN + FP + FN}\,.$$

- **Geometric mean**: The geometric mean (G-mean) is defined in the framework of automatic classification as the geometric mean of the true positive (TP_{rate}) and true negative (TN_{rate}) rates (López et al., 2013). These are the percentage of positive and negative examples correctly classified, respectively. It is one of the more employed performance measures in imbalanced classification scenarios (Triguero et al., 2017). Mathematically:

$$TP_{rate} = \frac{TP}{TP + FN}\,,\; TN_{rate} = \frac{TN}{FP + TN}\,.$$

Hence:

$$\textbf{G-mean} = \sqrt{TP_{rate} \cdot TN_{rate}}\,.$$

2.6 Machine Learning for Galaxy Classification

Astronomers have always been data miners, since the fundamental astronomy tasks resemble some of the classical ML areas (Ball and Brunner, 2010): characterisation of the data from their observations (clustering, class discovery, data pre-processing), labelling of detected objects (classification), or discovery of new events/objects (anomaly detection). Nowadays, ML algorithms are intended to mine pixels and explore catalogues, turning into automatic some of

these activities. As an initial step in the scientific process, classification is one of the main tasks in astronomical research, since it provides a method for organising information in a way that can be used to make further studies and confirm astronomical theories.

This section deploys an overview of the specialised literature concerning the use of ML approaches in galaxy image classification. First, FE-based approaches are reviewed in Section 2.6.1. Then, we examine the use of CNNs and their latest trends in galaxy image classification in Section 2.6.2. To conclude, in Section 2.6.3 AEs are discussed as a promising resource for FE over galaxy images, which is investigated in subsequent chapters of the thesis.

2.6.1 Feature Extraction *plus* Classifier

FE methods used in the classification of galaxy images can be grouped into two main categories: problem specific, which have been especially devised for this particular problem, and general, which cope with image classification regardless of the problem definition. Amongst the first, we review the use of physical parameters extracted from the image, whereas the second category is dominated by the WND-CHARM feature set.

The classification of galaxy images with ML started with the extraction of a reduced number of physical parameters from the image (Thonnat and Berthod, 1984; Lahav et al., 1995). These parameters accounted for properties such as galaxy ellipticity, surface brightness, or concentration. As a form of image features, they were then classified using ANNs (Basheer and Hajmeer, 2000). The feature set was then extended to a greater number of parameters and standardised with the so-called CAS (Concentration-Asymmetry-Smoothness) methods (Conselice, Bershady, and Jangren, 2000; Lotz, Primack, and Madau, 2004). ANNs were generally used for the classification, along with other general classifiers such as random forests and support vector machines (De La Calleja and Fuentes, 2004; Huertas-Company et al., 2008). Nonetheless, with the improvement of computational resources, general purpose feature extractors that compute longer sets of features have outperformed these first attempts.

Although it was originally developed as an image analysis tool for the classification of biological images (Orlov et al., 2008), WND-CHARM represents the state-of-the-art feature extractor for the classification of galaxy images (Shamir, 2012; Shamir, Holincheck, and Wallin, 2013; Kuminski and Shamir, 2016). First, through a FE phase, it computes a set of families of features from the

raw images in a one by one fashion. These are categorised as image content descriptors, image transforms, and compound image transforms (transformations of a previous image transformation), composing a feature vector for the image at hand (Shamir et al., 2008). Depending on the presence or not of colour in the image, two feature sets are available. Then, a feature selection (FS) phase chooses the most informative subset by calculating the Fisher discriminant score of each feature. Finally, the resulting feature vectors are classified using a modified nearest neighbour rule (Cover and Hart, 1967) that weights the distances using the Fisher scores previously computed. Whereas in traditional nearest neighbour only the closest (or k closest) examples determine the class, with WND-CHARM the distances to all training samples of each class are measured. In this work, we only consider the FE and FS phases of the WND-CHARM, the so-called WND-CHARM feature map, for the comparison of this method with the FE performed with AEs.

2.6.2 Convolutional Neural Networks

CNNs currently constitute the state-of-the-art in image classification (Szegedy et al., 2015; He et al., 2016), thanks to the ease of access to big datasets and the improvements in computational resources. Indeed, the galaxy classification problem is a perfect example of this fact, remarkably promoted by the wide availability of astronomical surveys on the web, the so-called *Virtual observatory* (Borne, 2013). CNNs have been widely employed in the classification of galaxy images, showing the limitations of FE methods when the classification is not restricted to the two main morphologies. One of the first successful implementations took place in the framework of a Kaggle competition, the Galaxy Challenge[9], which aimed at classifying a sample of ∼50,000 galaxies from the Galaxy Zoo 2 dataset (Willett et al., 2013). The goal was to predict the participants answers to a set of questions about morphological traits featured by the galaxy. The winner CNN architecture (Dieleman, Willett, and Dambre, 2015) established a benchmark for this problem that has been widely employed thereafter (Huertas-Company et al., 2015; Zhu et al., 2019). This network is composed of four couples of convolution and pooling layers that has obtained great results in the framework of galaxy classification. This particular architecture is presented in Figure 2.8.

However, these models make use of datasets of moderate size to make feasible their training and employ larger resources in terms of computational means

[9] https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge.

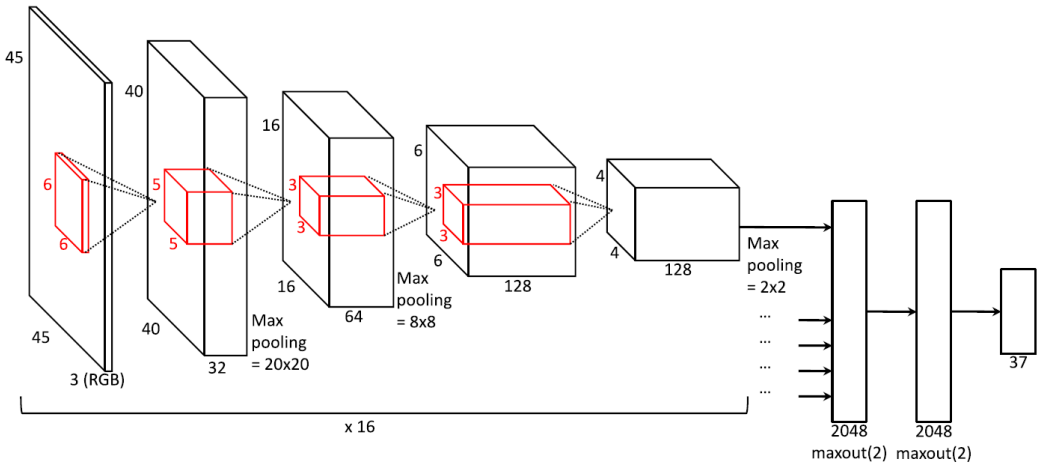


FIGURE 2.8: Example of CNN employed in galaxy classification. From (Dieleman, Willett, and Dambre, 2015).

and runtime. In this thesis, we explore the use of a simpler architecture, and compare it with a state-of-the-art deep CNN. We aim to investigate their differences in performance in terms of running time and the influence of several factors such as the image size, the presence of colour in the image, or the label used (amateur or expert). Then, this simpler architecture is leveraged in conjunction with an AE to learn from amateurs and experts, and unlabelled data.

2.6.3 Autoencoders

In astroinformatics, AEs show a great potential for the processing and storage of large datasets of images. First, their training is unsupervised, which is key due to the scarcity of reliable image labels. Additionally, they enable the encapsulation of the FE phase for the exploration of patterns in the data and large-scale storage and management of astronomical images. However, to the best of our knowledge, a comparative study of their use for FE on galaxy images has not been accomplished. The research developed in this thesis aims at exploring the use of AEs in conjunction with standard classifiers, comparing this approach versus the FE carried out using the state-of-the-art feature extractor for this specific classification problem.

2.7 Summary

This chapter has supplied the necessary background for an adequate comprehension of the conducted research throughout the thesis. The main areas of

study have been presented in more detail, namely: CzS, astroinformatics, and ML. It has been discussed the precedents of CzS, a brief description of the modern CzS implementation on the web has been provided, and the relevance of the galaxy classification problem in current astronomical research has been elucidated. These concepts will clarify the development of the methodology presented in Chapter 3, which aims at the refinement of CzS results prior to their use as input to a ML model in subsequent research stages. Then, the basic concepts around ML and DL techniques further employed through the experiments have been clarified, illustrating these techniques and facilitating their comprehension at a conceptual level. This will help to understand either the comparative study of ML approaches for image classification deployed in Chapter 4 as well as the novel DL-based methodology for the learning from amateur-labelled data, expert classifications, and raw data devised in Chapter 5. As a result, the gaps and motivation identified in the thesis introduction now stand clearer: the research developed in what follows will attempt to contribute with methods in pursuit of valid solutions.

This chapter paves the way for the experiments and discussions deployed in the following chapters. The exploitation of CzS data with suitable ML approaches will be based on the concepts and methods introduced above. As a first step in the research, the next chapter investigates the enhancement of the CzS data with the aid of expert classifications and the aggregation of information about the uncertainty in amateur classifications not exploited before.

Chapter 3

Handling Uncertainty in Citizen Science Data

3.1 Introduction

The use of ML approaches to exploit CzS data is expected to create more accurate models as they are provided with larger sets of high-quality annotated data. Nevertheless, expert analyses completed after the finalisation of CzS projects exhibit several weaknesses found into amateur classifications (Lintott et al., 2011; Willett et al., 2017). Hence, a correct pre-processing of the CzS data in off-line approaches is demanded, pursuing the best exploitation of such datasets either by professional researchers (Kosmala et al., 2016) or in conjunction with the desired automated approaches (Zevin et al., 2017). This will also ensure to maximise the potential utility of CzS projects along with its impact in research and the engagement of the general people in scientific progress (Watson and Floridi, 2018; Trouille, Lintott, and Fortson, 2019).

This chapter covers the first stage of the research. We present a novel methodology to improve the CzS data utility (Jiménez, Triguero, and John, 2019). Prior to its input by a ML classifier, here we explore the aggregation of valuable information included in CzS data that has been neglected within the state-of-the-art approaches, which informs about the uncertainty in amateur classifications. A novel multi-stage method to handle the uncertainty within CzS data is introduced, consisting of two main phases: first, a set of transformations that leverages the uncertainty in amateur classifications that is present in DK

votes and the distribution of the total number of votes awarded by the examples. Then, a hybridisation strategy provides the best aggregation of the transformed data for improving the quality and confidence in the results. The experiments with the GZ1 data are validated using two expert catalogues available (Section 2.4.1), and demonstrate that the proposed data transformations are able to boost the accuracy and classify more images with respect to current approaches with this CzS data.

The chapter is structured as follows. First, we analyse some fundamentals about the management of the uncertainty in Section 3.2. Then, our approach for the improvement of CzS data is explained in Section 3.3. The experiments completed with GZ1 data and expert classifications are presented in Section 3.4, and their results are analysed in Section 3.5. Finally, we summarise the achievements of this first stage in Section 3.6.

3.2 Preliminaries

The study of classification problems with uncertain labels has been developed adopting several approaches (Bouveyron and Girard, 2009; Li et al., 2012). Nonetheless, when the uncertainty comes from a set of independent judgements on the object being classified, fuzzy logic (Zadeh, 1965) provides a very suitable framework for a thorough study of this uncertainty. Fuzzy logic tackles, amongst multiple other subjects, the various forms of uncertainty exhibited in a varied range of problems. As *ambiguity* and *vagueness* (or *fuzziness*) we identify, respectively, the lack of specificity when a set of choices is available, and the difficulty of making sharp or precise judgements in real-world problems (Klir, 1987). These two concepts are deeply intertwined in CzS, since usual activities required to amateurs frequently involve unspecified tasks such as the crisp classification of vague classes, transcription of ambiguous information, or identification of patterns. On the one hand, this kind of tasks bring themselves some level of uncertainty within their definition; on the other hand, we eventually count with a set of independent opinions. These opinions have to be aggregated in the most proper way to get the best results and then maximise the utility of CzS projects.

Multi-criteria and Multi-expert decision making are well-studied categories of problems concerned with finding the best choice when a set of alternatives is available (Tsiporkova and Boeva, 2006; Fedrizzi and Pasi, 2008). Eventually, an aggregation method is needed to combine individual criteria into a final

decision, which is expected to contemplate all individual contributions. In data coming from CzS projects we often encounter this scenario, where there is available a set of opinions. However, while fuzzy models for decision making normally use information from a reduced number of experts on the problem, in the case of CzS data the uncertainty is more extreme for two reasons: firstly, amateurs (in contradistinction to experts) hold a wide range of backgrounds and varying expertise on the task, meaning more vagueness in their opinions; second, the number of judgements that need aggregation is much larger than in other typical group decision making problem. For instance, standard medical decision making has modelled the aggregation of $\sim$50 experts (Garibaldi and Ozen, 2007), whereas a typical CzS project engages up to hundreds of thousands of participants, each one providing several tens of opinions about a set of objects.

To the best of our knowledge, the enhancement of CzS results by using this kind of methods has not been fully addressed yet. A wide range of uncertainties is pervasive either through the problem definition, amateurs' votes, and in the process of aggregating these judgements to reach a final classification. Depending on the nature of the problem addressed, results provided by amateur participants can be aggregated using expert knowledge in the subject to take advantage of all resources available. Pursuing this target, here we propose a way for aggregating additional information about the uncertainty in the voting process that, despite its simplicity, is able to improve current results. We first identify three sources of uncertainty in CzS data that we address separately by a set of transformations that aim to enrich the original data. Then, we employ a hybridisation strategy that explores the most suitable combination of these individual transformations, providing more confident and accurate classifications. We eventually pursue a refinement of results so that they became more trustworthy and maximise the utility and outreach of CzS projects.

3.3 Methodology

In this section, we present our approach for handling the uncertainty spread within CzS data. We consider the whole dataset obtained after the project has finished collecting votes from participants, taking an off-line approach. Firstly, in Section 3.3.1 we introduce basic notation and motivate the adequacy of the method by distinguish three types of uncertainty present in this sort of data. Then, in Section 3.3.2 we present a set of mathematical transformations that

aim to leverage each of these uncertainty types. After this, in Section 3.3.3 we explain a hybridisation strategy that explores the best way to concatenate the three transformation stages in order to get the most convenient aggregation procedure.

3.3.1 Motivation: Three Sources of Uncertainty within CzS Data

In this section, we introduce the basic problem related to the employment of CzS results as well as some notation about CzS data taken in an off-line approach. Then, we provide brief explanations of the different ways the uncertainty is encountered within the data. This makes easier the later comprehension of the method.

This research focuses on CzS projects that signify a valuable aid for some scientific research in solving a certain classification problem. The classification task, which is the core goal of the project, generally involves the identification of a few classes across huge collections of images. However, as it has been explained above, the task is developed with the help of multiple amateur participants. Hence, the output is not a final label for each of the images released during the project running but a variable set of independent amateur votes. Using the data generated by this process, here we propose a better use of these results adopting an off-line approach and exploring how to leverage information about the uncertainty in amateur votes that is able to improve the quality of final classifications.

In order to facilitate the subsequent data analysis, amateur votes are usually converted into scores, which are numbers in the unit interval calculated dividing the number of votes in each category by the total number of votes received by the example. An example of such dataset is shown in Table 2.1. Thus, let $\mathbf{V} = (v_1, v_2, ..., v_C)$ be the vote vector for an instance in the dataset, containing the votes for each of the categories defined in the problem, with C the number of categories and $N = \sum_{i=1}^{C} v_i$ the total number of votes received by that object. We get the score vector $\mathbf{X} = (x_1, x_2, ..., x_C)$ by computing $x_i = \frac{v_i}{N}$, for $i \in \{1, 2, ..., C\}$. The score vector is typically used to obtain a final classification for the object by simply applying a threshold: the category which score is greater or equal than the threshold is assigned to that example. This procedure allows the expert to adjust the confidence in the classification: the higher is the threshold applied, the larger is the consensus amongst amateurs who

labelled that object, and objects holding a greater consensus are expected to be assigned more accurate classifications. However, the selection of the threshold is arbitrary, and even more importantly, it does not take into account the total number of votes, N. On the one hand, all objects which scores do not reach the threshold are left unlabelled (*uncertain*), making the process ineffective as we require higher confidence in the classifications. On the other hand, examples with similar scores may hold a totally divergent number of votes N. So we are neglecting a hidden disparity in confidence.

The main issue derived from the employment of CzS data is the prevalent uncertainty when a group of people provide a set of judgements about the same object. Amateur participants are not expected to agree on their classifications, and final labels depend on how this disagreement is handled. Additionally, we often encounter variability in the total number of votes received by the example, N. Our target here is to refine this amateur-labelled data in order to obtain better classifications and improve both the number of objects classified by applying a threshold and the quality of these classifications. We distinguish three different sources of uncertainty within the data:

- We refer to inherent uncertainty (IU) as the uncertainty due to the variation across amateurs' votes. Given an example displayed on the website, each participant is asked to classify it by clicking in the most appropriate category according to their opinion at the time. Therefore, the final outcome is not a classification but a record of votes for each of the categories, which spread tells us about the IU in that object. In the case all participants have voted for the same category, this class holds a 1.0 score and then the example presents zero IU. Conversely, if the votes are equally split across the categories, with scores equal to $\frac{1.0}{C}$, the IU reaches its highest value accounting for the greatest uncertainty.

- We denote as measured uncertainty (MU) the uncertainty directly quantified by the DK category. This option is normally offered as a form of ensuring every example gets a vote every time it is shown to a participant. This count of votes represents a measure of the uncertainty in the classification: as one object holds a greater number of DK votes, v_{DK}, it is expected to entail more ambiguity in its labelling. Hence, an example with $v_{DK} = 0$ ideally holds zero MU, getting bigger as v_{DK} takes on larger values.

- We refer to level of confidence (LC) as the uncertainty caused by the variability in the total number of votes, N, received by each of the examples

in the dataset. This quantity often follows an uneven distribution, being able to provide an estimation of the confidence in the classifications with respect to the whole set of examples: given an example, the higher is N in comparison with the rest of the set of objects (taking a metric, for instance, the mean number of votes, μ_N), the greater is our confidence in the set of scores for that example. Consequently, for scores similarly spread through the categories of the problem, the LC informs about the more or less confidence we can expect in regard to each particular example. For instance, in a votes' distribution with $\mu_N = 34$ votes, an image that received $N = 74$ votes should be awarded more weight in the final classification with respect to other example holding less confidence, e.g. $N = 5$ votes, and regardless of the particular distribution of N across the classes. In other words: whenever the total number of votes N keeps low, the errors induced by misleading classifications made by participants have a high impact in the final scores. The LC quantifies this imbalance.

These three sources of uncertainty are inevitably intertwined. The MU is part of the IU, which accounts for the spread of the votes through the whole set of categories, including the DK votes. The LC, in turn, is codified in the IU as well, since we can trust a finer variability in the scores given an example as more votes are available, that is, as N reaches greater values. Here we do not aim to study these concepts in depth. We only set a concise conceptual framework for the explanation of the method.

3.3.2 Three Transformations for Data Refinement

In the following subsections, we explain the basis of the proposed method, consisting of three independent mathematical transformations to be applied on the original scores. These transformations are intended to aggregate information about the uncertainties summarised above and not present per se in the set of scores obtained from amateur votes. For the sake of clarity, we label each one with a number tag (not related to any order or importance) and explain their application over the example data presented in Table 2.1. The method takes as input the whole set of vote and score vectors, $\mathbf{V}$ and $\mathbf{X}$, respectively, for each example in the dataset, and provides a modified score vector. Using the new scores, we can apply a threshold to assign a final class to the example, as it was explained above.

Normalisation: Reinforcement of Main Classes

The first transformation ({1}) consists of the normalisation of a subset of the scores. In CzS projects dealing with classification problems, we commonly find that some classes within the options available for voting covers the major part of the examples. These so-called *main classes* hold a greater importance with respect to the rest and represent the target of the problem, that is, to classify the sample according to these few main classes. For example, participants may be asked to recognise either shapes of celestial objects, patterns in a graph, or types of animals in a picture of the savanna, all of these previously defined as canonical types. In addition, other secondary classes are offered, corresponding to minority (less common or rare) classes in the problem or a DK response for the extreme cases in which the amateur is not able to decide. These secondary classes may be of interest in other problems. In this work we are focusing on the improvement in the classification of the main classes. Once the scores are computed, the minority classes tend to obtain negligible scores and therefore do not reach the threshold for the vast majority of the examples. However, these secondary scores contribute to lower the main classes scores, complicating the classification with a threshold. Hence, the normalisation of the main scores is intended to remove the "noise" due to votes received by secondary classes. We also obtain a representation of the IU restricted to the target classes of the problem and independent of the total number of votes received by the example, N: all instances with equal proportion of votes in the main categories are assigned identical scores after the normalisation.

Let $\mathbf{X} = (x_1, x_2, ..., x_C)$ be the whole score vector, we select the main scores, getting a reduced score vector $\widehat{\mathbf{X}} = (x_1, x_2, ..., x_M)$, with M the number of main categories ($M < C$). Once we have $\widehat{\mathbf{X}}$, the normalised score vector $\widehat{\mathbf{Z}} = (z_1, z_2, ..., z_M)$ is obtained performing a usual normalisation as shown in Equation 3.1, for $i \in \{1, 2, ...M\}$.

$$z_i = \frac{x_i}{\sum x_i} \tag{3.1}$$

The normalisation of the main scores ensures that $\sum_{i=1}^{M} z_i = 1$ for every example. This develops as well a cleaning of the main scores for a later aggregation of information about the MU and LC by the two other transformations. Taking as example the data presented in Table 2.1, the normalised scores for this data are shown in Table 3.1, assuming this is a problem with two main classes: *C1*

and *C2*. Each row includes the image identification (***ID***), total number of votes (***N***), original scores and normalised (***Norm.***) scores.

ID	*N*	*C1*	*C2*	*Norm. C1*	*Norm. C2*
0152948451	58	0.414	0.310	0.572	0.428
0152863349	14	0.643	0.214	0.750	0.250
0152878152	33	0.000	1.000	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.316	0.263	0.546	0.454

TABLE 3.1: Normalised scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1.

DK Votes Shift: Evaluation of Measured Uncertainty

The second transformation ({2}) modifies the main scores using the information held in DK votes. It aims to leverage the MU of the example by introducing a shift that favours one particular class and penalises the rest. In projects dealing with classification, we usually find an asymmetry in the main classes: one class is harder to identify than the rest. This occurs, for example, when the overall quality of the images is deficient because multiple factors (images of natural environments affected by weather conditions, space images that depend on the distance, etc.), or biases emerge in amateurs' skills (for instance, due to a unequal number of examples displayed for each of the classes, so more repeated classes become easily recognisable). In this case, the *Don't Know* category quantifies to some extent this uncertainty, when the number of DK votes keeps low with respect to the total amount of votes allocated to the main classes. The shift targets this imbalance: adds confidence to examples with moderate MU promoting the class selected as *favoured* class in the problem, which keeps fixed through the experiments. However, it vanishes for instances with high number of DK votes and prone to hold high IU and therefore be uncertain.

For the calculation, two quantities are incorporated. One is related to the example at hand and another one is taken as a global measure: the number of DK votes for the example, v_{DK}, along with the average number of DK votes across the entire dataset, μ_{DK}. These are combined for the computation of a quantity ϵ, as it is shown in Equation 3.2. Once the shift ϵ is calculated for each of the instances, it is added to the score of the selected as favoured class for the problem at hand. The remaining main scores are equally subtracted the proportional part of ϵ, depending on how many main categories

there are. Being x^* the favoured class in $\widehat{\mathbf{X}} = (x_1, x_2, ..., x^*, ..., x_M)$, the shifted score vector $\widehat{\mathbf{W}} = (w_1, w_2, ..., w_M)$ is computed as it is shown in Equation 3.3, for $i \in \{1, 2, ...M\}$.

$$\epsilon = \frac{\alpha \cdot \mu_{DK}}{\beta + v_{DK}} \tag{3.2}$$

$$\begin{cases} w_i = x_i + \epsilon_i, & \text{for} \quad x_i = x^* \\ w_i = x_i - \frac{\epsilon_i}{M-1}, & \text{in other case} \end{cases} \tag{3.3}$$

Additionally, this transformation uses two parameters that modulate the modification introduced to the scores, which can be adjusted depending on the nature of the problem. The parameter α works as a factor that regulates the influence of the shift over the original scores. The parameter β is added to the count of DK votes, v_{DK}, for the calculation of the shift. The functionality of β is two-fold: it modulates the influence of this transformation for examples with $v_{DK} = 0$, and also prevents the denominator in the definition of ϵ from vanishing to zero in such cases. These two parameters are optimised by testing a set of pair of values and assessing the modified scores with expert classifications as ground truth. Ultimately, we also restrict the range of application of the transformation: we discard examples with maximum IU (all scores equal to $1.0/M$) and zero IU (one category holds 1.0 score and the rest 0.0 score). The optimisation of parameters and range of application is illustrated within the experiments in Section 3.4.

Considering again the example data in Table 2.1, we demonstrate the application of this transformation {2} in Table 3.2. As an example, we use the values $\alpha = 0.1$, $\beta = 1$, and $\mu_{DK} = 1.5$ vote, as adjusted for this particular problem to compute the values of ϵ for each of the instances. Again, we assume there are two main classes, and *C1* is the favoured class. Consequently, the shift values ϵ are added to *C1* scores and subtracted from the *C2* scores. Each row includes the image identification (***ID***), total number of votes (***N***), DK score (***DK***), DK number of votes (v_{DK}), the value of the shift for the instance (ϵ), and original and modified (***S.***) scores. The DK votes are computed by multiplying the DK scores by the total number of votes.

ID	*N*	*DK*	v_{DK}	ϵ	*C1*	*C2*	*S. C1*	*S. C2*
0152948451	58	0.052	3	0.038	0.310	0.414	0.348	0.376
0152863349	14	0.071	1	0.075	0.643	0.214	0.718	0.139
0152878152	33	0.000	0	–	0.000	1.000	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.263	5	0.025	0.316	0.263	0.341	0.238

TABLE 3.2: Shifted scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1.

Votes Boost: Addition of Confidence to Highly-voted Examples

The third transformation ({3}) modifies the main scores employing the information present in the LC. In this case, each score is incremented using the distribution of the number of votes for the class across the entire dataset. Again, we are only interested in the main classes of the problem. However, unlike transformations {1} and {2}, this boost always increases the scores in accordance with the total number of votes received by the class.

The scores are modified as follows. In the first place, the number of votes for each of the main classes, v_i, are expressed in standard units as it is shown in Equation 3.4, taking the mean, μ_i, and standard deviation, σ_i, over the entire dataset for the selected class. After this, the score converted in standard units is weighted by the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. Finally, this result is multiplied by a parameter γ and added to the original score. We obtain by this way the transformed score vector $\widehat{\mathbf{R}} = (r_1, r_2, ..., r_M)$ as it is shown in Equation 3.5, for $i \in \{1, 2, ...M\}$.

$$v_i \quad \rightarrow \quad \tilde{v}_i = \frac{v_i - \mu_i}{\sigma_i} \tag{3.4}$$

$$r_i = x_i + \gamma sigmoid(\tilde{v}_i) \tag{3.5}$$

The parameter γ works as a factor that adjusts the influence of the boost depending on the particularities of the problem. It is optimised using the original scores and contrasting modified scores with expert classifications as ground truth, as it is shown in the experiments in Section 3.4. Extreme instances with maximum and zero IU ($\mathbf{1.0/M}$ and 1.0 scores, respectively) are not modified.

After the transformation is applied, some scores may result in values out of the unit interval, in case their number of votes are located in the right tail of the

votes' distribution. To avoid this, the transformed score vectors are re-scaled to the interval [0,1] after the transformation using the Equation 3.6. The x_{min} and x_{max} values are selected amongst the whole set of modified scores.

$$x_{[0,1]} = \frac{x - x_{max}}{x_{max} - x_{min}} + \mathbf{1.0} \tag{3.6}$$

The application of this transformation {3} over the example data in Table 2.1 is illustrated in Table 3.3. First, we calculate the standard units for the votes in the two classes, taking as example values $\mu_{C1} = \mathbf{20}$ votes, $\mu_{C2} = \mathbf{30}$ votes, $\sigma_{C1} = \mathbf{3}$ votes, and $\sigma_{C2} = \mathbf{5}$ votes. With these values, we take $\gamma = \mathbf{0.5}$ as adjusted for the problem to compute the boost for each example and obtain the modified scores. Since there are a few instances in the data, we omit here the re-scale process after the modification. Once more, we assume two main classes for this example problem. Each row includes the image identification (***ID***), total number of votes (***N***), original scores for the two classes (***C1*** and ***C2***, respectively) original number of votes (v_1 and v_2), number of votes in standard units ($\tilde{v}_1$ and $\tilde{v}_2$) and boosted scores (***B. C1*** and ***B. C2***). Here we skip the re-scale process due to the reduced number of examples.

ID	*N*	*C1*	*C2*	v_1	v_2	$\tilde{v}_1$	$\tilde{v}_2$	*B. C1*	*B. C2*
0152948451	58	0.414	0.310	24	18	1.33	- 2.40	0.809	0.352
0152863349	14	0.643	0.214	9	3	- 3.67	- 5.40	0.655	0.216
0152878152	33	0.000	1.000	0	33	- 6.67	0.60	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.316	0.263	6	5	- 4.67	- 5.00	0.321	0.266

TABLE 3.3: Modified scores obtained from Equation 3.5 for the two main classes ***C1*** and ***C2*** in the example data presented in Table 2.1.

3.3.3 Hybridisation Strategy

In this section, we introduce the final procedure leading to the target of the proposed method: exploring the best aggregation of information about the uncertainty in amateur classifications contained in the data. To this aim, we introduce a hybridisation strategy that operates with the three mathematical transformations explained above.

Each transformation tackles one particular expression of the intrinsic uncertainty present in the amateur-labelled data compiled after the project closure.

As it is discussed in Section 3.3.1, this uncertainty can be split into three distinguishable forms that are, generally speaking, independent of each other. Within a single instance, the DK votes take part in the distribution of votes across the complete set of classes. However, transformation {1} (Normalisation of the main scores) amends this issue: it restricts the IU to the main classes of the problem, neglecting the influence of other secondary classes; it works isolating the *horizontal* spread of votes across the main classes within a single example. Transformation {2} (DK votes shift), in turn, incorporates the MU codified in DK votes. Finally, transformation {3} looks at the distribution of votes through the entire population of examples within a same class and boost examples with high confidence. This *vertical* spread is unrelated to the previous two, and adds valuable information about how trustworthy are the main scores for that instance. The information within the data regarding the three uncertainties described here is represented in Figure 3.1.

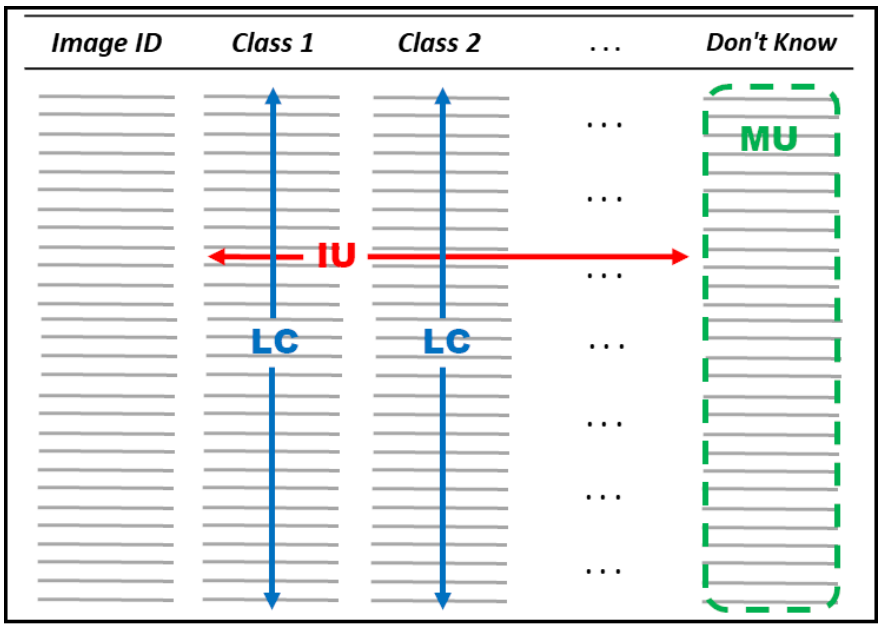


FIGURE 3.1: Three sources of uncertainty present in CzS data.

Hence, a proper blend of the three aggregations is desirable in order to employ the whole information present in the amateur-labelled data. However, the modified scores only focus on one of the uncertainties and ignore the others. Also, depending on the classification problem addressed, these uncertainty types may hold dissimilar relevance within the data. For example, the weight of the DK choice can vary in accordance with the nature of the problem and other factors such as the quality of the images shown to participants. Likewise, when the number of main classes is increased, the distribution of votes may naturally tend to be more uniform. Also, the variability in the LC

depends on the running of the project in case a fixed number of votes for each of the instances is required by project developers.

The method presented here proposes to hybridise the three transformations in all possible combinations to perform a posteriori selection of the best sequence for the problem. The hybridisation performs an independent, sequential, and cumulative application of the single transformations over the reduced score vector, $\widehat{\mathbf{X}}$. As a result, the three transformations are applied following a certain order and taking the modified scores as input of the next transformation. With the three transformations listed above ({1} **Normalisation**, {2} **DK votes shift**, and {3} **Votes boost**), a combinatorial calculation yields we can build a total of $\binom{3}{1} + \binom{3}{2} \cdot \mathbf{2!} + \binom{3}{3} \cdot \mathbf{3!} = \mathbf{15}$ different sequences, which we will denote explicitly from now on by the numerical sequence enclosed by keys[1].

The whole process is developed as depicted in Figure 3.2. Taking the amateur scores coming from CzS data as input (left), the method evaluates all hybrid transformations as they have been explained above (yellow rounded rectangle). Whenever a transformation sequence involves more than one transformation, the modified scores obtained work as input to the next transformation of the sequence. Moreover, the available set of expert classifications (right-bottom) is leveraged in two different ways: on the one hand, it serves as the necessary expert validation for the parameters' optimisation (which are necessary in transformations {2} and {3}). On the other hand, this expert-labelled data enables the ranking of all the transformation sequences obtained in terms of their quality, using an adequate metric. Both processes will be clarified in the experimental section. At the end of the process, the ranking provides a set of improved scores (*Refined Data*) for their later use to obtain final classifications for the objects classified by amateurs.

3.4 Case Study: Improving Galaxy Classification with CzS Data

In this section, we illustrate the proposed method using the GZ1 data. First, in Section 3.4.1 we present the particular features of GZ1 classifications. After this, Section 3.4.2 introduces the followed criteria for the assessment of the

[1]The factorial terms are a result of the no commutability of the single transformations taken alone. For example, the modified scores generated applying the DK-shift followed by a normalisation ({21} sequence) are not equal if the shift is applied *after* the normalisation ({12} sequence).

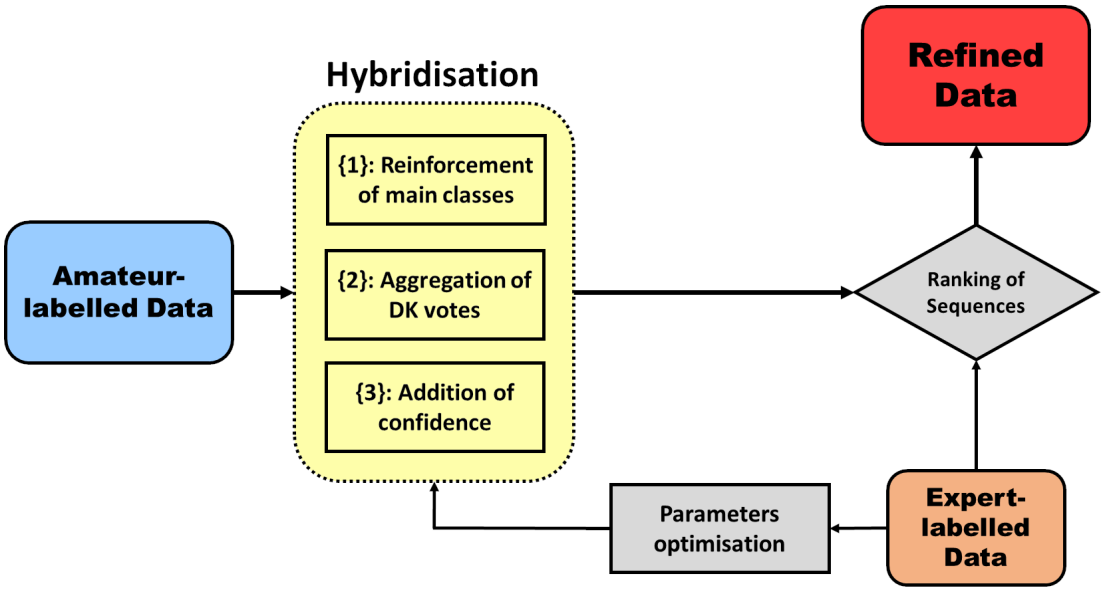


FIGURE 3.2: General workflow of the proposed approach for handling the uncertainty y CzS data.

proposed approach. Then, we describe the experiments implemented for the testing of the method in Section 3.4.3.

3.4.1 Galaxy Zoo Classifications

In GZ1 project participants were asked to classify the galaxy images displayed on the web selecting between one of six categories (Section 2.4.1). The focus of this first edition consisted in the distinction amongst elliptical and spiral morphologies. However, there are multiple factors that complicate this classification problem. Whereas elliptical galaxies present spherical symmetry, spirals hold plane symmetry. Consequently, the orientation of the galaxy plays a fundamental role in the identification of its morphology. In addition, the quality of the image strongly depends on several factors such as the distance to the galaxy, and its physical size and brightness. This brings a huge multiplicity of grades of difficulty that is reflected in the uncertainty within amateur classifications.

After the GZ1 data release (Lintott et al., 2011), the GZ1 team started analysing this data to evaluate the influence of biases in the classification task. This resulted in a thorough study included in Bamford et al. (2009) by which a (manual expert) transformation of the scores obtained from amateurs' votes was developed. Referred as *debiasing* of the scores, it was intended to counter the tendency of classifying blurred images of spiral galaxies as elliptical. As a result, the overall effect was to favour spiral classifications at the expense of

elliptical ones. For this amendment, the three spiral sub-categories were joint, giving a combined spiral score (the addition of the *Clockwise*, *Anti-clockwise*, and *Edge-on* scores), which we will refer to as *Spiral* score henceforth.

In addition to the amateur scores, the GZ1 team provides a set of final classifications known as *GZ1 flags*. These are generated via a process that involves the application of a 0.8 threshold over the debiased scores[2]. However, the debiasing process required an additional parameter[3] that was not available for the whole GZ1 dataset at the time. Therefore, the debiasing and thus the GZ1 flags were only computed for the GZ1-T2 portion of the GZ1 results. An exploratory analysis of the original and debiased scores included in the GZ1 data can be found in Appendix A.

3.4.2 Experiments Evaluation

To validate amateurs' performance through the GZ1, the developers team originally used two expert catalogues (Lintott et al., 2008), which conforms the GZ1-E sample (Section 2.4.1). These two expert catalogues, the MOSES (Schawinski et al., 2007) and Longo (Longo, 2011) expert catalogues, will operate as the ground truth needed for the evaluation and comparison of results, as the part the GZ1 data holding both expert and amateur classifications. Therefore, it is used to validate the GZ1 flags. Also, as the available expert knowledge on the problem at hand, this subset plays a fundamental role in order to assess the performance of our approach through the following experimental trials. From now on, we will take the expert classifications included in the GZ1-E subset as the ground truth for the evaluation of experiments.

Throughout the set of experiments, we use two metrics for the comparison and validation of results specially convenient for the study of this problem: accuracy (Acc) and rejection rate (RR). As the standard classification measure, the Acc computes the proportion of proper classifications with respect to the number of classified examples (Section 2.5.3). Nonetheless, these classifications are obtained applying a threshold over the scores. In case no score reaches the threshold, the example is annotated as *uncertain*. Hence, the RR measures the fraction of uncertain examples. Taking both measures, we perform a preliminary assessment of the GZ1 flags restricted to the GZ1-E subset. This provides

[2]Further details about how the GZ1 flags are produced can be found at `http://data.galaxyzoo.org`.

[3]This is the *redshift* of the galaxy, which works as an indicator of the distance to the object.

a benchmark for the subsequent experiments, and it is shown in Tables 3.4 and 3.5, respectively.

	MOSES	Longo	Joined
Present in GZ1 subset	16,375	25,049	41,424
Correctly flagged	4,181	20,385	24,566
Incorrectly flagged	1,040	26	1,066
Flagged as *uncertain*	11,154	4,638	15,792

TABLE 3.4: Evaluation of GZ1 flags using expert catalogues separately.

Accuracy	0.9584
Rejection Rate	0.3812

TABLE 3.5: Evaluation of GZ1 flags using the joint expert catalogue.

This way of validating results involves looking upon GZ1 as a binary classification problem. Under this view, *Elliptical* and *Spiral* are the main classes, working as negative and positive classes, respectively, since the identification of spiral patterns entail much more detail and observation. *Merger* and *Star / Don't Know* categories are regarded as secondary classes for which we do not count with any form of expert validation. However, we point to the employment of DK votes to improve the quality of the classifications for the two main classes.

3.4.3 Experimental Setting

Here we present and explain the set of experiments executed for the testing of our approach. In the first place, we illustrate the performance of the three transformations taken independently. After this, we test the hybridisation of the transformations over the GZ1-E subset.

In GZ1 there are two sets of main scores: first, we have the original scores directly obtained from the final count of amateur votes, which we will refer to as *raw* scores. Also, we have the *debiased* scores obtained after the debiasing process explained above. These debiased scores serve of a comparison method proposed by the experts in Bamford et al. (2009), as a manual transformation of the raw scores. Here we consider independently both sets of scores for the evaluation of the experiments results.

Similarly to the procedure followed by the GZ1 team, we apply a threshold over the scores in order to assign final classifications to the examples. However, we do not restrict the threshold to one single choice: we explore a series of thresholds in order to get a better intuition about the quality of the classifications provided after certain data transformation has been applied. Here we use six thresholds in the interval [0.5-1.0] taking 0.1 steps, that is, the set (0.5, 0.6, 0.7, 0.8, 0.9, 1.0). These values allow for a well-spread set of cuts that enables a fair comparison between unmodified and transformed scores over a wide range of consensus levels, ranging from 50% consensus (0.5 threshold) to full consensus (1.0 threshold) amongst participants required for the class to be assigned to the object. By this, objects with *Elliptical* or *Spiral* score greater or equal than the threshold being used are labelled as elliptical or spiral, respectively. In any other case, the galaxy is annotated as *uncertain* and counts as not classified. Therefore, each threshold gives one final label for each of the examples in the GZ1-E subset, so we can regard each of the thresholds as a single classifier. Likewise, the application of this series over the scores enables us to check the trade-off between Acc and RR as the IU varies across the sample. That is to say: the higher is the threshold, the larger is the amount of *uncertain* galaxies but more accurate the classifications provided.

By using this set of thresholds we compare the quality of the modified scores obtained after applying either a single transformation, or any hybrid combination of them. This is made according to the evaluation using expert classifications as ground truth explained above. To do this, we represent in an Acc-RR chart the (Acc, RR) points obtained for each of the thresholds in the [0.5-1.0] interval. In addition, for the sake of making the comparison easy and quantitative, along with Acc and RR we consider a third metric: the hypervolume (HV) (Zitzler et al., 2003) subtended by the set of (Acc, RR) points. Since we pursue a two-objective optimisation (we aim to maximise Acc and diminish RR), the HV enables a numerical comparison and ranking of different scores. For its calculation, we take as reference the optimum point (Acc = 1.0, RR = 0.0), that is, the right-bottom corner of the chart. Hence, we are after the minimisation of this measure: the smaller is the HV value, the better is the performance of the transformation.

The last key aspect in the experimental setting is the optimisation of parameters and ranges for transformations {2} and {3} (Equations 3.2 and 3.5). A fixed selection of parameters works well for a single application of these transformations. However, the hybridisation of transformations {2} and {3} in distinct

order shows that pre-fixed values are not appropriate when we mix these two in variable order. To overcome this issue, we conduct an independent optimisation of the parameters and range of application each time the transformation is applied in a sequence, and therefore using the input scores at the time. This means, for instance, that for sequence {213}, the γ parameter and range of application for transformation {3} is optimised using the scores obtained after application of {2} and {1} in that order, as the sequence establishes in this case. The optimisation is always performed in the same manner:

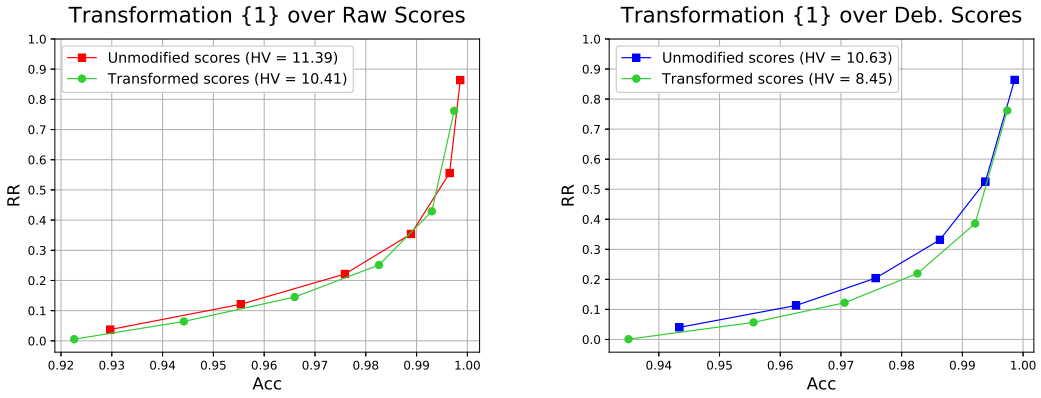
- First, we select the optimal range of application, squeezing the interval (0.5, 1.0) in 0.01 steps up to a minimum amplitude of 0.2. That is to say, we test the intervals (0.5, 1.0), (0.51, 0.99), (0.52, 0.98),... (0.65, 0.85), and choose that one according to the minimal HV value. In this phase, we use the fixed values $\boldsymbol{\alpha = 0.1}$, $\boldsymbol{\beta = 1.0}$ and $\boldsymbol{\gamma = 0.4}$.
- Secondly, we test the parameters in a range of values. For transformation {2}, parameter $\boldsymbol{\alpha}$ is first tested in the interval [0.01, 1.0] taking 0.01 steps, and then it is done the same with $\boldsymbol{\beta}$ in the interval [0.1, 10] taking 0.1 steps; for {3}, parameter γ is computed in the interval [0.01, 1.0] taking 0.01 steps as well. For all cases, the value that minimises the HV is selected.

In the hybridisation of transformations, the parameters' optimisation is developed implementing a 70/30 validation. We split the GZ1-E subset into two parts: the 70% of the sample is used to perform the parameter optimisation as explained above; the remaining 30% is used for the expert validation of the sequences obtained after completing the parameter optimisation process. This validation ensures that the model is not adjusting the values to the same data used for both optimisation of parameters and evaluation of final classifications.

Single Transformations Test

In this first trial of experiments, we test the behaviour of the three proposed transformations one at a time, over both raw and debiased scores. Here we adopt a fixed selection of values for the parameters that are found completing a grid optimisation over a reduced range of values. At this point we do not aim to optimise these values, but to use example values to show how the single transformations work.

Transformation {1} develops a normalisation of the main categories scores. Being $\boldsymbol{M = 2}$ for GZ1, with *Elliptical* and *Spiral* the two main categories, we take

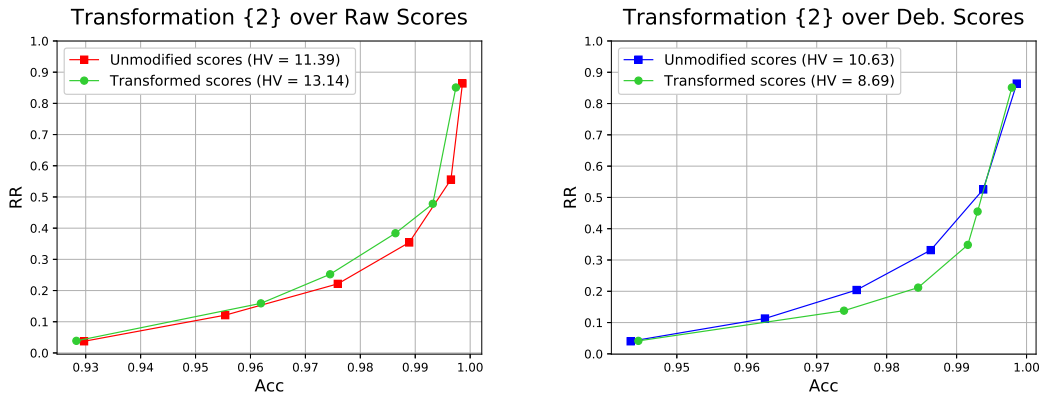


(A) (B)

FIGURE 3.3: Test of transformation {1} over raw (A) and debiased (B) scores.

the reduced score vector $\widehat{\mathbf{X}} = (x_{El}, x_{Sp})$ for each of the examples in the GZ1-E subset and calculate the normalised score vector $\widehat{\mathbf{Z}} = (z_{El}, z_{Sp})$ as expressed in Equation 3.1. For both raw and debiased scores, the application of the set of thresholds brings the charts shown in Figure 3.3. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. The HV values indicated in the legend are multiplied by a factor $\mathbf{10^3}$ for an easier comparison of the quantities.

Transformation {2} introduces a shift into the main scores and uses the DK votes present in the data. For this classification problem, *Elliptical* is selected as favoured class. The reason for this is simple: blurred images of spiral galaxies tend to look as ellipticals. Hence, the MU results in a direct quantification of this bias in favour of the *Spiral* class, helping to alleviate this imbalance present in amateur classifications. To better illustrate the adequacy of this procedure, we firstly check the distribution of this count of votes through the whole population of examples in the GZ1 data. Figure 3.5a shows the distribution of DK votes across the GZ1 subset. The average number of DK votes is $\mu_{DK} = \mathbf{2.82}$ votes, with a standard deviation of $\sigma_{DK} = \mathbf{3.55}$ votes. The maximum value is $n_{DK_{Max}} = \mathbf{81}$ votes, and there are 153,983 examples ($\sim$23% of the GZ1 subset) for which $n_{DK} = \mathbf{0}$ votes. As before with {1}, we take the reduced score vector $\widehat{\mathbf{X}} = (x_{El}, x_{Sp})$ for each of the examples in the GZ1-E subset and compute the shifted score vector $\widehat{\mathbf{W}} = (w_{El}, w_{Sp})$ as indicated in Equations 3.2 and 3.3. For both raw and debiased scores, the application of the set of thresholds brings the charts shown in Figure 3.4. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. The HV values indicated in the legend are multiplied by a factor $\mathbf{10^3}$ for an



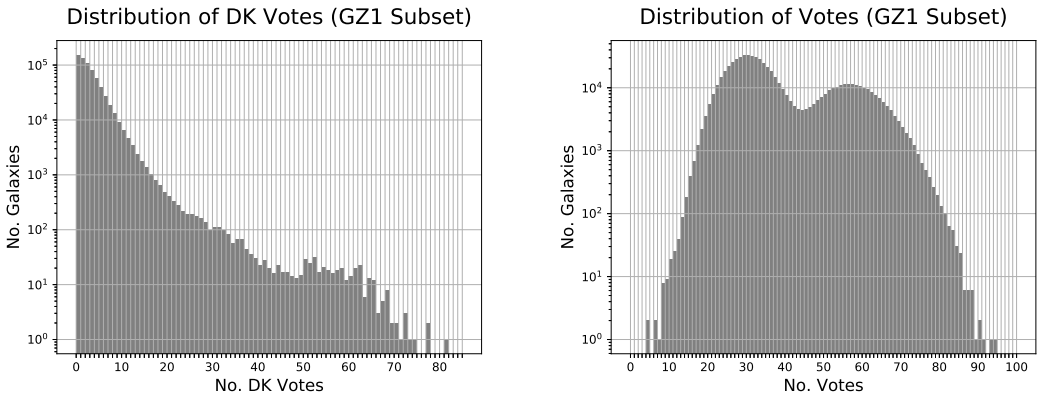
(A) (B)

FIGURE 3.4: Test of transformation {2} over raw (A) and debiased (B) scores.

easier comparison of the quantities. We adopt the values $\alpha = \mathbf{0.05}$ votes and $\beta = \mathbf{1.0}$ votes for calculating the shift (Equation 3.2), which we select after testing several pairs of values and minimising the HV using expert classifications to compare between original and modified scores. In the same manner, we restrict the range of application to the interval (0.6, 0.9). This means that any object with scores out of this interval is not modified.

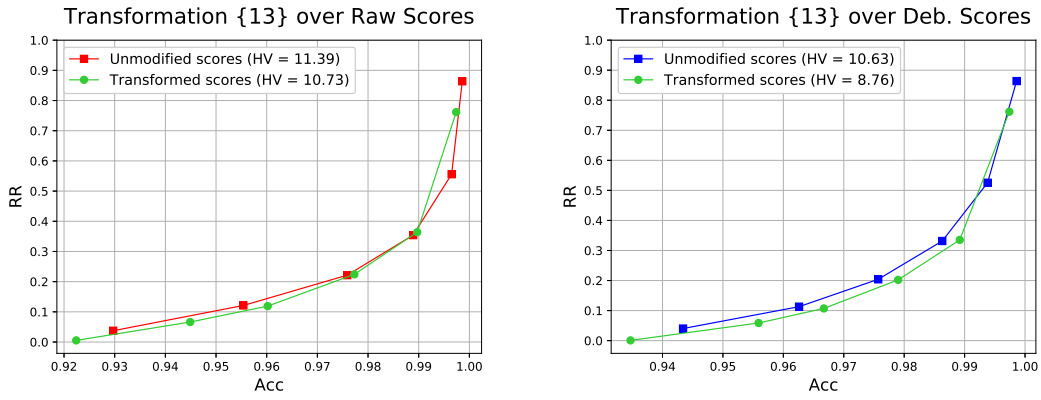
Transformation {3} converts the main categories scores, leveraging the distribution of votes in the category through the whole set of instances. This conversion intends to aggregate information codified in the count of votes, so that examples with similar scores but different number of votes can be disentangled for the labelling. As we did with {2}, we first check the distribution of the votes across the population and the two main categories in GZ1. Figure 3.5b presents the distribution of the total number of votes in the GZ1 subset. The average number of votes is $\mu_N = \mathbf{38.76}$ votes, with a standard deviation of $\sigma_N = \mathbf{13.83}$ votes. The maximum value encountered is $N_{Max} = \mathbf{94}$ and the minimum $N_{Min} = \mathbf{4}$ votes. For this case, transformation {3} shows a meaningful performance taking as input the normalised scores. Consequently, here we consider the hybrid sequence {13}: we take the normalised score vector $\widehat{\mathbf{Z}} = (z_{El}, z_{Sp})$ for each of the examples in the GZ1-E subset and compute the transformed score vector $\widehat{\mathbf{R}} = (r_{El}, r_{Sp})$ as it is shown in Equation 3.5. For both raw and debiased scores, the use of the same series of thresholds results in the charts shown in Figure 3.6. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. As in previous tests, the HV values indicated in the legend are multiplied by a factor $\mathbf{10^3}$ for an easier number handling. Here, we adopt the value $\gamma = \mathbf{0.4}$

(Equation 3.5), which we find after optimising the parameter by testing a range of values: we adopt the value that minimises the HV comparing with original scores and using expert classifications as ground truth. Following the same procedure, we also restrict the range of application to the interval (0.6, 0.9).



(A) (B)

FIGURE 3.5: Distribution of DK votes (A) and total number of votes (B) across the GZ1 subset in logarithmic scale.



(A) (B)

FIGURE 3.6: Test of sequence {13} over raw (A) and debiased (B) scores.

Hybridisation of Transformations

After the testing of the single transformations presented through the previous section, in the following we explain the hybridisation of transformations. In order to extract and combine all information present in GZ1 data, here we propose one hybridisation strategy in two steps: (1) first, we concatenate the three transformations in all their possible combinations; (2) second, we rank the resulting scores according to the HV metric and using the expert classifications.

This procedure ensures a proper blend of the transformations in order to aggregate both the information held in DK and total number of votes. We employ the notation $\{xyz\}$ meaning that transformation $\{x\}$ is applied to the scores, then the output scores are used as input to transformation $\{y\}$, and after this, in turn, the result is taken as input to transformation $\{z\}$. When three initial transformations, we can build a total of 15 different sequences: {1}, {2}, {3}, {12}, {13}, {23}, {21}, {31}, {32}, {123}, {132}, {213}, {231}, {312} and {321}. In addition, since in GZ1 there are two primary scores, raw and debiased, we take both score types available and compute the whole set of transformation sequences over them. Hence, this hybridisation provides a total of 30 different sets of (Acc, RR) points to compare, after validating each of the final scores obtained with expert classifications.

As a preliminary trial, we compute and rank the transformation sequences taking the same parameters values used in the previous section. We take $\boldsymbol{\alpha} = \mathbf{0.05}$ votes, $\boldsymbol{\beta} = \mathbf{1.0}$ votes, and $\boldsymbol{\gamma} = \mathbf{0.4}$, and restrict the application range to the interval (0.6, 0.9). Figure 3.7 shows this ranking of transformation sequences. Following the parameter optimisation and the 70/30 validation, we complete a second trial computing the same set of hybrid transformations. Figure 3.8 shows the ranking obtained with this validation for parameter optimisation. In both charts, the bar height represents the HV value of the sequence indicated in the base, with R standing for raw scores, and D for debiased scores. The red horizontal line represents the HV value for initial raw scores, and the blue horizontal line the same for debiased scores. Pursuing a fairer evaluation of the results, in the second ranking these benchmarks are calculated considering the 30% of the data employed for the expert validation of the sequences. All HV values are augmented by a $\mathbf{10^3}$ factor.

3.5 Discussion of Results

We have completed two sets of experiments for the testing of the method. Although the final goal is to obtain the best global transformation to be chosen amongst the set of hybrid sequences for the problem studied, the testing of the transformations alone illustrate how the method works. In broad terms, both experimental trials bring better trade-offs between Acc and RR with respect to the GZ1 benchmark (Table 3.5). Also, classifications provided by application of the proposed set of thresholds generally outperform the marks obtained by considering the original scores without modification. In the following, we

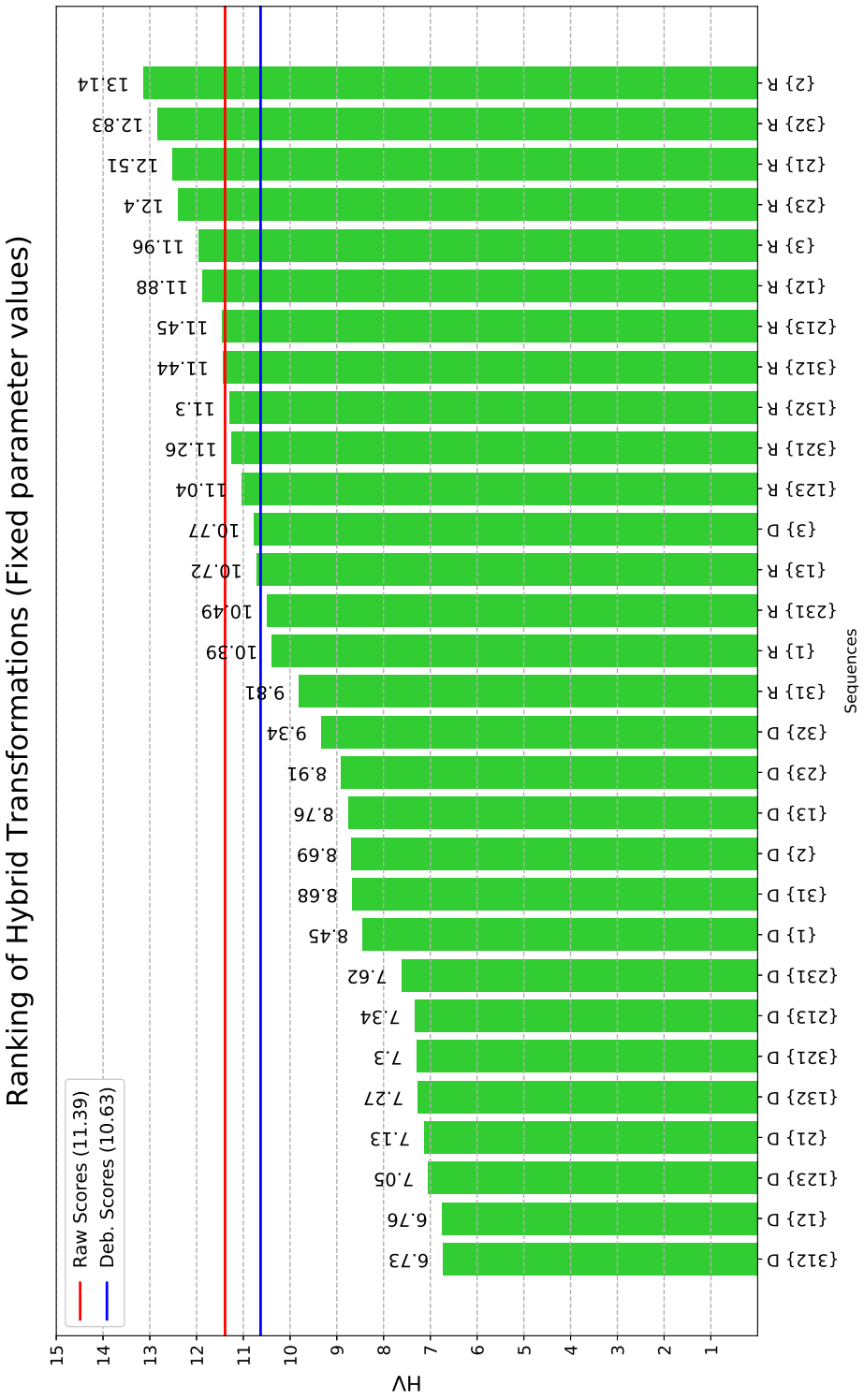


FIGURE 3.7: Ranking of hybrid sequences of transformations with fixed values of parameters.

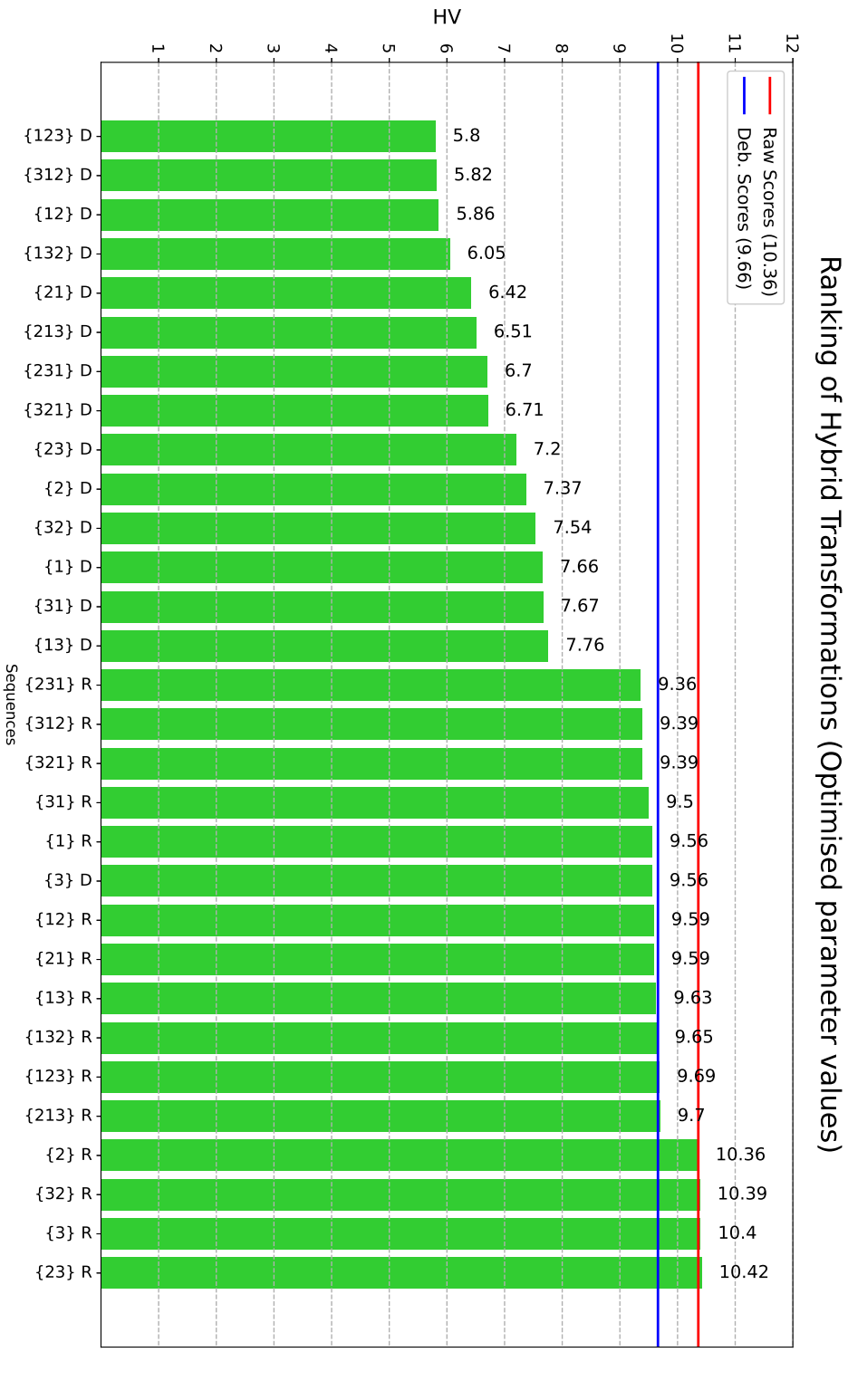


FIGURE 3.8: Ranking of hybrid sequences of transformations implementing parameter optimisation and 70/30 validation of the data. The Raw and Deb. benchmarks have been calculated on the 30% of the data used for expert validation, pursuing a more consistent comparison with the HV values obtained.

highlight the most meaningful results in accordance with the experiments presented above:

- GZ1 provides two sets of scores available, raw and debiased. As it is shown in Figures 3.3, 3.4 and 3.6, debiased scores reach better results compared with raw scores. This trend is maintained in the ranking of transformation sequences (Figures 3.7 and 3.8), for which all hybrid transformations applied over debiased scores outweigh those obtained from raw scores excepting one: the transformation {3} over debiased scores. These results confirm the critical importance of developing debiasing procedures depending on the problem, which also enables here a comparison between our proposal and a comparable method developed by experts in the field.

- Taken independently, transformation {1} is the only one able to provide a simultaneous improvement of both raw and debiased scores. Transformation {2} worsens the Acc-RR marks for raw scores, and transformation {3} does not provide any improvement to original scores, either raw or debiased. However, through the hybridisation process, it can be seen that {3} is key for the enhancement of the amateur-labelled data: it takes part in 7 of the 10 best sequences for both optimised and not-optimised rankings. This result shows a variety of transformations behaviour, justifying the hybridisation strategy in order to get the most convenient merge of information for the problem being studied.

- The optimisation of parameters provides a substantial reduction in the HV values, and thus considerably improves the performance of the transformation sequences (more galaxies classified with higher accuracy). However, the benchmarks also experience a decrease of $\sim$1.0 point when de test data is restricted to the 30% (from 11.39 and 10.63 with the 100% to 10.36 and 9.66 considering the 30% of the data, for raw and debiased scores, respectively). Hence, both set of HV results are not fairly comparable, which does not exclude that, in general terms, a greater number of transformation sequences surpass the debiased benchmark (the most restrictive of both) when the parameter optimisation is applied. This indicates the adequacy of such procedure. Nonetheless, finer optimisation strategies will have to be explored in future work.

These results confirm the potentiality behind this approach, as able to find an adequate adjustment for the aggregation of information about the uncertainty present in the data, taking the form of either MU or LC, and hidden in the DK

votes and distribution of votes through the main categories, respectively. This depends on the choice of metrics for the evaluation of results, and different metrics could lead to different optimal solutions. However, the results presented here ensure a wide margin of improvement using the proposed method, considering the state-of-art of the problem that is represented by the debiased scores computed by experts.

3.6 Summary

In this chapter, we have proposed a novel approach for a better employment of the data generated in the course of CzS projects that deal with classification problems. The main achievement of this approach is to be able to aggregate information about different types of uncertainty present in this sort of data: (1) inherent uncertainty, due to the lack of consensus amongst participants that annotate a same example; (2) uncertainty quantified by participants themselves and included as part of the data; and (3) the uncertainty codified in the distribution of votes through the whole dataset for the main classes of the problem. Using this information, our method has proposed three mathematical transformations that modify the original scores and a hybridisation of them that provides the best combined application in accordance with available expert classifications for the problem. To test our approach, we have analysed as case study one of the most representative CzS projects to date, the GZ1 project. We have presented two sets of experiments: the first one addresses the transformations alone, showing their performance in classifications generated using a threshold over the modified scores; the second implements the hybridisation of the three transformations, demonstrating the advantage of this procedure in order to explore the most adequate blending of them depending on the problem at hand.

As a result, the method has proven to enhance final classifications accuracy and diminish the amount of unclassified images, comparing with an existing method and using expert classifications as ground truth. These transformations will serve as a first building block in the following stages of the thesis. Once a better aggregation of information in CzS has been achieved, the next natural step entails the use of this data as input to a ML classifier. In the following chapter, we first compare automated approaches for galaxy image classification, and then explore a new setting to leverage CzS participants' and experts' expertise in the classification task. Eventually, we aim to study

the merging of all information available about the problem, pursuing the best results and utility of CzS outcomes for science and research.

Chapter 4

Galaxy Image Classification Based on Citizen Science Data

4.1 Introduction

The first achievement of this research has enabled a more integrated use of the data produced in the course of CzS projects. The exploration of different types of uncertainty present in CzS data, in conjunction with the utilisation of expert classifications, have led to the improvement of such CzS results in terms of accuracy in the classifications and number of images classified (Jiménez, Triguero, and John, 2019). However, the upcoming generation of specialised equipment announces the production of images at a rate not even feasible for the CzS approach (Tyson, 2019), and the inclusion of ML techniques pursuing an automated classification is thus considered mandatory (Bu et al., 2019; Zorich, Pichara, and Protopapas, 2020). Once the disadvantages posed by the prevalent uncertainty within CzS data have been ameliorated, the next natural step is to investigate the most adequate ML methodologies to maximise the particular traits exhibited by this sort of data.

In this second stage of the research, we shed light on the automated classification of galaxy images (Ball and Brunner, 2010). After improving the quality of CzS data by handling the pervasive uncertainty in such datasets (Chapter 3), we now focus on how to employ these improved amateur-labelled data in conjunction with an automated classifier. To this end, we start exploring the two general ML strategies for image classification, investigating their advantages and disadvantages in a common experimental framework within the galaxy image classification problem. Following the classical approach consisting of FE together with a classifier, we first compare the state-of-the-art feature extractor for this particular problem, the WND-CHARM (Orlov et al., 2008),

with our proposal based on the use of AEs (Charte et al., 2018). Then, we compare the performance of such an approach with an end-to-end classification using CNNs. We propose a simple yet effective CNN architecture and compare it with a deeper model, namely ResNet (He et al., 2016). To better leverage the CzS data, we also investigate a pre-training scheme that exploits both amateur- and expert-labelled data, which will constitute a key component of the learning methodology developed in the next research stage. For the first time, the entire GZ1 data is employed in the experiments, aiming to investigate the scalability of both classification approaches to a larger dataset. The completed experiments also reveal the most convenient automated approaches worth considering in the next research stage, and highlight the use of AEs as a promising approach to learn from unlabelled data.

This chapter is structured as follows. First, in Section 4.2 we present our proposed architectures of AEs and CNN for the FE and classification of galaxy images, respectively, including a brief explanation of the proposed pre-training of CNNs employing CzS data. After this, Section 4.3 details the experimental setup established in the experiments carried out. Section 4.4 provides the results obtained along with their discussion, and finally Section 4.5 summarises the achievements of this research stage.

4.2 Methodology: Galaxy Image Classification with Deep Learning

This section introduces the proposed AE and CNN architectures for developing the comparative study of ML approaches for galaxy image classification. First, in Section 4.2.1 we motivate the use of AEs and describe the two proposed AE models employed in the comparative study. Then, the CNN architecture used through the experiments is presented in Section 4.2.2. Finally, we explain our approach to make use of both amateur and expert labels using the usual pre-training and fine-tuning of CNNs in Section 4.2.3.

4.2.1 Autoencoders for Feature Extraction on Galaxy Images

In astronomy, in which the generation of data is continuous and a flexible processing of images is desirable (Tyson, 2010), AEs show a great potential. First, their training is unsupervised, which is key in a problem characterised by a restricted availability of reliable image labels. In addition, they allow for the

encapsulation of the FE phase (Pulgar et al., 2018), which can be advantageous for the exploration of inner patterns in the data and also enable an easier storage and management of huge collections of images. However, to the best of our knowledge their use for FE on galaxy images remains unexplored.

This comparative study tests two different AE models that showed the best performance amongst a wide set of topologies that were tested. The first one uses fully connected layers (Section 2.5.2) in a more classical approach, while the second implements a CNN-based architecture. These models are based on architectures originally designed for the classification of the MNIST dataset (LeCun et al., 1998).

- The deep AE (DAE) model deploys the architecture of a deep and undercomplete AE, that is, an AE with more than one hidden layer and the encoding having a lower dimensionality than the input (Charte et al., 2018). It holds two fully connected layers between input and encoding. Following this, the reverse structure is deployed from encoding to output layer, thus completing the symmetrical structure as it is shown in Figure 4.1. For this model, the encoding dimension is the same regardless of the input image size: 256 features. The activation function used in all neurons is the rectified linear unit (ReLU) (Nair and Hinton, 2010), except for the last layer (output), which applies the sigmoid function. The DAE architecture is specified in Table 4.1.

Layer type	Size (neurons)	Activation
Fully connected	1024	ReLU
Fully connected	512	ReLU
Fully connected	**256** (encoding)	ReLU
Fully connected	512	ReLU
Fully connected	1024	Sigmoid

TABLE 4.1: Topology of the proposed DAE.

- The convolutional AE (CAE) model uses the mechanism of CNNs for learning the encoding, implementing convolution and pooling layers. The proposed model deploys three pairs of convolution – pooling layers from input to encoding. After the third pooling layer, the resulting tensor is flattened to obtain the encoding, as Figure 4.2 shows. Unlike the DAE model, for CAE the number of features depends on the input image size. The first convolution layer holds 16 kernels and the remaining two hold 8 kernels. The receptive fields are 3×3 pixels size, and pooling layers implement max pooling with 2×2 pixels windows. As in

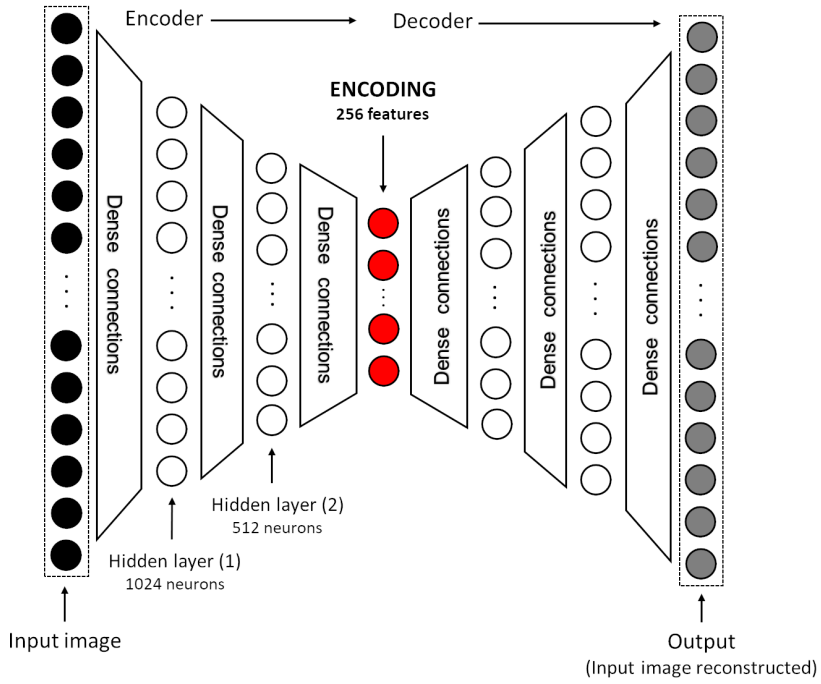


FIGURE 4.1: Architecture of the proposed deep autoencoder (DAE).

the DAE, ReLU activation functions are used along the network except in the output layer, which applies the sigmoid function. The complete specifications are presented in Table 4.2, where only are detailed the layers from input image to encoding. That is to say, the reverse structure is deployed for the decoder, using the sigmoid function in the last layer.

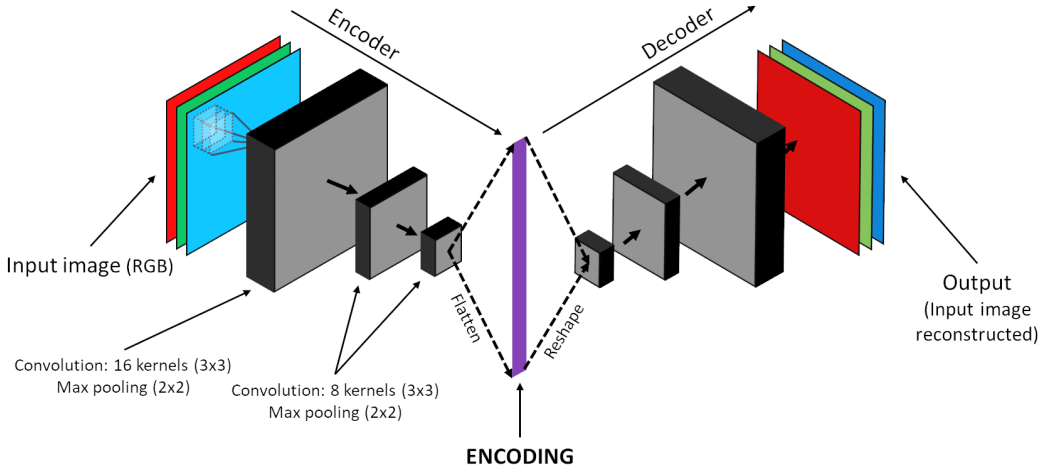


FIGURE 4.2: Architecture of the proposed convolutional autoencoder (CAE).

Layer type	Size	Stride	Activation
Convolution	16 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–

TABLE 4.2: Topology of the proposed CAE.

4.2.2 Convolutional Neural Network for Galaxy Image Classification

Pursuing a fair comparison with the classification using the two AE models introduced above, the CNN architecture proposed here resembles the CAE topology presented in the previous section. Hence, this consists of three consecutive pairs of convolution – pooling layers, which performs the FE phase, followed by three fully connected layers that complete the feature classification. As with the CAE, the network computes 16 feature maps in the first convolution layer, and then 8 in second and third convolution layers. Pooling layers implement the max pooling using 2×2 pixels windows. After this, two layers of 256 and 128 neurons hold dense connections and produce the network output, consisting of two neurons (binary classification). This output layer applies the SoftMax function, which converts the network output into probabilities. After this, these values are rounded to produce the final class labels. ReLU activation functions are used through the whole structure except for the output layer, as it has been pointed out above. The proposed architecture is shown in Figure 4.3, and all specifications presented in Table 4.3.

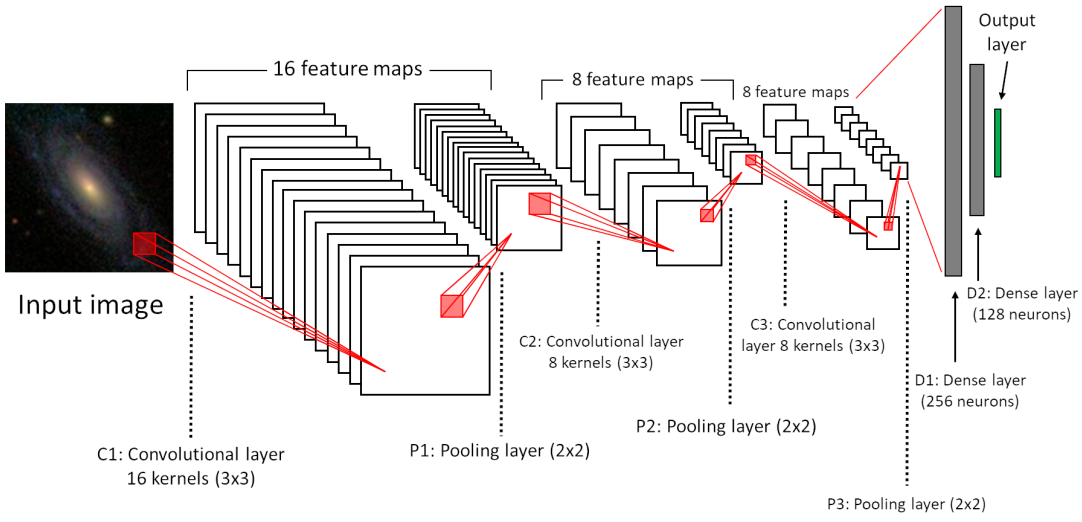


FIGURE 4.3: Architecture of the proposed CNN.

Layer type	Size	Stride	Activation
Convolution	16 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Fully connected	256	–	ReLU
Fully connected	128	–	ReLU
Fully connected	2	–	SoftMax

TABLE 4.3: Topology of the proposed CNN.

Deeper architectures were tested prior to the experiments, implementing up to six pairs of convolution – pooling layers and/or different numbers of feature maps. Nonetheless, the improvement was marginal or even diminished the accuracy in the results. Consequently, we opted to select the CAE's topology presented above and compare this simpler architecture against one of the state-of-the-art models more widely used in computer vision. Particularly, we selected the lighter implementation of ResNet, *ResNet50* (Mahdianpari et al., 2018), which is composed of fifty layers to exploit the residual learning blocks that characterise this DL approach (He et al., 2016).

4.2.3 Pre-training of CNNs with CzS Data

The final stage of our comparative study involves the pre-training and fine-tuning of the proposed CNN and the ResNet using CzS data. By this, we aim to investigate the learning of such models from both amateur and expert labels at the same time, thus leveraging the whole range of knowledge featured by CzS projects.

In the literature, pre-training has been used to alleviate the scarcity of labelled examples and also save the higher computational capacity and time required for the training of deep architectures (Kieffer et al., 2018). Hence, pre-trained networks, which have been previously *fully-trained* using a more generic dataset, are adapted (*fine-tuned*) to the particular classification problem at hand employing more specific data (Marmanis et al., 2016). In practical terms, this process generally entails the top dense layers of the network that performs the classification of the deep features extracted by the convolutional layers. Taking advantage of the set of weights previously learned in the pre-training step, the dense layers are modified and trained to fulfil the requirements of the targeted classification problem (Mahdianpari et al., 2018).

Inspired by the problem of coarse supervision (Taherkhani et al., 2019), in which image labels followed a hierarchical structure (e.g. coarse: big cat, finer: leopard, cheetah, tiger), here we explore a special case of pre-training by taking advantage of CzS data. First, we train the CNN using amateur-labelled data, which is more abundant with respect to expert-labelled data and considered less reliable. Then, we complete the fine-tuning with examples labelled by experts, which we take as ground truth for this problem.

4.3 Experimental Setup

In this section, we explain the experimental setup deployed for the comparative analysis of the classification approaches explored within the galaxy classification problem. First, in Sections 4.3.1 and 4.3.2 we present the selected standard classification algorithms used in conjunction with the FE-based approaches, and the two selected FS methods along with the algorithms' parameters, respectively. Both standard classifiers and FS methods are applied to either AEs or WND-CHARM feature sets. Then, in Section 4.3.3 we provide the experimental details of the DL-based models, which include both AEs, the proposed CNN, and ResNet. Finally, we introduce the GZ1 image data used and detail the experimental evaluation in Section 4.3.4.

4.3.1 Standard Classification Algorithms

The standard classification algorithms selected for the comparative study were made according to their proven accurate behaviour in many real-world applications (Wu et al., 2008). Here we use k-nearest neighbours, random forests, and support vector machines. The basics of each algorithm is introduced in the following and their configurations are summarised in Table 4.4. In these experiments, we are not interested in performing a fine-tuning of these algorithms. Hence, we employ some of the standard configurations for these classifiers as they are found in most settings within the specialised literature (Galar et al., 2011; García et al., 2012; Triguero, García, and Herrera, 2015).

- The k-nearest neighbours (kNN) (Cover and Hart, 1967) is a well-known algorithm in supervised learning. The kNN algorithm uses the whole training set to classify new instances via a similarity function, which is usually defined by a distance on the feature space. First, distances are computed from the new example to the entire training set, and then the k

closest training instances are selected. The final label is chosen in accordance to the prevalent class in the k-subset. Here, we use the Euclidean distance and $\boldsymbol{k = 3}$.

- Random forests (RF) (Breiman, 2001) is a classifier based on the decision tree algorithm (Witten and Frank, 2005). It trains a number n of decision tree classifiers and provides the majority class amongst them as a result. Each sub-tree, referred as an estimator, is trained using a sub-sample of the original training set. Here we run the experiments with $\boldsymbol{n = 100}$. To measure the quality of a split, the Gini impurity is used.
- Support vector machine (SVM) (Burges, 1998) is capable of learning a mapping from the input attributes to the set of classes by means of a higher-dimensional feature space. For this, a kernel function enables the computations of the inner product between two feature vectors. After the instances are mapped to the new feature space, the algorithm looks for the optimal separating hyperplane, that is, the one that maximises the distance to the different classes clusters. Here we train the SVM using the linear kernel.

Algorithm	Parameters
kNN	$\boldsymbol{k = 3}$ Distance: *Euclidean*
RF	$\boldsymbol{n = 100}$ Criterion: *gini*
SVM	Kernel type: *Linear* $\boldsymbol{C = 1.0}$, Tolerance parameter = $\boldsymbol{10^{-4}}$

TABLE 4.4: Parameters for the selected standard classifiers.

4.3.2 Feature Selection Methods

We consider two FS methods through the comparative study, aiming to investigate the impact of FS over the extracted feature sets in terms of classification runtime and accuracy:

- As mentioned in Section 2.6.1, a FS method is proposed as part of the WND-CHARM classifier (Orlov et al., 2008) and is based on a feature ranking involving the use of Fisher discriminant scores as feature weights (Bishop, 2006). First, the weights are calculated for each feature following Equation (4.1):

$$W_f = \frac{\sum_{c=1}^{N}(\overline{T_f} - \overline{T_{f,c}})^2}{\sum_{c=1}^{N}\sigma_{f,c}^2} \cdot \frac{N}{N-1}, \tag{4.1}$$

where W_f is the Fisher score of feature f, N is the number of classes in the problem, $\overline{T_f}$ is the mean of the values of feature f in the entire dataset, $\overline{T_{f,c}}$ is the mean of the values of feature f in the class c, and $\sigma_{f,c}^2$ is the variance of feature f values across all examples of class c. After the features are ranked according to this weighting, the 35% holding a lower Fisher score are rejected, as originally proposed in Orlov et al. (2008). Therefore, this Fisher FS method results in a fixed number of features.

- The second FS method proposed is part of the so-called embedded FS techniques, which highlights as a simple yet fast strategy especially suitable for high-dimensional data (Saeys, Inza, and Larrañaga, 2007). It implements a randomised decision tree with 50 trees to choose the most relevant features. This method also makes use of the labels, searching an optimal subset of features in the combined space of features and hypotheses. As such, the final number of features selected is variable, depending on the data sample considered (Wijnhoven and De With, 2010).

4.3.3 DL-based Models

In this comparative study we pursue a fair comparison between distinct approaches for the classification of galaxy images. Thus, as stated above, the DL-based models used in the experiments share a similar architecture and most parameters. Nonetheless, to compare the proposed CNN against a well-established deep neural network for end-to-end image classification, we introduce the ResNet (He et al., 2016) model as a comparison algorithm. ResNet is an architecture normally trained to distinguish between multiple classes. Given that we focus on a binary classification problem, we use one of the lightest versions of this network, the *ResNet50* (Mahdianpari et al., 2018). AE models, the proposed CNN and ResNet were trained over 100 epochs with a batch size of 256 examples. CAE, CNN and ResNet were optimised with stochastic gradient descent (Wijnhoven and De With, 2010), whereas DAE used the adadelta optimiser (Akila and Anitha, 2019). These parameters were selected according to the configurations proposed in relevant works that implement comparable architectures as the employed in the experiments conducted hereafter (LeCun et al., 1998; Dieleman, Willett, and Dambre, 2015; Charte et al., 2018).

4.3.4 Image Data and Evaluation of Experiments

The image data used in the experiments corresponds to the GZ1-T2 data (Lintott et al., 2011) in conjunction with the expert classifications available for a subset of this data. This GZ1-E sample (Section 2.4.1), used in the previous chapter, comprises an amount of 41,424 examples that were classified as elliptical or spiral by a team of expert astronomers. In this sample, we used the expert classifications available as image labels, which we take as ground truth for the problem. Conversely, the whole GZ1-T2 data provides the record of votes for the options offered on the GZ1 web to project participants (Lintott et al., 2008). Hence, when using the entire data we assigned the majority voted class amongst spiral and elliptical, considering the original amateur classifications. However, this criterion left 8,759 examples for which both scores coincided and thus could not be labelled in this way. We opted to remove these images from the GZ1-T2 data for consistency, then using the remaining 659,185 images. We refer to this sample as GZ1 amateur subset (GZ1-A) from now on. The class distribution of both GZ1-T2 data samples employed in the experiments are shown in Table 4.5.

Class	**GZ1 Expert** (GZ1-E)	**GZ1 Amateur** (GZ1-A)
Elliptical	16,375 (39.5%)	430,562 (65.32%)
Spiral	25,049 (60.5%)	228,623 (34.68%)
Total	41,424	659,185

TABLE 4.5: Data samples taken from the GZ1-T2 dataset used in the experiments.

To carry out the experiments, the entire GZ1-T2 image dataset was primarily downloaded from the SDSS CAS server[1]. In order to establish a fair comparison, we follow the original GZ1 project specifications as detailed in Lintott et al. (2008), taking 423×423 pixels JPEG images centred in the galaxy. The image scale is particular for each image and varies in accordance to the formula $\mathbf{0.024}R_p$ arcsec/pixel, where R_p is the Petrosian radius for the galaxy, that is, a good estimator of its physical size. However, we found that this automatic scaling tends to leave the galaxy isolated in the centre of the image, with a dominance of background pixels and/or other meaningless artefacts around the target object. Therefore, in order to speed up the image processing by both FE methods and the CNNs, we simplified the images' presentation and ended up considering two image sizes through the experiments, aiming to study the influence of the image size and resolution in the classification performance:

[1] http://cas.sdss.org

128×128 (128x) and 64×64 (64x) pixels images. To accomplish this, we first cropped the original images in the GZ1-T2 dataset to their half size (212×212 pixels) and converted them to TIFF format, keeping the galaxy in the centre of the image. After this, we compressed the resulting images to the two sizes referred above.

In all the experiments, we used a 5-fold cross-validation scheme (Section 2.5.1). For AEs, CNN and ResNet, the training set was split into 70/30 for training and validation. In the classification with extracted feature sets, the classifiers training were carried out consistently: same data partitions defined for the AEs training were used for the training and testing of the classifiers, aiming to resemble real working conditions for the classification of unseen data. Since the problem classes are balanced (Table 4.5), we drew upon the Acc measure (Section 2.5.3), taking as final value the average over the five test data partitions. We also analysed the performance in terms of runtime, aiming to estimate a runtime comparison between the different approaches studied here. We examined the FE runtime, taken by both AE models and WND-CHARM, and the classification runtime employed by the classifiers, including the CNN and ResNet. For AEs, the runtime shown accounted for the training and computation of features, whereas WND-CHARM directly computed the features with no training. The classification runtime presented accounts for the training and classification stages, considering negligible the time taken by the FS phase, when applied.

All experiments involving the classification of the feature sets were carried out in a single node with an Intel(R) Xeon(R) CPU E5-1650 v4 processor (12 cores) at 3.60GHz, and 64 GB of RAM. For the training of the DL-based models, we employed a NVIDIA Titan Xp GPU. In terms of software, the Keras[2] Python package was used for the AEs, CNN and ResNet, and the Scikit-learn[3] library for all experiments involving either the training and classification phases of the standard classifiers introduced above. The WND-CHARM implementation used here is freely available in Python language[4]. We acknowledge the disparity in using a GPU and a CPU in the experiments. The runtime analyses carried out are thus taken as an estimation, with no intention to set an exhaustive comparison of the algorithms in terms of the runtime employed.

[2] https://keras.io/
[3] https://scikit-learn.org
[4] https://github.com/wnd-charm/wnd-charm

4.4 Results and Analysis

This section presents the experimental results. First, we carry out a comparative study using the GZ1-E subset and considering both image sizes as well as colour and greyscale images in Section 4.4.1. Then, in Section 4.4.2, we focus on the GZ1-A subset to investigate the scalability of the best classifier from GZ1-E, and to leverage the larger number of amateur-based classifications with our proposed pre-training scheme for CNNs. Finally, we analyse the results obtained in Section 4.4.3.

4.4.1 GZ1-E: Expert Classifications

Due to its reduced size, we first used the GZ1-E sample with expert labels, investigating the performances of both AE models proposed and the influence of image size and use of colour in classification results. The CAE was tested with colour and greyscale images, whereas the DAE was only tested with greyscale images. For WND-CHARM, the two feature sets available for colour and greyscale images were computed. Thus, we obtained a distinct feature set for each FE method and image colour/size configuration.

The resultant number of features is indicated in Table 4.6, and a visual comparison of the runtime spent by these methods is presented in Figure 4.4. For visual illustration of the behaviour of both AEs, Figure 4.5 plots the image reconstructions performed for a selection of images from the GZ1-E sample. Top row presents the original 128x images from the GZ1-T2 dataset. Middle and bottom rows show their reconstructions performed by DAE and CAE architectures, respectively. As it is shown, the DAE model disregards colour channels and is less sensitive to the presence of artefacts in the image and galaxy contours. Conversely, the CAE model defines borders more accurately and partially replicates colour in the images.

	No. Features	
FE method	64x images	128x images
DAE	256	256
CAE Greyscale	512	2,048
CAE Colour	512	2,048
WND-CHARM Greyscale	2,919	2,919
WND-CHARM Colour	4,059	4,059

TABLE 4.6: Number of extracted features in GZ1-E subset.

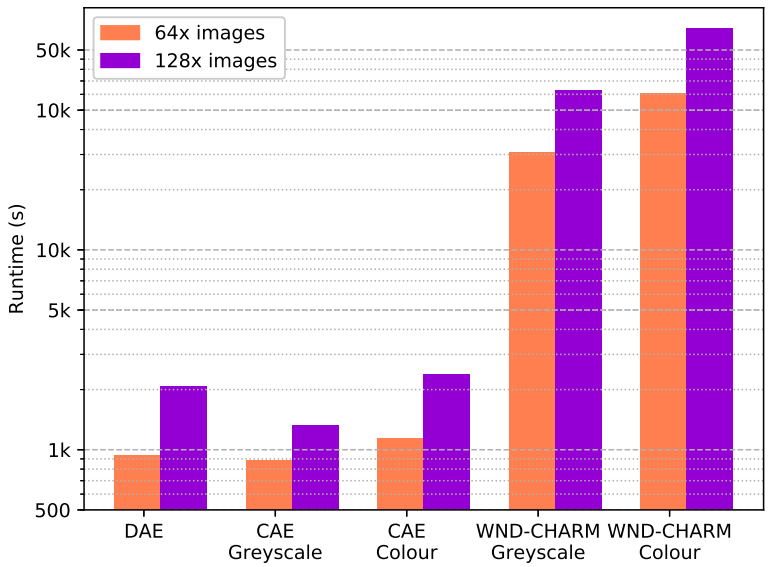


FIGURE 4.4: FE runtime in logarithmic scale for GZ1-E sample.



FIGURE 4.5: Sample of galaxy images reconstructed by the proposed AE models.

These feature sets were taken as input to the classifiers selected for the study. In first place, we carried out the classifications with no FS, pursuing a first comparison of the entire AEs and WND-CHARM feature sets for greyscale and colour images and both image sizes proposed. We then investigated the application of the two proposed FS methods aiming to speed up the classification phase. Acc results for these experiments are shown in Table 4.7 for both image sizes, with no FS (top sector of the table), Fisher scores FS (middle sector) and embedded FS (bottom sector). Comparative representations of the runtime are

presented in Figures 4.6 and 4.7 for 64x and 128x images, respectively, where the distinct classifiers and FS methods are represented by colour and intensity, respectively. These values correspond to the average of the classification runtime over the five data partitions.

		kNN		RF		SVM	
	FE method	64x	128x	64x	128x	64x	128x
No FS	**DAE**	0.8629	0.8648	0.8889	0.8898	0.8743	0.8699
	CAE Greyscale	0.8744	0.8677	0.9127	0.9128	0.8762	0.7685
	CAE Colour	0.8758	0.8742	0.9070	0.9253	0.8353	0.8540
	WND-CHARM Greyscale	0.8848	0.8822	0.9281	0.9255	0.9412	0.9379
	WND-CHARM Colour	0.9184	0.9161	0.9442	0.9444	**0.9485**	0.9482
Fisher FS	**DAE**	0.8639	0.8651	0.8776	0.8798	0.8765	0.8845
	CAE Greyscale	0.8737	0.8688	0.8505	0.8483	0.8695	0.7733
	CAE Colour	0.8751	0.8734	0.8352	0.8625	0.8299	0.8451
	WND-CHARM Greyscale	0.8816	0.8831	0.9265	0.9255	0.9408	0.9380
	WND-CHARM Colour	0.9193	0.9165	0.9439	0.9446	**0.9521**	0.9512
Embedded FS	**DAE**	0.8659	0.8648	0.8779	0.8767	0.8412	0.8789
	CAE Greyscale	0.8803	0.8824	0.8516	0.8418	0.8723	0.7120
	CAE Colour	0.8834	0.8914	0.8205	0.8604	0.8339	0.8609
	WND-CHARM Greyscale	0.8917	0.8945	0.9344	0.9314	0.9338	0.9318
	WND-CHARM Colour	0.9201	0.9203	**0.9495**	0.9490	0.9479	0.9475

TABLE 4.7: Results for 64x and 128x images of GZ1-E sample. The best result for each FS approach employed has been highlighted in bold style.

In third place, we performed the classification of the GZ1-E images using the proposed CNN and the ResNet model as a comparative algorithm. Here we also explored the use of greyscale and colour images and the two image sizes established for the study. These results are shown in Table 4.8.

	64x images		128x images	
	Acc	Runtime (s)	Acc	Runtime (s)
CNN Grey	0.9360	806.9	0.9381	1,539.5
ResNet Grey	0.9181	5,923.3	0.9472	9,136.6
CNN Colour	0.9483	1,207.6	0.9516	2,544.5
ResNet Colour	**0.9584**	5,879.2	**0.9633**	9,045.7

TABLE 4.8: Results of proposed CNN and ResNet for 64x and 128x images of the GZ1-E sample. The best Acc result for each image size has been highlighted in bold style.

Finally, we carried out a comparison amongst the total classification time of both strategies analysed using the GZ1-E subset. As an estimation of this time for the approaches with FE, we added the FE time of both AE models and WND-CHARM to the classification times obtained for each image configuration (Figures 4.4, 4.6 and 4.7). For the sake of simplicity, here we selected the tandem Embedded FS plus RF classifier for both AE models and

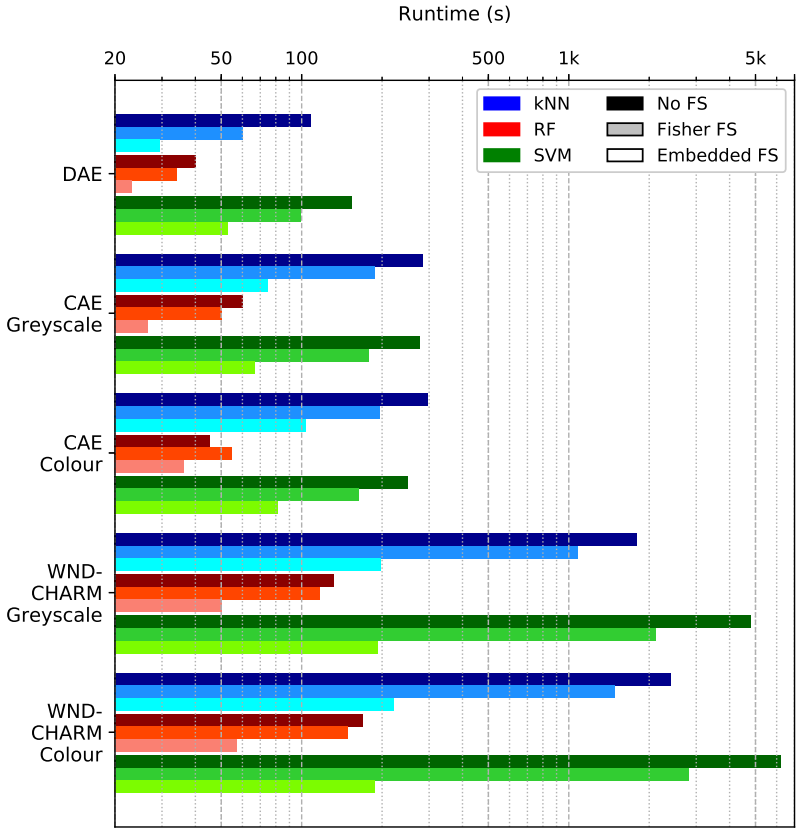


FIGURE 4.6: Classification runtimes in logarithmic scale for 64x images of the GZ1-E sample.

WND-CHARM feature sets classification, since this setting offered the best accuracy/runtime trade-off in the experiments presented above (Table 4.7 and Figures 4.6 and 4.7). This is represented in Figure 4.8. This runtime is subdivided in FE runtime and classification runtime.

For a visual illustration of the classification problem investigated and how the different approaches and algorithms work, Figure 4.9 displays a selection of images from the GZ1-E subset with a wide-range of image qualities and difficulty. For simplicity, we restrict the illustration to 128x colour images and select as well the combination Embedded FS plus RF classifier for both AE models and WND-CHARM feature sets classification. We include the result of DAE, CAE, and WND-CHARM FE methods, and both CNN and ResNet classifiers. Two top rows are classified by experts as *elliptical* and two bottom rows as *spiral*. For each of the images, we indicate which approaches misclassify (red) or correctly predict the label (blue), considering expert classifications as

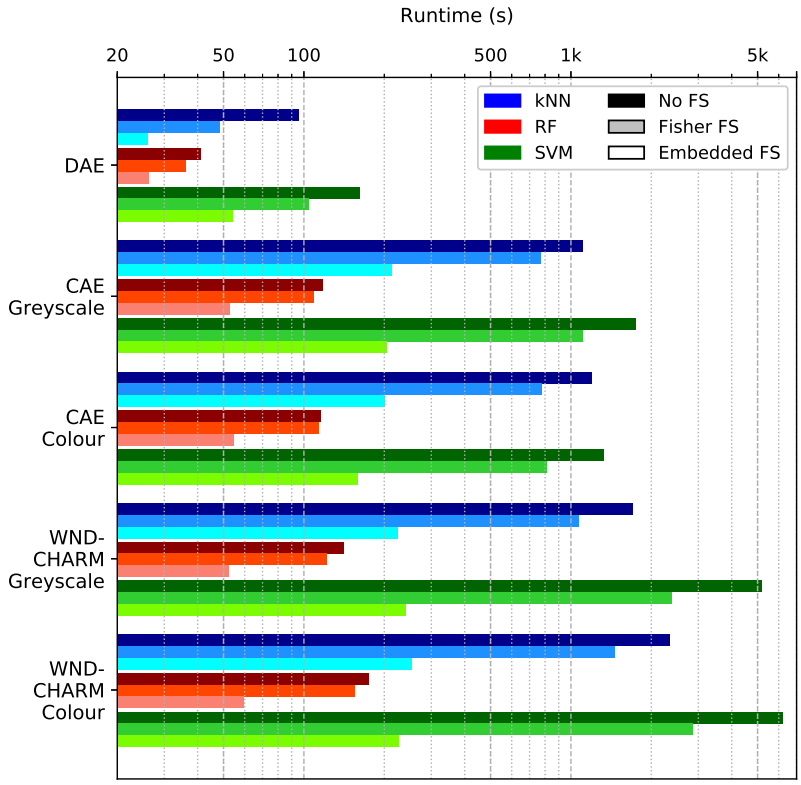


FIGURE 4.7: Classification runtimes in logarithmic scale for 128x images of the GZ1-E sample.

ground truth.

4.4.2 GZ1-A: Amateur Classifications

Using the GZ1-A subset, the aim of this subsection is two-fold. First, we use this bigger dataset to analyse the scalability of the methods compared in the experiments above. Second, we exploit the huge number of existing amateur-labelled galaxy images by pre-training both CNN and ResNet models on this dataset.

Scalability of Methods

After the first set of experiments using the GZ1-E subset, we extended the comparative study to the GZ1-A sample. For this larger dataset, we only compared the features obtained with the CAE and the WND-CHARM for 64x colour images. We completed the classification with RF algorithm, which showed the best balance between runtime and accuracy in the previous experiments with

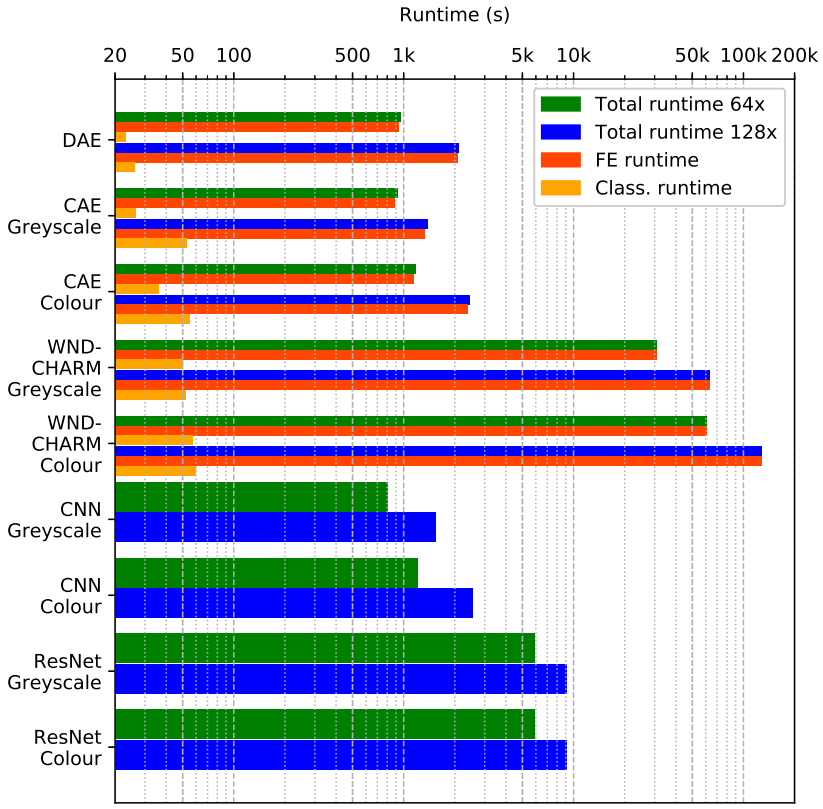


FIGURE 4.8: Total classification runtime in logarithmic scale for both strategies studied with the GZ1-E sample.

the GZ1-E sample. We also examined the application of both FS methods proposed. These results are shown in Table 4.9. The representation of classification runtime is presented in Figure 4.10, which also includes the FE runtime for the CAE and WND-CHARM. In line with the previous study of GZ1-E, we also carried out the classification with the proposed CNN and ResNet on this sample. Results of accuracy and runtime are shown in Table 4.10.

FE method	**No FS**	**Fisher FS**	**Embedded FS**
CAE	0.8481	0.8494	0.8456
WND-CHARM	0.8674	0.8696	0.8734

TABLE 4.9: Results for the GZ1-A sample.

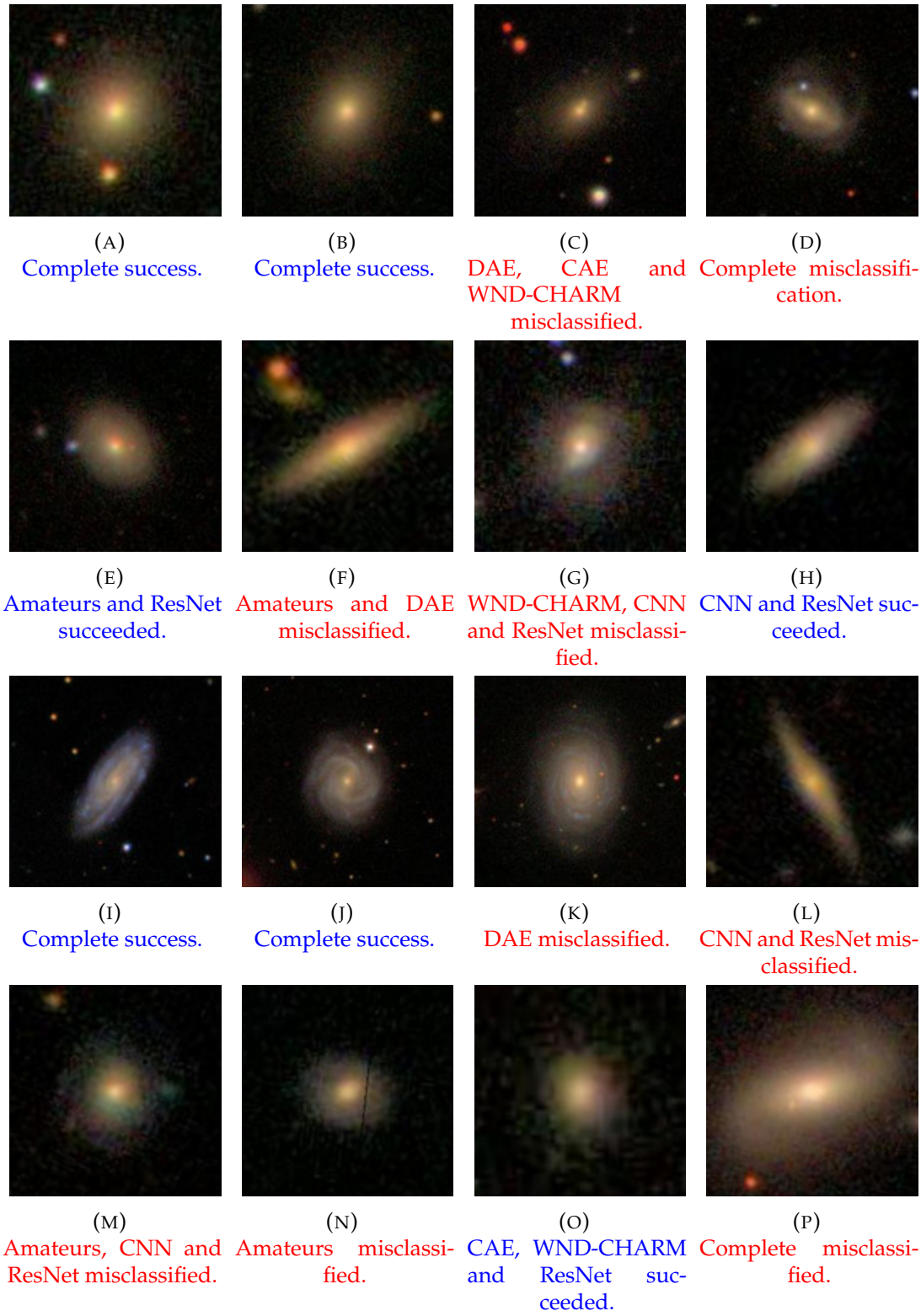


FIGURE 4.9: Sample of 128x colour images from the GZ1-E subset.

Finally, we compared an estimation of the total classification time of both approaches analysed with the GZ1-A sample. As we did with the GZ1-E subset, we added the FE time to the classification runtime for CAE and WND-CHARM. These results are represented in Figure 4.11. This runtime is subdivided in FE runtime and classification runtime.

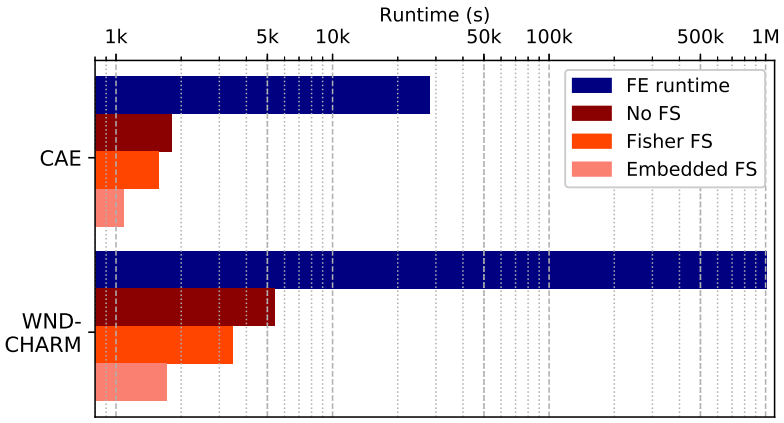


FIGURE 4.10: FE and classification runtimes in logarithmic scale for the GZ1-A sample.

Neural Network	**Acc**	**Runtime** (s)
CNN	0.8989	13,522.3
ResNet	0.9054	96,551.4

TABLE 4.10: Results of proposed CNN and ResNet for the GZ1-A sample.

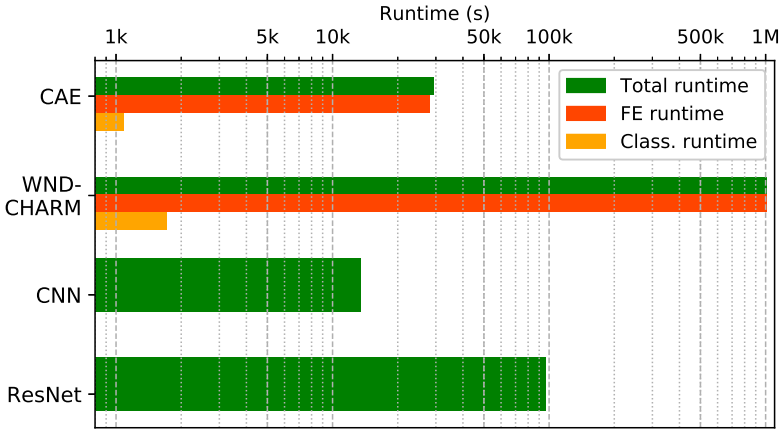


FIGURE 4.11: Total classification runtime in logarithmic scale for both strategies studied with the GZ1-A sample.

Pre-training with Amateur and Expert Labels

As we explain in Section 4.2.3, CzS projects enable a novel methodology for the pre-training and fine-tuning of CNNs. By making use of amateur and expert classifications, the network can be pre-trained employing amateur labels, which are expected to be higher in number and coarser in comparison with their expert counterparts. Then, the inclusion of expert labels permits the fine-tuning, occasionally re-defining the output (number of classes) of the network.

For this experiment, we considered 64x colour images. We first trained CNN and ResNet on GZ1-A using amateur labels, and then re-trained the network (from previously learned weights for *all* layers) on GZ1-E with expert labels carrying out the usual cross-validation established for all experiments. Since the smaller GZ1-E sample is included in GZ1-A, we removed the overlapping between both samples. That is to say, the pre-train phase skipped the examples in GZ1-A later used in the fine-tuning with GZ1-E. We refer to this data sample as GZ1-A*, which consisted of 617,986 examples. In contrast with usual pre-training approaches (Marmanis et al., 2016; Kieffer et al., 2018; Lopes and Valiati, 2017), here the number of classes does not change and therefore we kept the dense part of both networks for the fine-tuning. Results for this experiment are shown in Table 4.11.

		Runtime	**Runtime**	**No Pre-training**	
Neural Network	**Acc**	**Pre-training** (s)	**Fine-tuning** (s)	**Acc**	**Runtime** (s)
CNN	0.9574	55,419.3	841.8	0.8989	13,522.3
ResNet	**0.9643**	107,430.63	5,775.0	0.9054	96,551.4

TABLE 4.11: Results of proposed CNN and ResNet, implementing the pre-training and fine-tuning with the GZ1-A* and GZ1-E subsets, respectively, and with no pre-training process. The best global result has been highlighted y bold style.

4.4.3 Analysis of Results

An examination of the tables and charts presented above allowed us to conclude the following remarks after the experiments:

- Amongst the FE approaches compared, AEs have demonstrated to perform the extraction of features in a shorter amount of time, as it is shown in Figures 4.4 and 4.10. However, the classification with the WND-CHARM feature set provided better accuracy compared to AE features across both GZ1 data samples (Table 4.7 for GZ1-E and 4.9 for GZ1-A). In broad terms, RF generally outperformed the kNN and SVM algorithms and the use of 64x or 128x images did not make a big difference. Nonetheless, the presence of colour provided better accuracy with respect to greyscale images. This was accomplished at the cost of higher runtime as well, in particular for the WND-CHARM Colour feature extractor, which has proved to be the best feature set in terms of classification accuracy.
- The two proposed FS methods did not have a big impact on the classification accuracy. However, they considerably diminished the classification

runtime, especially for the WND-CHARM feature sets, as it is presented in Figures 4.6 and 4.7. The most promising results were obtained for the embedded FS method, probably due to the dynamic nature of this approach that selects a variable number of features. Conversely, the method based on Fisher scores always filters a fixed number of features, thus providing a runtime reduction that remained steady.

- Although the WND-CHARM Colour feature set yielded better classification accuracy, both AE models proposed here have proved to greatly accelerate the FE process (Figures 4.4 and 4.10), which could be decisive for the classification of big volumes of data. Amongst the proposed AE models, the CAE provided the best results in terms of accuracy with respect to the DAE, also enabling the use of colour images. For this architecture, the global classification time with the GZ1-E dataset is comparable to the classification using the proposed CNN (Figure 4.8). This confirms the potential utility of AEs in the classification of large amounts of image data, given that the AE's training would be completed only once.

- Both analysed CNNs provided the best performance in comparison with the three FE approaches studied in terms of accuracy/runtime balance for both data samples used in the study (Tables 4.8 and 4.10). Nonetheless, the WND-CHARM Colour feature set was able to obtain comparable accuracy in the GZ1-E subset when the classification was made using the SVM algorithm (Table 4.7). The difference was enlarged with the GZ1-A sample (Table 4.9), showing that CNNs coped better with the learning from larger amounts of data that probably contain more noise in the labels, and also revealing that amateur labels tended to degrade the classification accuracy.

- ResNet generally outperformed the proposed CNN model in terms of accuracy in the experiments with GZ1-E with expert labels. However, this was achieved at the cost of a much higher runtime that grew up to seven times for greyscale images (Table 4.8). In contrast, the improvement was marginal in the classification of GZ1-A with amateur labels, where a huge increase of the runtime did not provide a much better accuracy (Table 4.10). These experiments reveal that deeper architectures do not always translate in much more improved results, and that the selection of the model can be critical for an optimal DL classification approach in terms of time.

- The proposed pre-training and fine-tuning scheme, using both amateur and expert labels, showed a promising result as a way of leveraging all the potential of CzS projects (Table 4.11). Here the improvement in accuracy (comparing with the previous experiment with GZ1-E only, Table 4.8), was greater for the proposed CNN, indicating that the addition of more layers in the network did not provide any substantial improvement with coarser labels. However, the pre-training phase considerably enlarged the total runtime for both CNNs. This experiment confirmed the adequacy of considering expert and amateur labels to feed the learning of a ML approach, which specially applies in complex classification problems such as the one studied in this paper. For example, images 4.9d, 4.9e, 4.9l and 4.9n demonstrate that even the best approaches compared in the study are prone to fail (if we consider experts' judgements as ground truth), and an integrated use of all knowledge about the problem (e.g. expert classifications, additional astronomical data, CzS results) can be crucial.

4.5 Summary

In this chapter, we have presented a comparative study about the performance of two different strategies for the automated classification of galaxy images, either classifying a feature set obtained from the image or with CNNs. Through a set of experiments, we have compared the state-of-the-art feature extractor, the WND-CHARM, with the suitability of AEs for FE over galaxy images. We have then compared these results with the end-to-end classification provided by two models of CNNs under the same experimental setting. We have explored the impact of the image size and the presence or not of the colour channels in the classification results, also studying the effect of two distinct FS methods. The experiments have been run using two different samples from the GZ1 image dataset, also studying the scalability of both approaches to larger data and the influence of amateur and expert classifications in the accuracy results. In addition, we have tested the adequacy of a novel approach based on pre-training and fine-tuning of CNNs that have proven to take advantage of both label sets available for this problem, enabling the learning of a ML classifier of both label sets simultaneously for the first time.

The obtained results allow us to conclude that CNNs offer the best trade-off

between runtime and Acc, although the addition of a big depth and complexity in the network does not always provide a significant improvement in their prediction capabilities. Also, AEs represent a promising alternative for the classification of these images with the FE approach. This is a consequence of their ability to separate the FE and learning processes, which could eventually be beneficial when the amount of data to be classified expands. Finally, it has been shown that very promising results may eventually come from the learning of both amateur and expert label sets that CzS projects offer.

After the work presented here, in the following chapter we plan to enhance the learning phase of the proposed CNN model with the consideration of unlabelled data in conjunction with different levels of confidence in the images labelling. For this, the CAE model analysed will serve as a key piece to connect the learning from unlabelled and labelled data.

Chapter 5

CzSL: A Novel Learning Paradigm for Citizen Science

5.1 Introduction

In previous chapters, we have dealt with the main drawback derived from the commitment of amateurs in research-related tasks. The first stage of this thesis focused on the prevalent uncertainty within amateur-labelled data, enabling a more integrate use of this data and improving the CzS paradigm utility (Jiménez, Triguero, and John, 2019). As we have seen, CzS is gaining a great momentum as a valuable tool for the labelling of large repositories of image data by the general public, as the GZ1 and subsequent editions of the project have widely exemplified. Despite this, the CzS approach is prone to fail in the classification of the upcoming large volumes of data that are being generated by modern equipment (Ivezić et al., 2019; Deibe, Amor, and Doallo, 2020), and current CzS projects are being coupled with ML techniques, pursuing the automation of these classification tasks (Walmsley et al., 2020). Nonetheless, these approaches have neglected the fact that in this kind of problem, apart from amateur-labelled data, there is usually (limited) expert knowledge about the problem and vast amounts of unlabelled data that have not been exploited within a unified learning framework. Pursuing an improvement on this, in the second research stage we investigated the more suitable ML approaches for classification problems covered by CzS and have developed a first methodology to learn from amateurs and experts. After these previous achievements, the third stage deployed in this chapter aims to merge all previous findings, extending the learning from amateur and expert classifications to unlabelled data.

This chapter presents the Citizen Science Learning (CzSL) approach, a novel

ML methodology for CzS capable of simultaneously taking advantage of data either labelled by experts, amateurs, and unlabelled data. We leverage the achievements of preceding chapters to come up with a DL-based solution that investigates different ways to effectively use all levels of knowledge generated around the completion of CzS projects. On the one hand, the transformed data in which the prevalent uncertainty in CzS data has been ameliorated, investigated in Chapter 3 is leveraged as input to a ML classifier. On the other hand, we extend the methodology tested in Chapter 4, which is improved with the use of AEs that enable the learning from unlabelled data. For the first time, an approach is able to effectively use the different levels of knowledge available (Section 1.3), adopting a sort of *extended* semi-supervised learning (SSL) (Zhu and Goldberg, 2009) methodology capable of learning from unlabelled data and data holding two distinct confidence levels in the labelled examples.

As in previous chapters, the experiments are completed using the GZ1 data. Additionally, here we contemplate two more complex classification problems besides the previously investigated binary classification scheme. First, from the study developed by experts in Longo (2011) we derive two sub-classes of spiral galaxies, depending on the handedness of the spiral arms. Together with ellipticals, this results in a multi-class classification problem (Ou and Murphey, 2007). Second, the extensive study of galaxy mergers derived from the GZ1 results (Darg et al., 2010; Holincheck et al., 2016) suggests the employment of this rare class together with the two main morphologies, conforming an imbalanced classification environment (López et al., 2013). The obtained results confirm that the proposed CzSL method is able to improve the classification performance in a variety of empirical settings.

The chapter is structured as follows. First, we introduce the CzSL methodology to learn from amateur- and expert-labelled data, and unlabelled data along with the DL-based implementation used in Section 5.2. Then, the experiments completed are described in Section 5.3, including the new multi-class and imbalanced datasets introduced and several comparative approaches that enrich the test of the CzSL solution. Experimental results are presented together with their analysis in Section 5.4, and finally, the achievements of the third research stage are reviewed in Section 5.5.

5.2 CzSL: Learning from CzS Data, Experts and Unlabelled Data

This section describes the proposed CzSL model. First, in Section 5.2.1 we discuss the motivations that led us to the development of such an approach and justify several details of the proposed implementation. After this, we explain the followed methodology in Section 5.2.2 and provide the implementation details in Section 5.2.3.

5.2.1 Motivation: Levels of Knowledge

As we introduce in Section 1.1, around the CzS paradigm are generated two main resources worth considering for the development of an automated approach for the classification with CzS data. On the one hand, we have a pool of amateurs that volunteer their time and efforts to complete an image classification task through the completion of a CzS project on the web. They are people with a wide –and unknown– range of capabilities and backgrounds, ranging from children to seniors. Therefore, the resulting amateur-labelled data suffers from an inherent uncertainty that affects its quality, thus representing the *amateur knowledge* about the data. On the other hand, a cohort of experts in the field are able to validate and complement the amateurs' work. Conversely, these experts count on an extensive expertise on the problem as well as supplementary sources of information about the data being classified that support their judgements. They are thus expected to provide more reliable classifications in comparison with CzS participants, signifying the highest level of knowledge about the data, the so-called *expert knowledge*. In conjunction with expert and amateur classifications, the CzSL solution also delves into the exploitation of unlabelled data, that is, the lowest level of knowledge in this data framework (Section 1.3).

To date, many works have attempted the employment of CzS data although ignoring the potentiality held by expert-labelled and unlabelled data for the training of automated classifiers, either in off-line (Banerji et al., 2010; Beaumont et al., 2014; Bahaadini et al., 2018) or more recent on-line approaches (Zevin et al., 2017; Coughlin et al., 2019; Walmsley et al., 2020). Efforts in this direction have mainly targeted how to teach the machine to classify as amateurs (Banerji et al., 2010; Kuminski et al., 2014; Dieleman, Willett, and Dambre, 2015), thus emulating amateur classification skills but also replicating the biases and weaknesses latent within CzS data. On the other hand,

expert classifications have been leveraged to validate CzS results (Lintott et al., 2008; Willett et al., 2013; Willett et al., 2017). Nonetheless, the amounts of data either labelled by experts or in the course of CzS projects result negligible in comparison with the availability of unlabelled data, although no attempts of SSL-based approaches in conjunction with CzS data are found in the literature. Hence, no automated approach has been proposed to exploit all levels of knowledge available around the completion of CzS projects.

The proposed CzSL algorithm embraces the whole set of levels of knowledge. We model this scenario as a novel learning paradigm, proposing a unified framework to learn from both expert and amateur labels, and unlabelled data. To this aim, the development of a DL-based implementation enables the learning from amateur and expert labels making use of the pre-training and fine-tuning of a CNN, which has been thoroughly tested with success in previous experiments (Section 4.4.2). Instead of using amateur and expert classifications separately, CzSL is able to extend the expert knowledge encapsulated in expert classifications to original amateur scores included in CzS data, resulting in a larger amount of high-quality labelled data for the training of a CNN classifier. In addition, we utilise the whole data in conjunction with a CAE to learn patterns from the images without considering any label set. Nonetheless, the CzSL approach does not modify the original CzS data as the previous data transformations propose (Chapter 3). Conversely, CzSL is able to analyse the correlation between expert labels and scores to come up with enhanced amateur labels, which avoids the persistent problem of producing final labels from the CzS data (Section A.5). For this, we propose a multi-layer perceptron (MLP) that is employed in the experiments carried out to test the CzSL approach, although any other ML classifier could be adequate for that. In the following section, the CzSL learning methodology is thoroughly described.

5.2.2 CzSL: A Unified Deep Learning Approach for Citizen Science

In this section we complete a formal description of the proposed CzSL learning methodology. Let $\boldsymbol{TR}$ be the training set disposed for the learning from expert classifications, CzS data and unlabelled data, and $\boldsymbol{TS}$ the test set reserved for testing. Within $\boldsymbol{TR}$, we consider an *amateur* subset $\boldsymbol{A}$ of these images that is labelled as CzS data (by amateurs), and the rest of the set as unlabelled data. That is to say, in $\boldsymbol{A}$ we hold amateur scores accounting for the votes performed by the project participants on the web, as it is shown in Table 2.1 (Section 2.2).

Additionally, we assume to own expert classifications for some part of the $\boldsymbol{A}$ subset, an *expert* subset $\boldsymbol{E}$, thus completing the three levels of knowledge referred above. This $\boldsymbol{E}$ subset of $\boldsymbol{A}$ contains the portion of data both labelled by experts and amateurs. This is a basic assumption of the CzS approach: we only consider the available expert classifications for data previously classified in the course of a CzS project. Therefore, $\boldsymbol{E} \subset \boldsymbol{A}$ by definition. Although we do not establish the relative size of these sets, it is assumed that $|\boldsymbol{E}| \ll |\boldsymbol{A}| \ll |\boldsymbol{TR}|$, as derives from actual CzS projects (Lintott et al., 2008; Willett et al., 2013) and given that $\boldsymbol{TR}$ also contains unlabelled data, which is generally abundant in many disciplines. CzSL focuses on exploiting the expert-labelled part of the data, $\boldsymbol{E}$, which hold the highest confidence, and how this labelled set is enlarged throughout the CzSL learning process. Hence, in the subsequent experiments we analyse the prediction capabilities of the model in the rest of the $\boldsymbol{TR}$ data, named as $\boldsymbol{U}$, similarly as it is done in experiments with SSL methods (Triguero, García, and Herrera, 2015). Therefore, $\boldsymbol{E} \cup \boldsymbol{U} = \boldsymbol{TR}$.

The CzSL approach consists of three phases that sequentially learn from unlabelled data, amateur-labelled data, and expert-labelled data in last place. The DL-based implementation developed enables a bottom-up learning: it learns from the bottom upwards in the knowledge framework described above. This is opposed to self-labelling strategies in SSL, which extend the labelling from labelled to unlabelled examples (Triguero, García, and Herrera, 2015). However, such approaches tend to incur in misclassifications in the labelled set enlarging process, also increasing the computational cost in terms of runtime due to the repeated trainings. Conversely, CzSL first tackles the unlabelled data to learn afterwards from amateur and expert labels. By this, the whole $\boldsymbol{TR}$ data aids the CNN pre-training without the use of noisy labels across $\boldsymbol{U}$.

The implementation of the CzSL is based on two key elements: the pre-training and fine-tuning of a CNN aided by a CAE, and a MLP. The former is used in the first and second stages of the method to learn from unlabelled and amateur-labelled data using the CAE and CNN, respectively, which share a common architecture that enables the transfer of weights across the three stages of the method. The latter is employed in the third phase to extend the expert knowledge to the whole CzS data available. In the following paragraphs, we describe the different learning stages of the method, which is represented as a whole in Figure 5.1.

- **Stage 1: CAE-CNN pre-training.** *Exploiting unlabelled data with a CAE.* The CzSL starts by employing the whole set of images in $\boldsymbol{TR}$ set with

no consideration to the image labels. The proposed CAE learns patterns from the data, enabling by this way the unsupervised pre-training of the CNN employed in the subsequent stages of the method. To this aim, the feature extraction layers of the CNN must match the CAE structure, thus enabling the transfer of weights to the next stage (dashed arrows from Stage 1 to Stage 2 in Figure 5.1). We refer to this process as CAE-CNN pre-training, which allows for the exploitation of the entire data available.

- **Stage 2: CNN pre-training.** *Learning from rough amateur labels*. In this stage, the CNN loaded with the CAE weights trained in the previous phase, is pre-trained over the $\boldsymbol{A}$ subset using the amateur labels derived from the scores included in the CzS data. The labels are generated adopting a simple majority procedure: the class holding the highest score is assigned to the example. In case of a tie between the classes, the final label is assigned at random. In summary, this phase pursues the pre-train of the CNN with coarse (rough) amateur labels, which are more numerous than the expert counterparts although noisier as well. These CNN weights will be transferred to the last stage of the learning process. Also, the CzS data is first leveraged in a rough mode (majority criterion), in contrast with the following phase that focus on enhancing the amateur knowledge with the aid of expert classifications.

- **Stage 3.a: MLP extension.** *Extending expert knowledge to CzS data*. Prior to the final fine-tuning of the CNN, this stage develops the extension of expert knowledge in $\boldsymbol{E}$ to the rest of examples in $\boldsymbol{A}$ (note that $\boldsymbol{E} \subset \boldsymbol{A}$, hence this affects to the examples in $\boldsymbol{A} \setminus \boldsymbol{E}$). For this we propose to implement a MLP that learns the correlation amongst expert labels and amateur scores included in CzS data. Hence, the MLP is trained over the expert-labelled portion of the data, that is, $\boldsymbol{E}$. Amateur scores included in CzS data are employed as the MLP input, and expert labels as output. However, as it has been pointed out above, this stage is not dependent on the selected solution to carry out the knowledge extension process, and other suitable classifier could be applied with success in this phase of the method.

- **Stage 3.b: CNN fine-tuning.** *Learning from expert and enhanced amateur labels*. Once the expert knowledge extension has been completed using the amateur scores and expert labels in $\boldsymbol{E}$, the MLP predictions are employed as enhanced amateur labels for the rest of the images in $\boldsymbol{A}$. Thus,

the $\boldsymbol{A}$ image subset is labelled with expert labels for examples in $\boldsymbol{E}$, and enhanced amateur labels for examples in $\boldsymbol{A} \setminus \boldsymbol{E}$. Using these labels, the CNN loaded with the learned weights in Stage 2 (dashed arrows from Stage 2 to Stage 3.b in Figure 5.1) is trained on the entire $\boldsymbol{A}$ subset of images. This completes the CNN fine-tuning with the finest labels, leveraging the highest level of knowledge achieved from expert labels and amateur scores. On the other hand, the CNN has captured the information present in (rough) amateur labels and patterns learnt in previous stages 2 and 1, respectively.

After completing fine-tuning process, the CNN would be ready to predict unseen examples in $\boldsymbol{TS}$ set. Although the CzSL integrates both amateur and expert levels of knowledge, we are also interested in verifying how the treatment of the uncertainty in CzS data addressed by the previously proposed data transformations (Chapter 3) behave in conjunction with the CzSL learning process. To this end, we employ either original or transformed scores as input to the MLP in the Stage 3.a of the method. This aspect will be tested in the experimental section of the chapter. In the following section, we detail the implementation of the elements involved in the CzSL learning stages described above.

5.2.3 Implementation

The CzSL implementation is based on the pre-training and fine-tuning of the CNN employed using a CAE for the learning from unlabelled data, and a MLP for expert knowledge extension to CzS data. The CNN, CAE and MLP architectures presented have been tested in previous experiments (Jiménez, M., M. Torres Torres, R. John, and I. Triguero, 2020) with the GZ1 data and are not proposed as a final or strict implementation for the CzSL approach. Conversely, other CNN or MLP topologies could be beneficial depending on the classification problem being tackled. In the following, the particular CAE, CNN and MLP implementations used in the experiments are detailed.

CAE-CNN Pre-training and CNN Fine-tuning

AEs are normally implemented for the extraction of features from images, which is desirable in multiple classification scenarios. They are ANNs deploying a symmetrical structure of layers that are trained to reconstruct the input pattern into the output, providing an encoding of the input at the end of the learning process (Charte et al., 2018). However, here we pursue to train

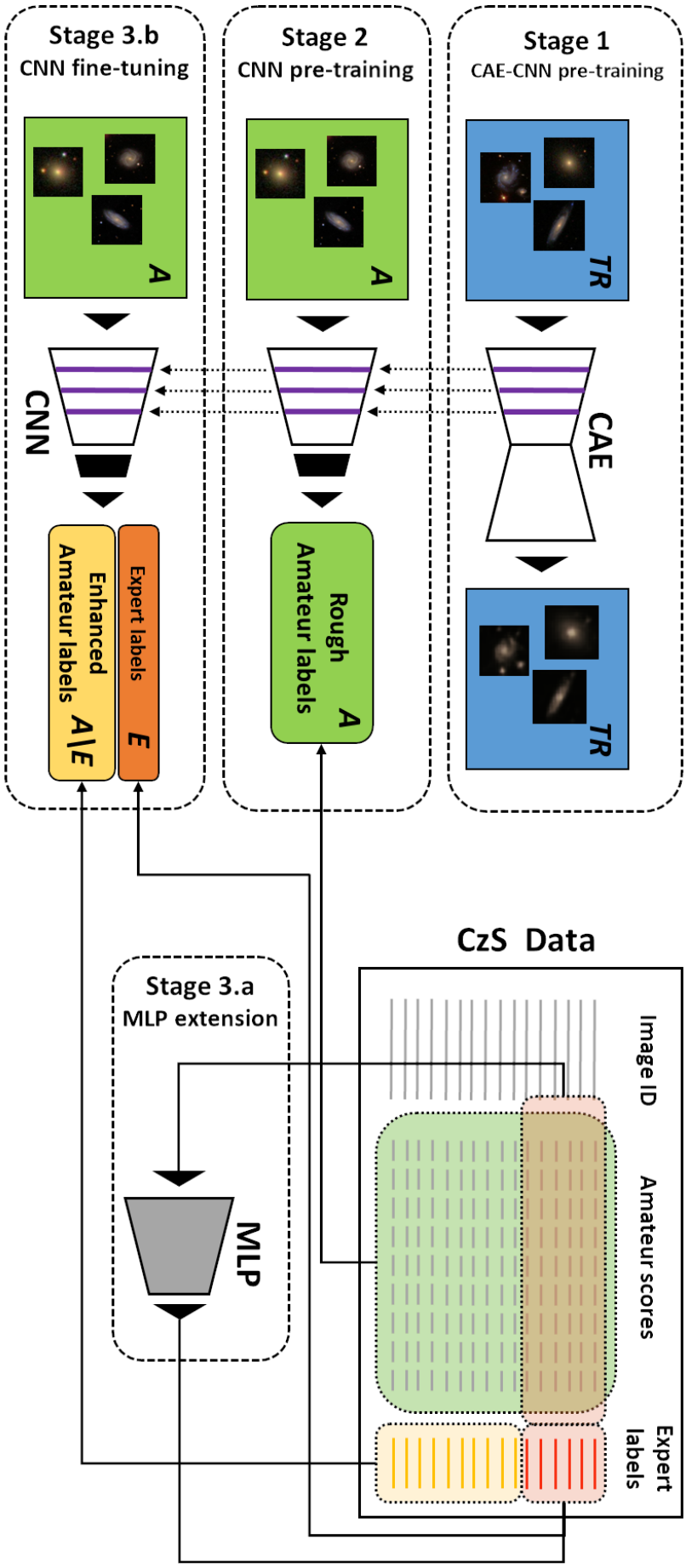


FIGURE 5.1: The CzSL learning workflow.

the CAE over all images to transfer the CAE weights to the CNN used in the subsequent phases of the method with amateur and expert labels.

For the experiments with the CzSL, we implement CAE and CNN models that have been tested in previous experiments within the galaxy image classification problem in Chapter 4. The CAE deploys three pairs of convolution – pooling layers between input and encoding. The complete specifications are summarised in Table 4.2, and the architecture is represented in Figure 4.2. The CNN thus resembles this layer structure. It is composed of three pairs of convolution – pooling layers (Section 2.5.2) as well, followed by two dense layers of 256 and 128 neurons and the output layer with same number of neurons as classification classes defined in the problem. This architecture is summarised in Table 4.3 and represented in Figure 4.3.

MLP for Expert Knowledge Extension

MLPs represent the simplest type of feed-forward artificial neural networks (Lecun, Bengio, and Hinton, 2015), consisting of input and output layers of neurons with a variable number of hidden layers in between. All neurons deployed through the layers are densely connected, and the structure is trained by the back-propagation algorithm (Rumelhart, Hinton, and Williams, 1986).

In order to select the best MLP architecture through experimentation, we conduct a grid search to find an adequate model for the experiments with the classification problems covered. First, we establish a set of architectures involving varied number of layers and neurons. After this, these models are evaluated using the expert-labelled data, $\boldsymbol{E}$, pursuing a fair model selection under the datasets described above. In our experiments, a nested cross-validation (Varma and Simon, 2006) is implemented using this data: we establish training and test sets splitting the $\boldsymbol{E}$ data in 80/20, respectively. In this way, the MLP architectures are first trained using the 80% and then tested in the remaining 20%, using the expert labels as ground truth. Following this procedure, the best-performing MLP implementation is composed of four hidden layers with 8, 7, 5, and 3 neurons, using backpropagation algorithm for optimisation. As with the CNN implemented, SoftMax is applied to the last layer to provide class probabilities. This MLP architecture is shown in Figure 5.2.

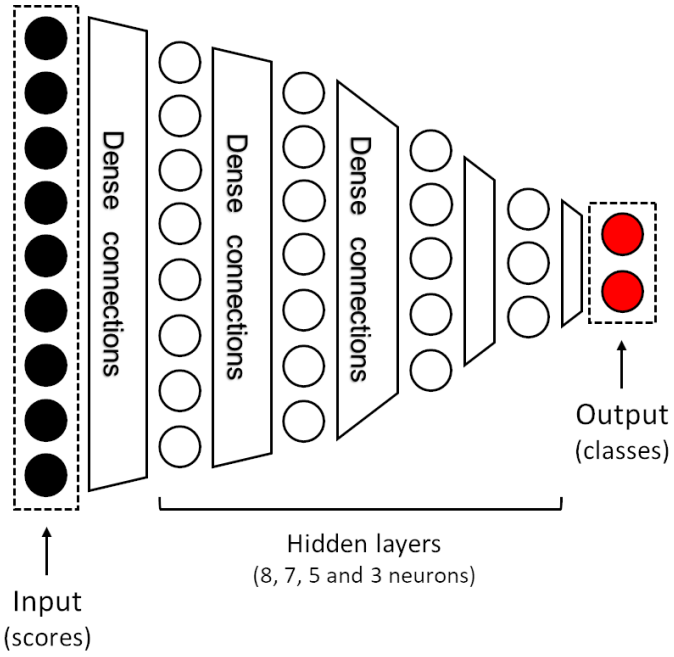


FIGURE 5.2: Architecture of the MLP implemented in the CzSL approach.

5.3 Experimental Framework

This section presents the set of experiments carried out to test the proposed CzSL approach. First, we explain the employed datasets in Section 5.3.1. Then, we describe the evaluation of experiments established in Section 5.3.2. Finally, we introduce a set of comparative approaches that enables a well-grounded analysis of the proposed CzSL approach and specify several implementation parameters in Section 5.3.3.

5.3.1 Datasets

The experiments employ different classification problems derived from the GZ1 project results, aiming to expose the proposed learning methodology under stress in varied environments. As in the experiments completed in preceding chapters, the amateur classifications employed are part of the GZ1-T2 dataset, which does not provide a final classification for the galaxies but a set of votes for each of the images displayed on the GZ1 website. Previous uses of this data have drawn on the application of different thresholds over these scores in order to obtain samples of varied quality. For example, the *clean* and *superclean* samples in (Lintott et al., 2008) are defined selecting examples with 80% and 95% consensus between amateur votes for elliptical and combined spiral classes, respectively. However, this procedure neglects a great part of

the GZ1 results (Jiménez, Triguero, and John, 2019). In contrast, here we apply a strict major criterion to obtain the rough amateur labels leveraged by the CzSL methodology, as follows: the class (morphology) holding the highest number of votes is assigned to the object, allowing us to leverage the whole GZ1 project results. In the case of a tie between the scores (e.g. elliptical and spiral scores both equal to 0.5), we assign a random choice to the example.

On the other hand, the expert classifications considered are primarily obtained from two expert catalogues, which were originally employed by the GZ1 developers for the assessment of the results (Lintott et al., 2008). These are the MOSES catalogue (Schawinski et al., 2007), for elliptical galaxies, and the Longo catalogue (Longo, 2011), for spirals. The combination of both samples conforms a total of 41,424 galaxies classified as elliptical or spiral by a team of expert astronomers that has been referred to as GZ1-E dataset in previous experiments across Chapters 3 and 4. Furthermore, the Longo catalogue also provides two sub-classifications for spirals depending on the galaxy handedness of its spiral arms, which enables us to define a multi-class problem by splitting the spiral galaxies included in GZ1-E into two classes. Thus, from the GZ1-E dataset we define two classification settings: on the one hand, a binary classification problem that covers spiral and elliptical classes; on the other hand, a multi-class problem that considers right-handed spiral, left-handed spiral, and elliptical classes. For simplicity, we refer to these datasets as GZ1-(SE) and GZ1-(RLE), respectively.

In addition, here we also look upon an expert catalogue generated for the study of galaxy mergers within the GZ1 data. This is the catalogue published in Darg et al. (2010), which includes a set of 3,003 visually selected pairs of merging galaxies. For the elaboration of an expert-labelled set for this imbalanced classification problem, we compared the Darg catalogue against the GZ1-E sample: galaxies both included in the Darg catalogue and GZ1-E are taken as mergers[1] (MG), whereas the rest of GZ1-E examples are kept as no mergers (No-MG). In addition, the rest of examples in the Darg catalogue are added to the final sample. More precisely, there is an overlap of 496 examples, and 2,355 examples not present in GZ1-E are added. We refer to this imbalanced dataset as GZ1-(MG) from now on, which consists of 43,779 examples.

[1]This derives from the fact that a galaxy merger represents a collision between two or more galaxies itself. Hence, one merger example could be tagged as spiral or elliptical with validity depending on the morphology of the most prominent component of the merger.

Dataset	No. examples	Class distribution (%)	
		Expert labels	Amateur labels
GZ1-(SE)	40,902	60.7 (S) : 39.3 (E)	58.6 (S) : 41.4 (E)
GZ1-(RLE)	40,902	30.1 (R) : 30.6 (L) : 39.3 (E)	28.8 (R) : 29.9 (L) : 41.3 (E)
GZ1-(MG)	43,779	6.5 (MG) : 93.5 (No-MG)	6.5 (MG) : 93.5 (No-MG)

TABLE 5.1: Datasets used in the experiments with their distribution of classes in percentage.

In GZ1-(RLE) and GZ1-(MG) samples, the majority criterion is applied consistently to obtain amateur labels: right-handed spiral and left-handed spiral classes are assigned to the class with the highest score between Clockwise and Anti-clockwise spiral scores, for examples whose combined spiral score is greater than their elliptical score. Likewise, the MG class is allocated to examples holding a merger score greater than 0.5. As before, ties are handled assigning a random class. In the experiments, we employ either the original GZ1 scores or the transformed ones to generate the rough labels utilised by the CzSL approach in the Stage 3.a of learning. By this, we can eventually test the performance of the data transformations previously investigated (Section 3.3.2) when the improved data is used as input to an automated classifier. In summary, the number of examples and distribution of classes, either considering expert or amateur labels, are presented in Table 5.1 for the three datasets introduced above.

The images utilised through the experiments derive from the original images classified in the course of the GZ1 project. We thus employ the GZ1-T2 images following the original specifications (Lintott et al., 2008). However, the comparative study developed in the previous chapter led us to conclude that the performance of CAE and CNN used in the experiments is not significantly affected by the image size. Similarly as it is proposed in other approaches (Dieleman, Willett, and Dambre, 2015), we first converted the images to TIFF format and cropped them to their half size, keeping the galaxy in the centre. After this, we compressed them to 64x64 pixels size to accelerate the training of the aforementioned models. We thus employ 64x64 pixels RGB images in TIFF format in all the experiments.

5.3.2 Experimental Evaluation

To evaluate the performance of the proposed approach in the datasets described above, we aim to approximate the real conditions for the availability of image labels, either expert or amateur. First, the datasets were partitioned

using a 5-fold cross-validation scheme (Section 2.5.1), so that 80% of the data was used for training ($\boldsymbol{TR}$) and the remaining 20% for testing ($\boldsymbol{TS}$). As it is usually done in experiments with SSL (Triguero, García, and Herrera, 2015), the training set is composed of a small amount of labelled data and a larger set of unlabelled data. Following the nomenclature previously introduced (Section 5.2.2), the vast majority of the examples (75%) are included in the $\boldsymbol{U}$ set and taken as unlabelled (which is simulated ignoring the labels), and the rest (25%) is assumed to hold amateur labels based on CzS, conforming the $\boldsymbol{A}$ set. Of the data included in $\boldsymbol{A}$, we consider as expert-labelled $\boldsymbol{E}$ a small amount (5%) of the entire $\boldsymbol{TR}$. Therefore, the CzSL takes the $\boldsymbol{E}$ set entirely included in $\boldsymbol{A}$: we assume to hold expert classifications for the 20% of the data labelled by amateurs. Mathematically: $\boldsymbol{E} \cup \boldsymbol{U} = \boldsymbol{TR}$, and $\boldsymbol{E} \subset \boldsymbol{A}$. These sets remain fixed within each of the five training partitions arranged as a result of the 5-fold cross-validation. This setting is sufficiently representative and keeps the number of experiments to a reasonable amount. Nevertheless, future work will consider the exploration of 25% and 5% random samples of $\boldsymbol{TR}$ to produce the $\boldsymbol{A}$ and $\boldsymbol{E}$ sets, respectively.

Expert classifications are regarded through all the experiments as ground truth for results evaluation. Although this expert-based labelling is also prone to include noise and errors in the results, we rely on experts judgements as ground truth for the assessment of all algorithms investigated, including the proposed CzSL approach. For GZ1-(SE) and GZ1-(RLE) datasets, given that the class distribution are well-balanced (Table 5.1), we employ the Acc measure as the metric to compare the performance of the different approaches. Acc simply accounts for the ration of correct classifications, the number of correctly predicted examples over the total number of classified examples. However, this measure is less representative in imbalanced classification frameworks, given that the classes are not equally weighted in such cases. Alternatively, we use the G-mean (López et al., 2013) for the experiments involving the GZ1-(MG) dataset. For a formal definition of both performance measures, see Section 2.5.3.

Besides the classification performance, we also look at the runtime required by the approaches. In order to minimise the effect of the randomness as a consequence of the weights' initialisation, we completed ten executions of every single experiment. Hence, the metrics' values presented in the following correspond to the average over ten independent executions of the same experiment. We take as representative the average runtime of the training over one single

partition (***TR*** data). At this respect, we just record the CNN training time, as this is the more significant phase of the experiment in terms of total runtime and accounts for a stage that is common to the three datasets examined. In this respect, all the experiments were run in a single node with an Intel(R) Xeon(R) CPU E5-1650 v4 processor (12 cores) at 3.60GHz, and 64 GB of RAM. The training of CAE and CNN models implemented either in the CzSL approach or other algorithms employed a NVIDIA Titan Xp GPU. In terms of software, we used the Keras[2] and Scikit-learn[3] Python packages for the implementation of CAE and CNN models and evaluation metrics.

5.3.3 Comparative Approaches and Parameters

Aiming to develop an exhaustive investigation of the proposed CzSL algorithm performance, we define a set of comparative approaches that learn from different configurations of the data framework introduced above. All of them make use of the implemented CNN under the same experimental conditions, thus allowing us to analyse how well we are able to predict ***TS*** examples by learning from expert and amateur label sets alone or in combination, and also evaluate the proposed MLP in a pure supervised approach. We also introduce an additional approach inspired in usual self-labelling techniques (Triguero, García, and Herrera, 2015) that takes advantage of the MLP ability to extend the expert knowledge to amateur-labelled data. We first define some approaches that serve as upper- and lower-bounds. Then, other approaches using the MLP and implementing a CNN self-training are described. The complete set of comparative approaches are described in the following points.

- *Oracles*. First of all, we contemplate an ideal situation in which we would hold either expert or amateur classifications for all the data. We refer to these approaches as *oracles*, as they are meant to serve as upper-bounds for the rest of experiments. For both **Expert Oracle** and **Amateur Oracle**, the CNN takes the whole ***TR*** set with expert and amateur labels, respectively, although the evaluation in ***TS*** is always carried out using expert labels (ground truth). These two approaches do not only establish a maximum level for Acc and G-mean values in the following, but also serve as an assessment of the rough amateur labels generated from CzS data.

[2] https://keras.io/
[3] https://scikit-learn.org

- *Pure supervised learning*. We then set out the learning from expert and amateur labels, either independently or together, based on the data partitions defined for the experiments above. By this, we investigate in the simplest way the performance of the selected CNN in a supervised training under the data framework established, testing the learning from only the expert and amateur sets without exploiting any unlabelled data. Furthermore, these approaches will also clarify the analysis of results, providing information about the relative difficulty of the distinct classification problems explored and the potential improvements obtained by the CzSL model, as a sort of lower-bounds. In these **Expert**, **Amateur**, and **Expert-Amateur** approaches, the CNN learns from $\boldsymbol{E}$, $\boldsymbol{A}$, and $\boldsymbol{E} \cup \boldsymbol{A}$ sets, respectively. In the last case, expert labels are considered for examples in $\boldsymbol{E}$ and amateur labels in the rest ($\boldsymbol{A} \setminus \boldsymbol{E}$).

We also test the performance of the proposed MLP regardless of the CzSL model in two comparative approaches. The proposed CzSL model is able to learn patterns from unlabelled data (Stage 1), and then from amateur labels (Stage 2), expert labels and MLP predictions (Stages 3.a and 3.b), leveraging the full range of knowledge available in a bottom-up fashion. Conversely, the two comparative approaches proposed with the MLP illustrate top-down strategies, which learn from expert labels to enhance amateur labels and then extend the labelling to raw data. Hence, these two comparative approaches using the MLP allow us to compare the effectiveness of these two learning strategies.

- *MLP*. The first one, referred to as **MLP**, is intended to check the MLP capabilities to replicate expert labels by using amateur scores only, enlarging the $\boldsymbol{E}$ set with additional examples in $\boldsymbol{A}$. Hence, the MLP is trained in $\boldsymbol{E}$ using expert labels, and then its predictions are used as enhanced amateur labels in $\boldsymbol{A} \setminus \boldsymbol{E}$. The CNN is then trained in $\boldsymbol{E} \cup \boldsymbol{A}$ using expert labels ($\boldsymbol{E}$) and MLP predictions ($\boldsymbol{A} \setminus \boldsymbol{E}$).

- *MLP Self-labelling*. The second approach reports a test of the MLP within a different learning algorithm, allowing us to evaluate this component of the CzSL model in a different setting. This approach, referred to as MLP-Self-labelling and designated from now on as **MLP-SL**, extends the learning algorithm explained above to a basic self-training approach (Yarowsky, n.d.): after completing the propagation of labels as in the MLP approach, the CNN is trained over the entire $\boldsymbol{TR}$ set using expert labels ($\boldsymbol{E}$), MLP predictions ($\boldsymbol{A} \setminus \boldsymbol{E}$) and the CNN predictions for the rest

of examples in $\boldsymbol{U}$, that is, the 75% of the data not included either in $\boldsymbol{E}$ nor $\boldsymbol{A}$. We only complete one single iteration, given that multiple iterations of the employed CNN would invalidate such an approach in terms of training time.

As it has been pointed out, with the MLP and MLP-SL comparative approaches we also investigate the adequacy of the data transformations developed in Chapter 3, which aim to aggregate additional information about the uncertainty present in the original GZ1-T2 dataset. We compare the performance of these approaches employing the transformed scores or the original ones as input to the MLP in the Stage 3.b of the CzSL workflow. The complete set of comparative approaches is summarised in Table 5.2.

Approach	Learns from (% of $\boldsymbol{TR}$)	Labels used	Description
Expert Oracle	$\boldsymbol{TR}$ (100%)	Expert	CNN trained over the entire $\boldsymbol{TR}$ set with expert labels
Amateur Oracle	$\boldsymbol{TR}$ (100%)	Amateur	CNN trained over the entire $\boldsymbol{TR}$ set with amateur labels
Expert	$\boldsymbol{E}$ (5%)	Expert	CNN trained over the $\boldsymbol{E}$ set with expert labels
Amateur	$\boldsymbol{A}$ (25%)	Amateur	CNN trained over the $\boldsymbol{A}$ set with amateur labels
Expert-Amateur	$\boldsymbol{E} \cup \boldsymbol{A}$ (25%)	Expert and Amateur	CNN trained over $\boldsymbol{E} \cup \boldsymbol{A}$ using expert labels for examples in $\boldsymbol{E}$ and amateur labels for the rest ($\boldsymbol{A} \setminus \boldsymbol{E}$)
MLP	$\boldsymbol{E} \cup \boldsymbol{A}$ (25%)	Expert and MLP predictions	CNN trained over $\boldsymbol{E} \cup \boldsymbol{A}$ using expert labels for examples in $\boldsymbol{E}$ and MLP predictions for the rest ($\boldsymbol{A} \setminus \boldsymbol{E}$)
MLP-SL	$\boldsymbol{TR}$ (100%)	Expert, MLP predictions and CNN predictions	CNN first trained over $\boldsymbol{E} \cup \boldsymbol{A}$ using expert labels ($\boldsymbol{E}$) and MLP predictions ($\boldsymbol{A} \setminus \boldsymbol{E}$), and then over the entire $\boldsymbol{TR}$ adding its own predictions for the rest of examples in $\boldsymbol{U}$ ($\boldsymbol{U} \setminus \boldsymbol{A}$)

TABLE 5.2: Comparative approaches established for the evaluation of the CzSL model.

Finally, for the experiments with the GZ1-(MG) dataset, we provide two additional settings using this dataset in order to verify the experimental results free from the added difficulty of imbalanced classification. Amongst the plethora of data pre-processing techniques that have been widely studied within the literature (Thabtah et al., 2020), here we select the random over-sampling (ROS) and random under-sampling (RUS), two very simple approaches that have extensively provided good results and allow us to understand the learning in this environment (Triguero et al., 2015; Galar et al., 2012). ROS and RUS are sampling techniques that adjust in the simplest way the imbalance ratio (IR), defined as the ratio between the majority and minority classes numbers of examples in the data (López et al., 2013). For ROS, instances of the minority class are randomly replicated until certain IR is achieved. In the case of RUS, examples of the majority class are randomly removed to balance the IR. Here we do not intend to test these techniques or present a representative application of them. Conversely, we aim to provide a deeper exploration of the different approaches' performance under the imbalanced classification problem. In our experiments with the GZ1-(MG) dataset, ROS and RUS are applied to reach IR

= 1.0, that is to say, same number of MG examples and No-MG examples in the data.

In terms of DL models' parameters, CAE and CNN were trained over 100 epochs with a batch size of 256 instances, using stochastic gradient descent for optimisation (Wijnhoven and De With, 2010), and ***TR*** was split into 70/30 for training and validation. The MLP was trained over 300 iterations, employing the ReLU activation function (Nair and Hinton, 2010) and stochastic optimisation (Wijnhoven and De With, 2010). As original, the scores for *Elliptical*, *Combined Spiral*, *Clockwise Spiral*, *Anti-clockwise Spiral*, *Edge-on Spiral*, *Don't Know*, and *Merger* classes were employed as input to the MLP (Stage 3.a of the CzSL training). When the transformed scores are employed, the elliptical and combined spiral original scores were replaced by the scores obtained after applying the data transformation that best performed in the previous research (Section 3.4). More precisely, this transformation sequence consists of a combination of *Normalisation*, *DK Votes Shift*, and*Votes Boost* single transformations, completed in this order.

5.4 Results and Analysis

This section presents the results of the experiments conducted for the testing of the proposed CzSL approach in accordance with the experimental setup and comparative approaches described in the preceding section. First, in Section 5.4.1 we display the results for oracles, which represent the upper limits for the experiments involving CzSL and other approaches. Then, we summarise all the experimental results along with their analysis for the three covered datasets in Sections 5.4.2 and 5.4.3.

5.4.1 Oracles

Oracles employ the entire ***TR*** set for completing the CNN training in a pure supervised way, either using expert or amateur labels. These are meant to provide reference points for the experimental analysis in several ways. First, taking a global scope, they allow us to differentiate between the difficulty of the three classification problems investigated, since the experimental setting remains unchanged across the three problems investigated: same classifier,

image features, and data amounts[4]. Second, from a more local perspective, expert oracles set upper bounds for the comparison of results within the same classification scenario, thus enabling an objective evaluation of the improvements either provided by the comparative approaches or the proposed CzSL method. In addition, they imply a true evaluation of the amateur label set quality versus the ground truth considered (expert labels), given that all other approaches either combine the learning from both label sets or learn from dissimilar amounts of data. These results are shown in Tables 5.3 for Acc and G-mean values in ***TS*** set, and Figure 5.3 for average training times for the CNN over one single ***TR*** partition (average runtime over the ten executions completed). The computational time taken by ROS and RUS pre-processing phases is considered negligible.

Dataset	**Expert labels**	**Amateur labels**
	Acc_{TS}	Acc_{TS}
GZ1-(SE)	0.9510 ± 0.0037	0.9169 ± 0.0053
GZ1-(RLE)	0.8181 ± 0.0166	0.8065 ± 0.0165
	G-mean_{TS}	G-mean_{TS}
GZ1-(MG)	0.7592 ± 0.0274	0.5768 ± 0.0613
GZ1-(MG) + ROS	0.8487 ± 0.0026	0.7669 ± 0.0048
GZ1-(MG) + RUS	0.8738 ± 0.0092	0.8754 ± 0.0050

TABLE 5.3: Results for oracles in the three datasets investigated.

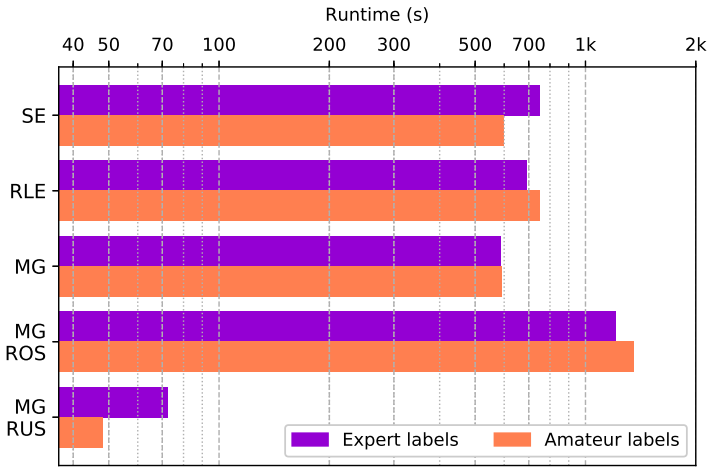


FIGURE 5.3: Execution times in logarithmic scale for oracles.

Although the three classification problems revolve around the same sort of images, experimental results show a wide range of difficulties in the recognition

[4]Strictly speaking, the GZ1-(MG) dataset is 2,877 images bigger (∼7.0%) than GZ1-(SE) and GZ1-(RLE), but we do not consider this difference as significant, and it is in favour of the most difficult problem at first.

of patterns by the employed CNN. As expected, these spread throughout the distinction amongst elliptical and spiral as the easiest, then the split between right-hand and left-hand spirals, and the identification of galaxy mergers as the hardest problem. The pre-processing techniques used with the GZ1-(MG) dataset also result critical in terms of G-mean and runtime values: RUS is able to provide the best performance employing a noteworthy shorter runtime. In general, the learning using expert labels yields better results, except for classification of mergers with RUS, which the classification performance is comparable using either expert or amateur labels. This can be due to the inherent randomness of the process in conjunction with the shortage of data in the under-sampling process, which tends to diminish the distinction between amateur and expert labels when the number of examples used in the learning is minimal.

5.4.2 Binary and Multi-class Classification

The experiments carried out with GZ1-(SE) and GZ1-(RLE) datasets entail binary and multi-class classification problems, respectively. In the former case, the CNN is trained to distinguish amongst elliptical and spiral classes. In the latter, it also must discern the handedness of the spiral galaxy's arms. Therefore, the image data employed is identical for these two problems; only the labels are altered. As referred above, we also investigate the influence of the data transformations proposed within the first stage of the thesis, comparing between *Original* (unmodified) and *Transformed* scores, as it is indicated besides the comparative approach designation. This involves the MLP, CzSL, and MLP-SL approaches. The results for binary and multi-class classification problems are presented in Table 5.4 for average Acc values for transductive ($\boldsymbol{U}$) and inductive ($\boldsymbol{TS}$) learning, where the best results in each of the columns are highlighted. Figure 5.4 represents the training times for these experiments, where top and bottom bars represent the MLP input with original and transformed scores, respectively. For CzSL and MLP-SL, the total execution times are detailed in Tables 5.5 and 5.6, respectively, indicating the different phases' average runtimes of such approaches over the ten executions completed with their standard deviation. All values are indicated in seconds.

These results report the following facts:

- Globally, the improvement provided by the CzSL model is greater for the multi-class problem in comparison with the binary classification setting, where there is also a wider margin of improvement. Clearly, MLP-SL and

Approach	GZ1-(SE)		GZ1-(RLE)	
	Acc_U	Acc_{TS}	Acc_U	Acc_{TS}
Expert	0.9222 ± 0.0040	0.9225 ± 0.0047	0.6320 ± 0.0058	0.6367 ± 0.0071
Amateur	0.9162 ± 0.0028	0.9143 ± 0.0044	0.6520 ± 0.0074	0.6468 ± 0.0055
Expert-Amateur	0.9288 ± 0.0033	0.9285 ± 0.0038	0.6595 ± 0.0037	0.6555 ± 0.0033
MLP Original	0.9392 ± 0.0026	0.9381 ± 0.0023	0.6601 ± 0.0022	0.6563 ± 0.0030
MLP Transformed	0.9411 ± 0.0017	0.9402 ± 0.0023	0.6622 ± 0.0054	0.6585 ± 0.0042
CzSL Original	0.9396 ± 0.0025	0.9372 ± 0.0029	0.7334 ± 0.0062	0.7167 ± 0.0061
CzSL Transformed	0.9435 ± 0.0028	0.9405 ± 0.0036	**0.7380 ± 0.0049**	**0.7192 ± 0.0052**
MLP-SL Original	0.9419 ± 0.0016	0.9407 ± 0.0018	0.6735 ± 0.0059	0.6665 ± 0.0057
MLP-SL Transformed	**0.9440 ± 0.0018**	**0.9430 ± 0.0023**	0.6733 ± 0.0065	0.6656 ± 0.0058

TABLE 5.4: Results of all approaches investigated with GZ1-(SE) and GZ1-(RLE) datasets.

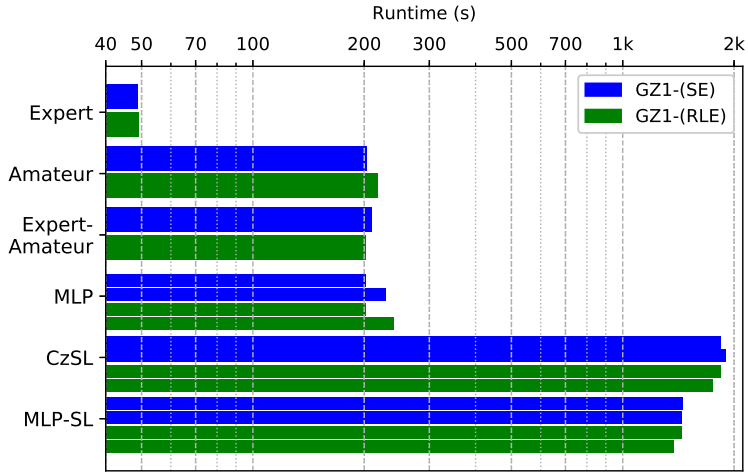


FIGURE 5.4: Execution times in logarithmic scale for GZ1-(SE) and GZ1-(RLE) datasets.

Dataset	MLP input scores	CzSL			
		CAE Training	Pre-training	Fine-tuning	Total
GZ1-(SE)	Original	1,422.3 ± 65.4	207.7 ± 7.7	209.0 ± 7.5	1,839.0 ± 80.6
	Transformed		236.1 ± 3.9	236.2 ± 4.7	1,894.6 ± 74.0
GZ1-(RLE)	Original		208.7 ± 5.3	211.5 ± 4.7	1,842.5 ± 75.4
	Transformed		167.3 ± 9.2	165.0 ± 6.8	1,754.6 ± 81.4

TABLE 5.5: Breakdown of CzSL model runtime for GZ1-(SE) and GZ1-(RLE) datasets.

Dataset	MLP input scores	MLP-SL		
		Training $\boldsymbol{E} \cup \boldsymbol{A}$	Training $\boldsymbol{TR}$	Total
GZ1-(SE)	Original	272.0 ± 3.9	1,185.0 ± 16.5	1,457.0 ± 20.4
	Transformed	264.5 ± 4.1	1,180.5 ± 18.9	1,445.0 ± 23.0
GZ1-(RLE)	Original	268.6 ± 7.1	1,176.8 ± 16.3	1,445.4 ± 23.4
	Transformed	251.0 ± 8.4	1,121.5 ± 15.7	1,372.5 ± 24.1

TABLE 5.6: Breakdown MLP-SL model runtime for GZ1-(SE) and GZ1-(RLE) datasets.

CzSL offer a great advantage in the classification performance, learning from additional knowledge versus the small datasets labelled either by experts or amateurs.

- The MLP input using transformed scores outperforms the use of original scores for all approaches except for MLP-SL approach in GZ1-(RLE), indicating that the score transformations clearly refine the CzSL data in the binary classification case. In fact, the disagreement in the multi-class classification can be explained considering that these transformations were designed to be used in the context of binary classification. Hence, the distinction between right-handed and left-handed spiral classes might not be benefited from the transformations as they were proposed.

- Within the classification of GZ1-(SE), the best performing approach is MLP-SL using transformed scores, which surpass the CzSL in terms of Acc. On the other hand, the MLP approach performs similarly as CzSL, showing that the CNN pre-training is probably less decisive in this problem. It is remarkable how the use of expert labels is able to improve Acc values with respect to the use of amateur labels alone: the inclusion of expert labels boosts the Acc result in 1.42% amongst Amateur and Expert-Amateur approaches, demonstrating that the employment of both label sets results crucial. However, due to the tight improvement margin in this problem, these results lose significance in absolute terms in comparison with the multi-class scenario.

- In the multi-class classification with GZ1-(RLE), expert labels seems to be less informative: the Amateur approach is able to provide better Acc than the Expert, indicating that the quantity of data is probably more relevant that the quality of the labelling for this dataset (also supported by the thinner difference amongst expert and amateur oracles). This case, the CzSL is the best performing approach, greatly surpassing the MLP and MLP-SL models in $\sim$6% of Acc, whose results are similar.

- In general terms, CzSL increases the required runtime in $\sim$22% with respect to MLP-SL, mainly due to the CAE training phase over the whole $\boldsymbol{TR}$ set. As expected, the runtime is increased as more additional knowledge (data) is employed, and same approaches take comparable time either using GZ1-(SE) or GZ1-(RLE) sets, given that both datasets hold the same size. For the rest of approaches, execution times are similar for both classification problems and proportional to the amount of data employed in the training: Expert approach is the fastest (learning from 5% of $\boldsymbol{TR}$),

followed by Amateur, Amateur-Expert and MLP approaches that obtain similar runtime values (25% of ***TR***).

5.4.3 Imbalanced Classification

The experiments with the GZ1-(MG) dataset hold the difficulty inherent to imbalanced classification scenarios, in which the classifier performance tends to be biased towards the majority class. For this reason, we provide two complementary experimental settings using ROS and RUS pre-processing techniques. In this problem, the CNN is trained to differentiate between merger and No-MG objects, thus implementing the same architecture used in the binary classification. These results are displayed in Table 5.7 for average G-mean values for transductive (***U***) and inductive (***TS***) learning, where the best results in each of the columns are highlighted. Figure 5.5 represents the training times for these experiments. Similarly as with GZ1-(SE) and GZ1-(RLE), the total execution times of CzSL and MLP-SL approaches are detailed in Tables 5.8 and 5.9, respectively, indicating the different phases' runtimes along with their standard deviations. All values are indicated in seconds. In this problem there is available original scores only.

Approach	GZ1-(MG)		GZ1-(MG) + ROS		GZ1-(MG) + RUS	
	G-mean$_U$	G-mean$_{TS}$	G-mean$_U$	G-mean$_{TS}$	G-mean$_U$	G-mean$_{TS}$
Expert	0.2103 ± 0.0947	0.2096 ± 0.0973	0.7802 ± 0.0087	0.7834 ± 0.0091	0.6537 ± 0.0735	0.6537 ± 0.0736
Amateur	0.2650 ± 0.1061	0.2624 ± 0.1063	0.7713 ± 0.0045	0.7482 ± 0.0066	0.8137 ± 0.0253	0.8107 ± 0.0226
Expert-Amateur	0.4252 ± 0.0869	0.4188 ± 0.0850	0.7951 ± 0.0130	0.7777 ± 0.0136	0.8182 ± 0.0185	0.8158 ± 0.0194
MLP Original	0.7054 ± 0.0391	0.6957 ± 0.0377	0.8653 ± 0.0040	0.8333 ± 0.0069	0.8479 ± 0.0143	0.8439 ± 0.0141
CzSL Original	**0.7649 ± 0.0188**	**0.7431 ± 0.0217**	0.8508 ± 0.0039	0.8144 ± 0.0049	**0.8693 ± 0.0071**	**0.8631 ± 0.0079**
MLP-SL Original	0.7255 ± 0.0402	0.7053 ± 0.0422	**0.8701 ± 0.0037**	**0.8373 ± 0.0056**	0.8629 ± 0.0073	0.8573 ± 0.0086

TABLE 5.7: Results of all approaches investigated for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques.

The analysis of these results allows us to conclude the following remarks:

- As expected, the inclusion of ROS and RUS techniques has proven to be crucial through both oracles, comparative approaches and CzSL. The classification with CzSL and RUS is able to improve the G-mean over ***TS*** value in 12.0%, providing the best global result in ***TS*** classification. As

Dataset	CzSL			
	CAE Training	Pre-training	Fine-tuning	Total
GZ1-(MG)		230.6 ± 4.7	231.1 ± 6.5	2,173.7 ± 136.0
GZ1-(MG) + ROS	1,712.0 ± 124.8	291.9 ± 11.2	280.0 ± 7.7	2,283.9 ± 143.7
GZ1-(MG) + RUS		13.8 ± 0.6	21.3 ± 0.2	1,747.1 ± 125.6

TABLE 5.8: Breakdown of CzSL model runtime for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques.

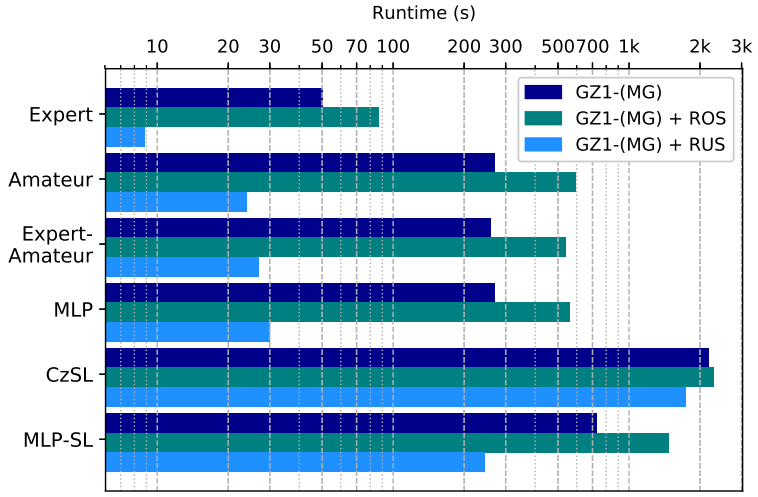


FIGURE 5.5: Graphical representation of experiments' runtime in logarithmic scale for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques.

Dataset	**MLP-SL**		
	Training $\boldsymbol{E} \cup \boldsymbol{A}$	Training $\boldsymbol{TR}$	Total
GZ1-(MG)	143.5 $\pm$ 3.7	584.2 $\pm$ 6.4	727.7 $\pm$ 10.1
GZ1-(MG) + ROS	290.3 $\pm$ 18.6	1,191.0 $\pm$ 44.9	1,481.3 $\pm$ 63.5
GZ1-(MG) + RUS	24.7 $\pm$ 2.0	221.0 $\pm$ 43.6	245.7 $\pm$ 45.6

TABLE 5.9: Breakdown of MLP-SL model runtime for GZ1-(MG) dataset with no pre-processing, ROS and RUS techniques.

occurs in the GZ1-(SE) classification, the best results across the three sub-problems reach comparable performance to oracles, which demonstrate a greatly more advantageous use of the data labelling with the MLP proposed: employing only 5% and 25% of $\boldsymbol{TR}$ data with expert and amateur labels, respectively, CzSL and MLP-SL approaches are able to provide similar performance as the CNN trained over the 100% of the data using expert labels.

- CzSL reports the best performing result for the classification with no pre-processing and RUS. Conversely, MLP-SL approach beats CzSL in the classification with ROS, providing a result equivalent to the one obtained with MLP. This might indicate that the CAE-CNN and CNN pre-trainings included in CzSL lose weight with the over-sampling of the data. Nonetheless, the best global mark is provided by CzSL with RUS, improving the best MLP-SL performance in nearly 3% of G-mean.

- Similarly to the GZ1-(SE) and GZ1-(RLE) classification, CzSL increases the total runtime in comparison to MLP-SL due to the CAE training phase

using the whole ***TR*** set. This time, the difference is magnified with the pre-processing with ROS and RUS. For the rest of approaches, the execution times follow the same pattern within the same approach and globally, always proportional to the amount of data used for training.

5.5 Summary

In this chapter, we have presented the learning paradigm deployed around the completion of CzS projects. This combines the learning from data either labelled by experts and amateurs, and unlabelled data, which represent the three levels of knowledge available and worth considering whether an improved automated classification with CzS data is desired. We have proposed as a unified ML approach the Citizen Science Learning, a unified learning methodology able to take advantage of expert-labelled data, CzS data, and complementary unlabelled data. The CzSL proposed algorithm has demonstrated, through the comparison against a wide selection of comparative approaches and using three different classification problems derived from the GZ1 results, a more integrated use of CzS data and classifications validated by experts in the field of study, which had been attempted with success in the preceding chapter leveraging the well-known pre-training and fine-tuning of CNNs. Going a step further, the solution proposed has enabled the learning from unlabelled data utilising the previously explored CAE, not attempted before in the framework of CzS. In conjunction with these elements, CzSL has based its functionality in the extension of expert knowledge to CzS data using a MLP, which also has ameliorated the problem of generating of amateur labels from amateur scores.

As the final stage of the thesis, the developed methodology also has leveraged the achievements obtained in the previous two stages of the research, thus confirming the global utility of the entire research completed. These have consisted in the improvement of the use of CzS data and the exploration of ML methods for image classification problems under the umbrella of CzS, respectively. First, the proposed data transformations for the handling of uncertainty within CzS data have confirmed their utility, generating the transformed data that has been used as input to the CzSL algorithm and has demonstrated better results in comparison with the employment of GZ1 original scores. Then, the CAE proposed in the comparative study of ML approaches previously developed has resulted in a key component of the CzSL proposed solution along

with the investigated CNN-based approach for the learning from amateurs and experts.

To conclude the dissertation, in the following chapter we outline the final conclusions for the thesis after the main body of research conducted. The contributions of the thesis are emphasised, and we also summarise a set of limitations of the research carried out and highlight several potential lines of future work.

Chapter 6

Conclusions

This chapter presents the final conclusions from the research carried out in this work. First, we outline the set of contributions provided in this doctoral dissertation in Section 6.1. This also includes a set of verification steps in order to adapt the CzSL methodology to other problems, illustrating the high-level approach proposed towards an improved image classification with CzS. Then, in Section 6.2 we point out several limitations that are found after the review of the contributions. Finally, we conclude the chapter with some reflections on potential lines of future work in Section 6.3.

6.1 Summary of Contributions

As it has been discussed in previous chapters, CzS has arrived to the forefront of scientific research as a valuable method for large-scale processing of data in fields such as astronomy or bio-sciences, where novel technologies are generating a flow of data for which a thorough expert analysis is no longer feasible. However, processing this data by amateurs carries several drawbacks that hinder the exploitation of such data, complicating either its utility in research or the employment of these results for the training of automated approaches. It has been shown that there is a need of novel methods to take advantage of CzS data and the existing expert knowledge around the CzS projects under a unified learning framework, also dealing with the inherent uncertainty present within the CzS results.

The research developed as part of this dissertation has contributed towards the improvement of automated approaches for classification problems covered by CzS projects. The experiments have been conducted employing the classification of galaxy images as a case study, the GZ1 project, which still signifies the pioneering implementation of modern CzS on the Internet. A thorough review

of the state-of-art about CzS and the chosen case study have been completed, including related work on ML approaches for this classification problem. After this, three research stages have been developed following a progressive order, in which early findings take part in the final research outcomes. The main contributions deployed through them are revisited in the following paragraphs, where they are also connected to the thesis objectives established at the beginning of this dissertation (Section 1.3). After this, we summarise the key aspects to be checked for the application of the CzSL methodology to other problems (Section 6.1.1).

- The first contribution has addressed the development of a novel methodology for the exploitation of the prevalent uncertainty in CzS data (Chapter 3). The key novelty of the proposed approach lies in the aggregation of information to deal with this uncertainty, also defining three types of uncertainty present in this sort of datasets (Section 3.3.2): the IU, due to the lack of consensus amongst participants in the project; the MU, a direct measure of the uncertainty in every image as a result of the clicks awarded by the DK option on the web; and LC, the uncertainty codified within the distribution of total number of votes achieved by the images. This information has been leveraged by a hybridisation of three mathematical transformations that modify the original scores, employing the available expert classifications for the same data as ground truth, and providing improved classifications for more images with respect to state-of-the-art approaches that used this data (Lintott et al., 2008). In this chapter, it has also been introduced the use of expert classifications for the first time towards the refinement of CzS results themselves and the improvement of their utility in conjunction with ML approaches, which has constituted the main motivation behind the enhancement of CzS data, later confirmed in follow-up experiments (Chapter 5). These results confirm the achievement of the **Objective 1** of the thesis, since the CzS utility has been greatly improved with the support of expert classifications available (expert knowledge) for this classification problem.

- After the refinement of CzS data, a second contribution has encompassed a comprehensive study of ML approaches for image classification, also introducing the use of AEs for FE on galaxy images (Chapter 4). Such comparative study had never been completed for this particular classification problem. It has been analysed the influence of varied aspects in the classification performance of distinct ML strategies under the same datasets, exploring

the image size, use of colour, and distinguishing amongst amateur and expert labels. Additionally, AEs have been proposed for FE of galaxy images (Section 4.2.1), showing a meaningful reduction of runtime and computational cost in comparison with the main state-of-the-art approach (Orlov et al., 2008). As a result, CNNs have been selected as the preferable ML classifier for galaxy images, which together with AEs have enabled the development of the novel classification methodology that leverages amateur and expert label sets in conjunction with unlabelled data in a subsequent research stage (Chapter 5). These achievements have covered the **Objective 2** of the thesis: the most suitable ML approaches for image classification have been explored in the framework of the case study proposed. In addition, fruitful insights have been obtained, which have resulted key for the development of the most general CzSL methodology in subsequent research stages.

- As part of the comparative study of ML approaches that has been conducted (Chapter 4), expert and amateur classifications have been exploited together in the learning of a ML classifier, signifying the first attempt of this *combined* learning process promoted by CzS (Section 4.2.3). This has been developed adopting the well-established pre-training and fine-tuning methods for CNNs, which are usually employed to save computational costs in the use of very deep ANNs. In contrast, we have demonstrated a novel exploitation of these approaches that have shown a smarter use of amateur and expert classifications in comparison with the learning from CzS data only (Section 4.4.2). These contributions encompass **Objective 1** of the thesis, given that the CzS utility has been improved with the use of amateur-based classifications for the training of ML classifiers. Nonetheless, they also anticipate some achievements towards the **Objective 3**, since the final CzSL approach is built upon the aforementioned methodology that has been tested for the training of CNNs using both expert- and amateur-based data.

- The preceding investigation of ML approaches has contributed a solid evidence that supports the selection of the most adequate DL-based models for the elaboration of a novel approach that unifies the learning from amateur- and expert-labelled data, and unlabelled data under the same methodology. Then, a novel learning paradigm has been introduced (Chapter 5) along with a solution that implements the AE and CNN models previously tested (Sections 4.2.1 and 4.2.2). The referred to as CzSL approach (Section 5.2) has embraced a DL-based setting able to exploit all available levels of knowledge generated around CzS projects, constituting the very first automated

solution that has been proposed with these capabilities. CzSL has shown a superior performance when it is has been compared against a set of approaches using three distinct classification scenarios derived from the GZ1 project results, namely: binary, multi-class, and imbalanced classification. In addition, this investigation has also tested the adequacy of the methodology previously proposed for the CzS data improvement (Chapter 3), confirming in a unified set of experiments all contributions included in this doctoral dissertation. These results entirely cover the **Objective 3** of the thesis: a unified approach has been developed to leverage all levels of knowledge available at the completion of a CzS project, thus embracing as well the general aim of the thesis.

In conclusion, it can be stated that the contributions summarised above have thoroughly fulfilled the thesis objectives previously established, and the general aim of the thesis has also been achieved. Indeed, the global output of the conducted research has contributed towards the development of novel automated classification methods for CzS, taking advantage of the three levels of knowledge available around the realisation of CzS projects: expert- and amateur-labelled data, and unlabelled data.

6.1.1 Adaptation of CzSL to Other Problems

The proposed CzSL methodology, in conjunction with the data transformations developed in an earlier stage, hold the necessary generality to be applied to a wide scope of problems covered under the umbrella of the CzS approach. These are crowdsourcing projects involving the classification of images to an arbitrary number of classes, provided that certain conditions are met in terms of both the data being classified and the CzS data resulting from the project finalisation. In the following, we present a detailed checklist that is meant to summarise the main steps for verification in order to adapt the CzSL approach to a new scenario, also displaying the high-level approach provided as a result of the research completed in this thesis.

① In first place, several initial conditions must be assured, as follows:

- We are interested in CzS/crowdsourcing projects dealing with classification of images to a variable number of classes. These can be either actual classes of interest in a given scientific problem (e.g. *Elliptical*, *Spiral*, *Merger*), or a direct measure of the uncertainty in the classification (e.g. *Don't Know*, *Unclear*, *Artefact*, etc.). As it has been

shown, the implementation of CzS projects on the web generally exploits a combination of them (Section 2.2).

- There must be accessible the project results once it has been closed down (*off-line* approach, see Section 2.2): this entails the final count of amateur votes for all the options offered to participants in the project website, properly tabulated, and next to the image ID (as an example, see Table 2.1).
- Expert classifications must be available for some part of the CzS data. This is a first constraint of the CzSL methodology (Section 5.2.2). In other words: only the expert-labelled data that is *overlapped* with the CzS data is usable in the CzSL learning.
- It is desirable the largest amount of unlabelled data, which is used in the last stage of the CzSL learning (Figure 5.1). This aspect will not be a problem in most cases, given that the modern CzS approach itself emerges as a solution to take advantage of the overabundance of automatically generated data in many fields of study.

(2) Having fulfilled the initial conditions referred above, it is a great advantage if the project implementation on the web includes an option that quantifies in some way the uncertainty in amateurs votes. For example, this has been the case of the GZ1 project: the *Don't Know* (DK) votes have enabled the quantification of some type of uncertainty, the Measured Uncertainty (MU), as it has been thoroughly discussed (Section 3.3.1). With this data included within the CzS results, the transformation {2} (*DK Votes Shift*) is able to exploit this information and improve the final classification performance in conjunction with the other two transformations and the hybridisation strategy proposed (Section 3.3.2). Nonetheless, in the absence of the MU, it is always possible to perform the other two transformations with success and test their hybridisation, given that they operate independently. This will depend on the particular classification problem being addressed.

(3) Once the CzS data and expert classifications are ready to use, it is possible to *learn* the data transformations for the problem, on the basis of the requirements detailed above. Employing the expert classifications (expert knowledge) as ground truth, there will be possible to test the set of transformations provided along with all their hybridisation sequences, as it has been explained in the methodology proposed (Section 3.3). To that

end, several aspects must be explored for the classification problem being addressed, namely: **(1)** identification of main classes, **(2)** selection of the favoured class, and **(3)** optimisation of transformations' parameters. In this regard, the experiments developed offer an illustrative example of the application of the methodology to a real classification problem (Section 3.4). This data pre-processing incorporates available information about the uncertainty in amateur classifications, improving the quality of the learning from CzS data by the CzSL method. However, it is worth noting that the CzSL model is able to learn from the unmodified scores as well (this has been shown with the multi-label and imbalanced datasets in the experiments in Chapter 5, indeed). Consequently, this step could be skipped without loss of continuity in the application of the global approach. Nonetheless, it has been widely shown that the data transformations are able to improve the performance of the CzSL learning whenever it has been possible their application (Section 5.4).

④ After the CzS data has been refined with the aggregation of information about the uncertainty in amateur classifications included within the CzS data, the next step is to use this pre-processed data in conjunction with the CzSL approach, leveraging the three levels of knowledge available. To this aim, a preliminary step entails the testing of DL-based models needed for the implementation of CzSL:

- First, there are the CAE and CNN employed in the first and second phases of the CzSL learning, respectively (Figure 5.1). At this point we encounter a second constraint, namely: the CAE and CNN models selected must share the convolutional section of their architectures, that is, the set of layers that learn the features from the image via Convolution layers, Pooling layers, etc. (Section 2.6.2). This is an obvious result derived from the fact that the learning from unlabelled data is implemented using the CNN pre-training with the CAE. Hence, for a proper transfer of weights between CAE and CNN both models must replicate the convolutional section within their topologies (Section 5.2.2). The selection of proper DL models for the problem at hand will depend on certain aspects such as the number of classes defined in the problem, the difficulty in their identification, etc. Previous studies will be necessary to come up with a suitable architecture to be shared by both CAE and CNN.
- After this, there is the MLP used in the last learning phase of the

CzSL for the extension of expert knowledge to amateur-labelled data. First of all, it should be noted that this process is not restricted to the use of an MLP: other classifiers may appropriately do the same job, that is, to *learn* the correlation between amateur scores (either raw or transformed) and expert labels. Therefore, this sub-step of the global CzSL method is open to a wide range of trials in order to find an algorithm that works properly according to the problem being addressed. As an example, in our experiments we develop a grid search between a set of MLP architectures implementing a nested cross-validation (Section 5.2.3).

⑤ Finally, once all the steps detailed above have been completed, the CzSL methodology is ready to be adequately implemented and trained (Figure 5.1). Although the experiments carried out to test the methodology (Section 5.3) originally kept fixed the size of the different labels sets (the so-called $\boldsymbol{E}$, $\boldsymbol{A}$, and $\boldsymbol{U}$ datasets), the extension of the method to other classification problems obviously will overcome this particular experimental setting. The more data available will most likely improve the performance of the global approach. This will depend, once more, on the problem covered and the availability of labelled and unlabelled data.

These progressive steps encompass all required aspects worthy of attention for a proper adaptation of the CzSL methodology to other classification problems developed in the framework of CzS or crowdsourcing initiatives. As a result, the global contribution of the thesis has been summarised in technical terms, showing the high-level approach provided towards an enhanced automated image classification with CzS.

6.2 Limitations

In relation to the contributions summarised above, the research completed in this thesis also presents limitations that need to be mentioned. In this section, we highlight some aspects that would reinforce the research conducted and may be part of future work.

- The experiments in which standard classifiers and DL-based models have been employed lack an adequate optimisation of these algorithms' parameters. This is the case of the comparative study of ML approaches for image classification developed in the second research stage (Chapter 4). As it has been acknowledged in the experimental setup, the parameters selected

were a consequence of the review of some relevant work that deploys similar comparative studies amongst such classifiers or implements DL architectures that are comparable to those proposed for the experiments (Sections 4.3.1 and 4.3.3). On the other hand, a parameter fine-tuning is not relevant for the results, since its absence might equally affect all models involved and it was not the target of the study. Nonetheless, whenever a further refinement of the experimental results is desired, a hyper-parameter tuning of these models could greatly boost the confidence in the results and provide new insights on the comparative study carried out.

- The proposed method for the enhancement of CzS data (Chapter 3) has leveraged distinct types of uncertainty present within the CzS data itself. However, it is worth noting that for this research it has only been employed the GZ1 data that was made public after the project closure (Section 2.4.1). This data includes the final counts of amateur votes for the classes shown. No further information about participants performance or related aspects about the timing distribution of the classifications on the web are provided. As a result, all votes are awarded the same confidence. However, if we were provided with the information about who voted what, it could be feasible to strengthen the proposed mechanisms to handle the uncertainty taking into consideration this increased granularity within the CzS data. For instance, it has been shown that participants tend to develop their skills as they accumulate more experience in the task (Sprinks et al., 2017; Trouille, Lintott, and Fortson, 2019). At this respect, the consideration of additional data about participants' performance could considerably extend our investigations and establish new frameworks for the refinement of CzS data either for research or its utilisation in automated approaches.

- The proposed CzSL learning algorithm is tested using different classification problems encountered within GZ1 results. Besides the binary classification problem, which is the most common in the research, we carry out the experiments to put such algorithm under stress introducing the multi-class and imbalanced datasets. However, these two additional classification scenarios have been treated adopting standard procedures, given that the aim has been to test the robustness of the CzSL approach under more complex classification scenarios. For example, the study of the multi-class classification could be complemented investigating different strategies to tackle these problems such as one-vs-all or one-vs-one (Galar et al., 2011) and ensemble-learning, or examining the detailed performance per class.

On the other hand, the classification of mergers (imbalanced classification) has been tackled using standard over-sampling and under-sampling pre-processing techniques. The inclusion of more complex approaches, such as the use of Generative Adversarial Networks (GANs) (Douzas and Bacao, 2018) for the over-sampling, or data augmentation procedures as have been proposed in Dieleman, Willett, and Dambre (2015), may bring new results that could extend the range of applicability of the CzSL approach.

- This thesis has focused on image classification problems, as the main task generally demanded by CzS projects. The GZ1 website clearly symbolises this framework: one image is displayed, along with a few options to decide, and one single click to be allocated. However, other problems that are nowadays being covered by the CzS approach do not necessarily adjust to this scheme. For instance, projects tackling the identification of patterns in graphical data (Zevin et al., 2017), or dependent on tasks involving counting or measuring (Silverberg et al., 2018) would generate data of a very different nature. This might lead to several modifications and adjustments of the methodologies that have been proposed.

- The entire conducted research has employed as case study the classification of galaxy images, as the classification problem holding the largest corpus of results and research works provided by the CzS paradigm for a unified scientific problem (Fortson et al., 2012). As an advantage, this has enabled the easy accessibility to the image data and classifications needed for the experiments, making possible an aggregation of the whole set of insights obtained in this research that may cause a positive impact on this particular problem. However, the application of CzS in recent years has been extended to varied research fields ranging from bio-sciences (Reis et al., 2015) to the study of climate change (Knapp et al., 2016) or distinct topics in history (Barber, 2018; Blickhan et al., 2019; Eckstein, 2019). The inclusion of these projects' results may require the adaptation of the methods developed under the selected case study.

These limitations in the research conducted in this thesis represent a set of improvements that may suggest immediate lines of work. In the next section, we relate our research with current areas of investigation that could entail its continuation or potential application of the insights developed in the preceding chapters.

6.3 Future Work

The research conducted in this thesis has aimed to bridge current gaps in the intersection between three main areas of study, namely: CzS, ML, and astroinformatics. The conceived methods and investigations carried out could support future research in such areas, and also extend their applicability to other fields that could be influenced by the realisation of CzS projects. In the following, we remark several ideas of future work that derive from the completed research and could open new directions.

- As an immediate direction for future work, it is worth considering the enhancement of the methodologies developed throughout the research stages completed. In particular, the optimisation of parameters, which has involved several experiments, offers a wide margin of improvement. This aspect has also been acknowledged above as an inner limitation of the experiments carried out as part of the comparative study of ML approaches for image classification performed (Chapter 4). However, in the case of the method proposed for the refinement of CzS data (Chapter 3), the improvement of the suggested procedure for the optimisation of transformations' parameters (a grid search) could considerably enrich the whole methodology, globally depicted in Figure 3.2. These parameters play a fundamental role in the effectiveness of such data transformations, as the analysis of results unveils. Consequently, the development of a upgraded approach for this purpose may have a positive impact on the performance of this methodology, and thus extend the outreach of the entire research.

- Currently, one of the most prominent areas of investigation around the development of CzS projects addresses the engagement of participants along with their performance on the task demanded. Many works have focused on several factors that could affect the success of CzS in the long run such as the study of participation (Herodotou and Robinson, 2020), the different motivations that lead *citizen scientists* (Jackson et al., 2015; Lee et al., 2018), or the best design of projects to enrich the volunteer experience and enhance the CzS outreach (Spiers et al., 2019; Sprinks et al., 2019). Another aspect that has proven to influence amateurs' performance is user fatigue, widely studied in the literature (Jiménez, Gallardo, and Moller, 2018; Jungbluth, Yeng, and Vives, 2018). It has been demonstrated to be a key component having an impact on the quality of amateur classifications, which could be counteracted with different approaches such as the offer of different forms

of rewards to keep participants motivated during the task, or the development of user-friendly and intuitive designs for the web environments. In these investigations, the management of the uncertainty introduced by amateurs can be handled prior to the generation of the CzS data, given that it highly depends on the way the problem is presented and taught to the project participants. Hence, the research on these variables exhibits several areas of improvement that could uphold and extend the insights presented here.

- The development of improved learning systems in the domain of on-line approaches for CzS is capturing a lot of attention in recent works. These approaches tackle varied aspects such as the combined learning from experts and amateurs during the course of the project (Walmsley et al., 2020), or the training of volunteers using ML techniques towards the enhancement of the quality of their contributions (Crowston et al., 2020; Jackson et al., 2020). These advances clearly present a wide range of improvement in the combined action of ML, expert knowledge, and amateur endeavours at the time the project is running. In this sense, the extension of the CzSL method or some of its stages to embrace the on-line learning poses a great challenge for future work.

- In line with the limitation mentioned above about the use of advanced preprocessing techniques in the study of galaxy mergers, many works are investigating the use of GANs for the over-sampling in imbalanced classification problems (Karras et al., 2018; Wang et al., 2019b; Mansourifar, Chen, and Shi, 2019). However, these approaches could benefit from the inclusion of amateur and expert labels, which would increase the amount of labelled data available for the training of these ANNs. This scenario suggests a novel work direction in which the extension of expert knowledge to amateur-labelled data could promote the generation of new images within the galaxy classification problem in distinct environments such as the imbalanced classification of mergers, simulation of galaxy evolution models, generation of morphological sequences, etc.

- Many current research topics are affected by the scarcity of high-quality data for the training of efficient and robust automated approaches. With the advent of internet facilities worldwide and the dissemination of smartphones, wearables, internet of things, the world is becoming very rich in data but poor in terms of extracted knowledge and profit from this data. The current CzS paradigm offers a set of tools that could contribute to ameliorate

this imbalance, and several efforts are taking place in this direction (Zeng et al., 2020; Starkey et al., 2017). For instance, the current research around the investigation of Covid-19 pandemic could benefit from the large-scale generation of labelled data coordinated via CzS-based environments across the web. The research included here may help as a first attempt towards the inclusion of professional science in the daily life of general people, inspiring the development of novel methods to exploit all kind of labelled data to either assist fundamental research or reinforce the learning of automated approaches.

Appendix A

Galaxy Zoo 1: Data Exploratory Analysis

A.1 Introduction

In this appendix, we aim to provide a deeper understanding of the GZ1-T2 dataset, along with the process by which it was created: the amateurs' clicks on the GZ1 website[1]. In the following sections, we focus on these main points of interest:

- Study of the scattering of galaxies across the two main scores, ***Elliptical*** and ***Spiral***, as well as across the total number of votes received by the examples. This first understanding will tell us about the mass of galaxies prone to an improved classification, since the great part of them are eventually labelled as ***uncertain***. On the other hand, this information will also inform about the potentiality behind the addition of more information from the total number of votes received by each galaxy, in order to perform more consistent and accurate classifications.
- Study of the debiasing process developed by Bamford et al. (2009), how it affects the population of galaxies, and how galaxies are classified by the GZ1 team using these debiased scores. Whereas it was supposed that classifications were performed applying a threshold (Lintott et al., 2011), the published data shows several disagreements.
- Evaluation of the GZ1 classifications. To this aim, two expert catalogues are introduced, conforming the *expert knowledge* on the galaxy classification problem; GZ1 classifications coming from amateurs' votes, on the other side, represent the *amateur knowledge*.

[1] http://zoo1.galaxyzoo.org.

- The aforementioned expert knowledge also allows us to go further and evaluate other classifications generated using different thresholds, either over original or debiased scores. In addition, we propose the normalisation of *Elliptical* and *Spiral* scores, which introduces two more score types into play. This will constitute part of the data transformations developed through the research (Chapter 3).

The data considered through this exploratory analysis corresponds to the so-called GZ1 Table 2 of the GZ1 data release[2]. This dataset contains identification number (ID), total number of votes, scores across six categories[3], *Combined Spiral* score[4], debiased *Elliptical* and *Spiral* scores, and final classifications for a population of 667,944 galaxies.

The rest of the appendix is structured as follows. Section A.2 presents two set of histograms on the distribution of galaxies across *Elliptical* and *Spiral* scores, and total number of votes received. Then, amateur classifications are examined in Section A.3 and expert classifications are introduced in Section A.4, which serve to perform an evaluation of the GZ1 classifications by applying different thresholds as well as an assessment of the proposed scores normalisation in Section A.5. Finally, we summarise the contents of this appendix in Section A.6.

A.2 Distribution of the data

First, in Section A.2.1, we intend to shed light over the scattering within the population of galaxies across the two main morphological classes, *Elliptical* and *Spiral*, along with the two score types, *original* and *debiased*. Then, in Section A.2.2 we examine the spread of the population across total number of votes received by each galaxy image. The former allows us to get an intuition about the transformation of the scores after the debiasing procedure (Bamford et al., 2009), whereas the latter tells us about a confidence margin held by the data and not previously used.

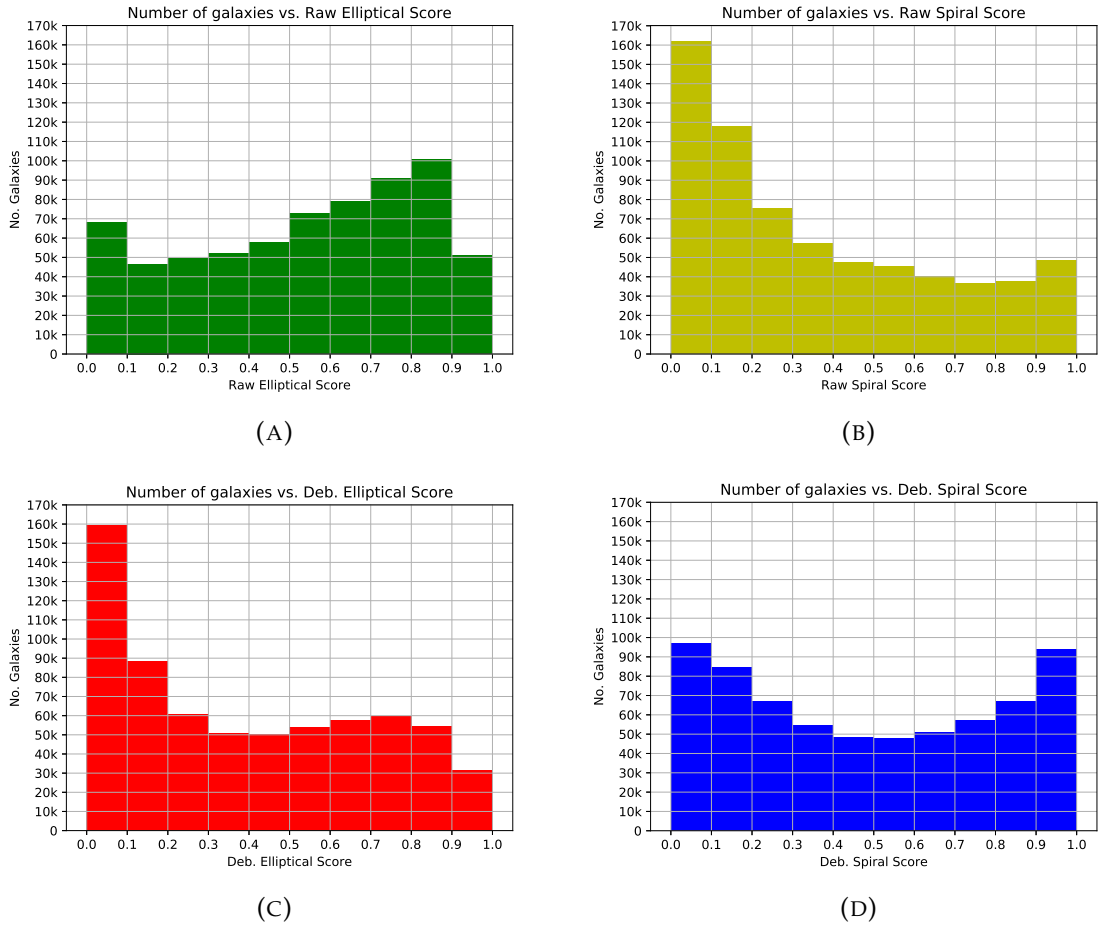
[2]This data is publicly available at `http://data.galaxyzoo.org`.

[3]The six original categories people were asked to choose between are: *Elliptical*, *Clockwise Spiral*, *Anticlockwise Spiral*, *Edge-on Spiral*, *Don't Know* and *Merger*.

[4]The three original spiral scores are eventually added by the GZ1 team to form a *Combined Spiral* score, which we will refer to from here on as *Spiral* score.

A.2.1 Original and Debiased Scores

This first series of histograms present the distribution of galaxies across the four score types included in the GZ1-T2 data. Figure A.1 represents total number of galaxies per bin of 0.1 score width, for original *Elliptical* (A), original *Spiral* (B), debiased *Elliptical* (C) and debiased *Spiral* (D) scores. Each bin includes galaxies from the low mark and just below the upper mark. For instance, second bin contains galaxies with score in the interval [0.1,0.2). In order to get a better understanding on the spreading of the data, same histograms are also presented in cumulative form in Figure A.2, and basic statistics of each score distribution are provided in Table A.1.



(A) (B)

(C) (D)

FIGURE A.1: Number of galaxies across original and debiased score types.

An inspection of histograms and tabulated data above draws these characteristics:

– There is an evident imbalance between *Elliptical* and *Spiral* original scores

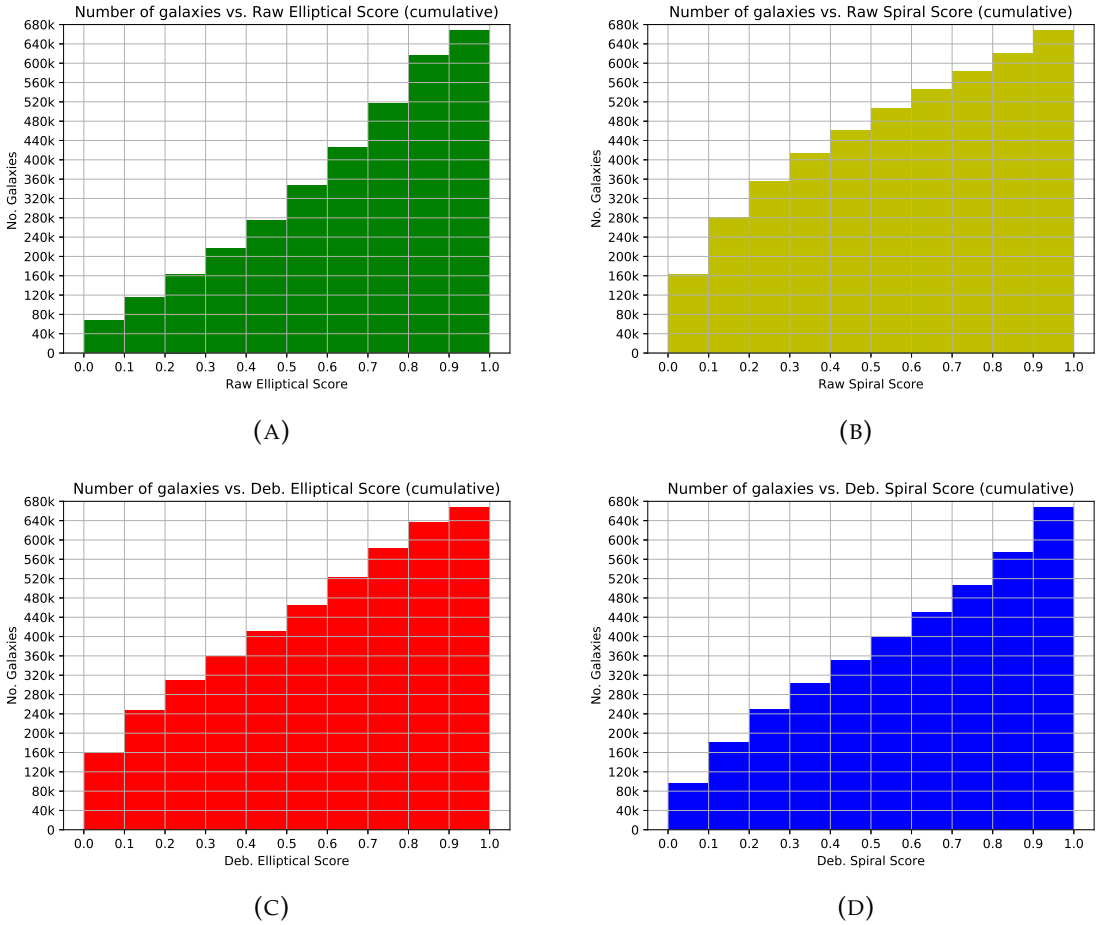


FIGURE A.2: Number of galaxies across original and debiased score types (cumulative).

Score type	Average	Std. Deviation	Coeff. of Variation
Original *Elliptical*	0.536	0.284	53.05%
Original *Spiral*	0.360	0.299	82.99%
Debiased *Elliptical*	0.394	0.302	76.64%
Debiased *Spiral*	0.483	0.322	66.55%

TABLE A.1: Basic statistics for original and debiased score types.

(Figures A.1a and A.1b), by which the population of *Elliptical* highly-scored galaxies is remarkably larger previous to the debiasing. On the contrary, original *Spiral* scores tend to concentrate towards the left of the chart, supporting that tendency by which non well-resolved images are more likely to be classified by amateurs as *Elliptical*. However, objects in the (0.9-1.0) bin are approximately equal for both categories ($\sim$50k). This is an interesting symmetry between these classes, which is openly broken after the debiasing (Figures A.1c and A.1d). It indicates that the number of exemplars with great consensus were similar after the count of votes, and not skewed towards one of the classes.

– The histograms offer a more intuitive comprehension of the debiasing net effect: it partially corrects the imbalance between *Elliptical* and *Spiral* original scores. As explained in Lintott et al. (2011), *"The overall result of applying the correction is to lower the probability that a given galaxy will be classified as early type and increase the chance that it will be classified as spiral"*. Following this trend, the number of *Elliptical* low-rated galaxies (Figure A.1c) is widely increased, whereas the distribution of the debiased *Spiral* score (Figure A.1d) is almost balanced, with two maxima around ∼90k galaxies at both ends of the chart.

A.2.2 Total Number of Votes

In this section, we present a different set of histograms on the distribution of galaxies across the total number of votes, one of the features included in the GZ1-T2 dataset. The first two plots, Figures A.3 and A.4, represent the same histogram in lineal and logarithmic scales, respectively. In these charts, each of the bins includes galaxies which received a number of votes greater or equal than the low mark and just smaller the upper mark. For instance, the (20-25) bin contains galaxies with total number of votes in the interval [20,24). However, since the variability through this feature compared with the number of galaxies per bin is extremely large, it makes both tails of the lineal histogram to vanish, as it is shown in Figure A.3. Thus, using logarithmic scale, the two ends of the plot are thoroughly visible in Figure A.4. Next, a cumulative histogram presented in Figure A.5, which summarises the total count of galaxies up to a particular number of votes. In this chart, the value of the bin is explicitly indicated on top: each bin includes the galaxies which received a number of votes greater or equal than the low mark and just smaller the upper mark. For instance, the (20-25) bin contains galaxies with total number of votes greater or equal to 20, and smaller or equal to 24. After this, Table A.2 finally presents basic statistics of this feature over the entire dataset.

Average	**Std. Deviation**	**Median**	**Coeff. of Variation**
38.76	13.83	34.00	35.67%

TABLE A.2: Basic statistics for total number of votes received by each galaxy image.

After an inspection of the three histograms and statistical information presented above, we draw the following remarks:

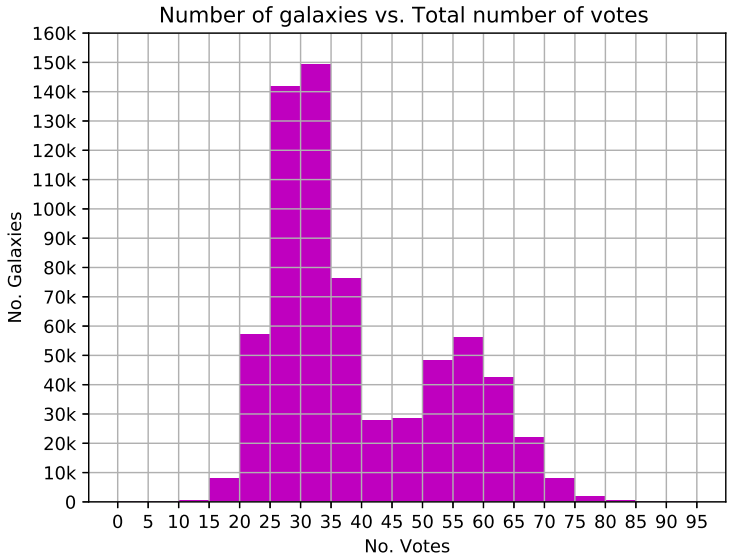


FIGURE A.3: Number of galaxies across the total number of votes received.

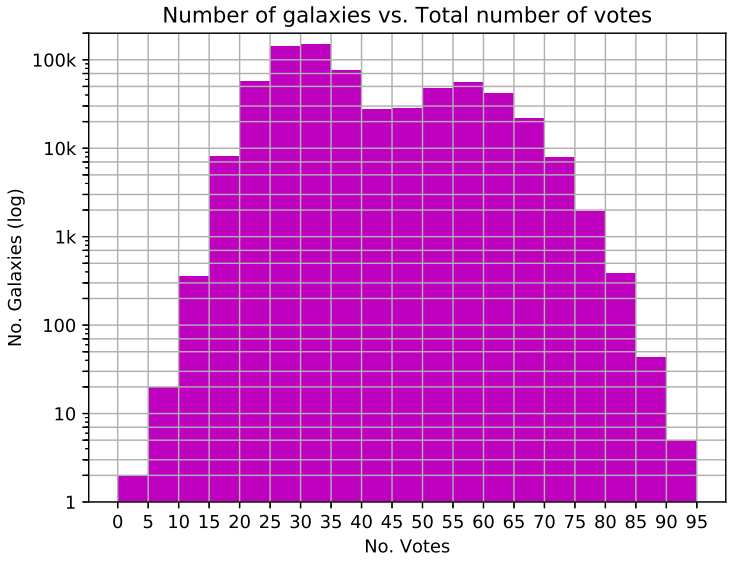


FIGURE A.4: Number of galaxies across the total number of votes received in logarithmic scale.

– The representation of total number of votes (Figure A.3) shows and intriguing feature: the scattering of galaxies through the number of votes does not follow a normal distribution, as it is shown in Figure A.6, where a Normal distribution of equal μ and σ is represented. On the contrary, the plot clearly presents a double peek, which is still visible in logarithmic scale (Figure A.4). Reasons for this peculiar shape may derive from a random pattern followed in the collection of votes, since the number of

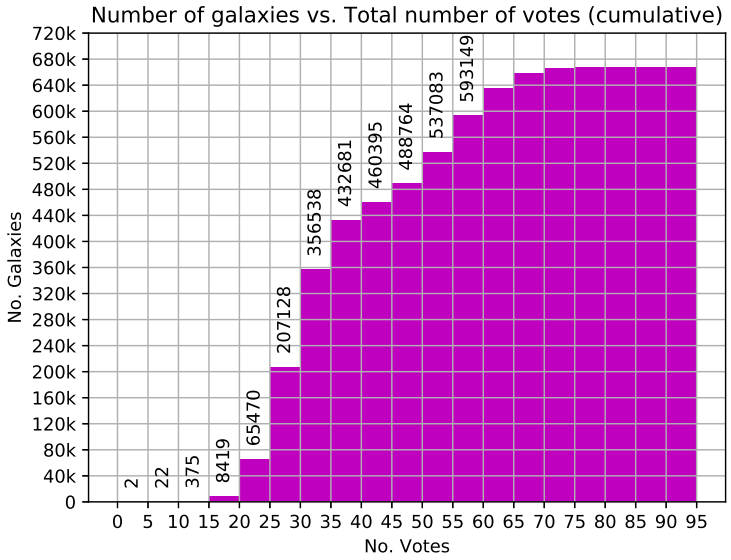


FIGURE A.5: Number of galaxies across the total number of votes received (cumulative).

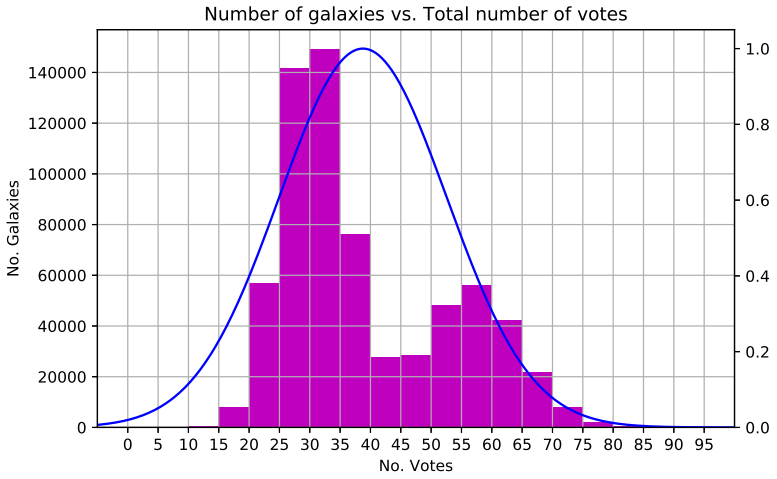


FIGURE A.6: Number of galaxies across total number of votes received with a Normal distribution overlapped.

galaxies with a mean number of votes (Table A.2) strongly drops. As a consequence, the median number of votes is a more adequate indicator as a reference level.

– Since the number of votes is highly uneven through the population of galaxies, this treat of each classified image hides a different level of confidence, which has been openly neglected thus far. Galaxies in the (30-35) bin (Figure A.3) should not be equally analysed as the exemplars in the (60-65) bin. This variability through the data, now confirmed, calls for

the inclusion of this information into the galaxy final classification, as a sort of confidence level.

A.3 GZ1 Classifications

The data analysed through this appendix was released by the GZ1 team once the project finished collecting votes in February 2009. After the count of amateurs' votes and the application of bias correction developed by Bamford et al. (2009), some sort of 0.8 threshold was computed over the debiased scores in order to assign final classifications to galaxies: objects holding 0.8 or greater debiased *Elliptical* or *Spiral* score were classified as such. Galaxies not reaching this mark in neither of the debiased scores were grouped as *uncertain*[5] (Lintott et al., 2011). From now on we will refer to these classifications performed by the GZ1 team as *GZ1 flags*.

However, the process conducted to award classifications seems to be more complex. The resulting GZ1 flags do not match the exact application of a 0.8 cut over the debiased scores. The disagreement is presented in Table A.3.

	Elliptical	***Spiral***	***Uncertain***
GZ1 flags	62,190	190,225	415,529
Real 0.8 threshold	86,102	160,766	421,076

TABLE A.3: Disagreement between GZ1 flags and a 0.8 threshold over debiased scores.

As it is presented, a noticeable amount of *elliptical* galaxies "disappear" when the threshold is applied. The opposite trend is found through the *Spiral* category: a great part of *Spiral*-flagged galaxies do not reach a 0.8 debiased *Spiral* score. In order to disentangle this disagreement, we compute the GZ1 flags of these objects, for both *Elliptical* and *Spiral* categories. We also present a singular kind of objects, which we will refer to from now on as *prototypes*: galaxies with 1.0 score. They are a total of 3,956 ellipticals and 11,673 spirals. Holding a 100% of consensus between amateurs who voted that galaxy, they embody the set with maximum confidence within the *amateur knowledge*. In addition, as zeros and ones are not changed through the debiasing procedure, they work as a bridge between original and debiased scores: their confidence is common to both sets (Tables A.4 and A.5). The counts are cumulative, i.e., the objects

[5]Strictly speaking, *uncertain* is not a classification category, since it gathers the undecidable objects.

included in second rows are contained in the row just below, and so forth.

Deb. *Elliptical* Score	Elliptical	Spiral	Uncertain	Total
1.0	**3,956**	0	0	3,956
$\geq$0.9	29,159	4	2,217	31,380
$\geq$0.8	57,755	165	28,182	86,102
$<$0.8	4,435	–	–	–

TABLE A.4: Count of galaxies as classified by the GZ1 flags with respect to the application of different thresholds over the debiased *Elliptical* score.

Deb. *Spiral* Score	Elliptical	Spiral	Uncertain	Total
1.0	0	**11,673**	0	11,673
$\geq$0.9	0	92,378	1,444	93,822
$\geq$0.8	0	147,138	13,628	160,766
$<$0.8	–	43,087	–	–

TABLE A.5: Count of galaxies as classified by the GZ1 flags with respect to the application of different thresholds over the debiased *Spiral* score.

Therefore, the assignment of flags clearly does not follow a strict threshold. Part of the galaxies scored over 0.8 are labelled as *uncertain*, for both categories. Indeed, 165 *Spiral*-flagged galaxies possess debiased *Elliptical* score greater than 0.8 (Table A.4, third column). Another different criteria is supposed to be applied over the debiased scores, so that the GZ1 flags could be derived from them.

A.4 Expert Classifications

Two expert catalogues were firstly used by the GZ1 team in order to evaluate amateurs' performance in the classifications (Lintott et al., 2008). The MOSES[6] catalogue (Schawinski et al., 2007) contains a total of 16,520 galaxies classified by expert astronomers as *Elliptical*, of which 16,516 are included in the GZ1-T2 dataset. On the other hand, the Longo catalogue (Longo, 2011) is a set of 25,561 galaxies classified by expert astronomers as *Spiral*, of which 25,190 are present in the GZ1-T2 as well.

However, it is remarkable the fact that between these two expert catalogues there is an overlapping of 141 galaxies, which have been removed from the

[6] *MOSES* is acronym of **MO**rphologically **S**elected **E**lliptical in **S**DSS.

whole dataset for consistency. After this correction, we consider the joint catalogue out of MOSES and Longo as ground truth on the classifications, which allows us to perform an evaluation of GZ1 classifications. These results are presented in Tables A.6 and A.7.

	MOSES	**Longo**	**Total**
In GZ1 dataset	16,375	25,049	41,424
Correctly flagged	4,181	20,385	24,566
Incorrectly flagged	1,040	26	1,066
Flagged as *Uncertain*	11,154	4,638	15,792

TABLE A.6: Comparison between GZ1 flags and expert classifications from MOSES and Longo catalogues.

True Elliptical	4,181
True Spiral	20,385
False Elliptical	26
False Spiral	1,040
Not classified	15,792
Total	41,424

TABLE A.7: Evaluation of GZ1 flags considering the joint catalogue from MOSES and Longo as ground truth.

By this terminology proposed here, *True Elliptical* refers to objects flagged as elliptical and contained in MOSES catalogue; *True Spiral* corresponds to objects flagged as spiral and contained in Longo catalogue; *False Elliptical* refers to elliptical-flagged galaxies in Longo, and *False Spiral* alludes to spiral-flagged galaxies which are present in MOSES; *Not classified* brings together galaxies in either MOSES or Longo, and flagged as *uncertain*.

These results allow us to calculate classification Acc and RR measures (Section 3.4.2), as a form of evaluation of GZ1 flags. Acc informs about the proportion of correctly classifications with respect to the total number of classified objects. The RR, instead, is calculated as the proportion of non-classified objects with respect to the total. These values, indicated in Table A.8, will serve as a baseline for subsequent classifications with different thresholds, as it is presented in the following section.

A.5 Evaluation of Different Thresholds

The joint expert catalogue obtained after the combination of both expert catalogues, after the removal of overlapping galaxies, conforms a sort of *ground*

GZ1 classifications	
Accuracy	0.9584
Rejection Rate	0.3812

TABLE A.8: Acc and RR values for GZ1 classifications (GZ1 flags) considering the joint catalogue from MOSES and Longo as ground truth.

truth which allows for the evaluation of other classifications made out of the GZ1 original scores. In this section, we present Acc and RR values for different classifications performed by application of several thresholds, ranging from 0.5 up to 1.0 in 0.1 steps. In turn, we consider up to four score types: original and debiased scores, as presented in GZ1-T2 dataset, and these same scores after a normalisation process. This consists in re-scaling the *elliptical* and *spiral* original scores, dividing each value by their addition so after that they equal 1.

The restriction to *elliptical* and *spiral* classes after the computation of the votes, clearly calls for a re-normalisation of the scores. In other case, the proportional part of the score coming from *Merger* and *Don't Know* classes is clearly neglected, since these two scores are residual through galaxies evaluated by the joint expert catalogue[7]. Moreover, the absence of a normalisation causes that *elliptical* and *spiral* scores frequently do not add to one, which occurs to more than 550,000 galaxies. As a result, the normalisation process also promotes the emergence of a lot of new prototypes, from systems that we refer to as ***anti-Prototypes***: galaxies with *elliptical* or *spiral* score equal to zero. For these galaxies, we do not have a whole certainty about pertaining to one class; however, none of the amateurs voting that galaxy did include it in the opposite class. The resulting number of old prototypes and new ones from anti-prototypes is summarised in Table A.9, detailing the number of prototypes in the original GZ1-T2 data (left column), after the normalisation process (center column), and within the joint expert catalogue (right column).

	GZ1 prototypes	anti-Prototypes	Expert
Elliptical	3,956	19,972	1,705
Spiral	11,673	9,519	8,164

TABLE A.9: Number of prototypes.

This comparison between the four score types available is shown in Figure A.7, where the GZ1 marks (Table A.8) are represented by the magenta star. This

[7] As both expert catalogues are made of a set of clear examples of each morphology, it is expected that *Don't Know* and *Merger* scores are pretty small for these objects.

demonstrates a notable improvement in the GZ1 classifications: the RR value is able to reduce to 0.0 with a 0.5 threshold. The current baseline is broadly improved by the normalised debiased scores, as a good trade-off is reached with a threshold in the interval (0.7-0.8).

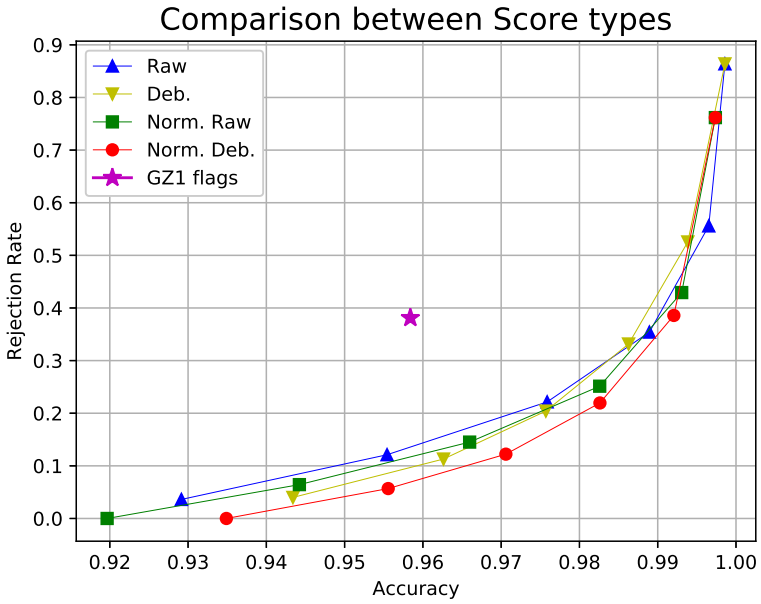


FIGURE A.7: Comparative representation of Acc and RR achieved by each score type across a selected set of thresholds.

A.6 Summary

In this appendix we have presented a set of statistical analyses of the data included in the GZ1-T2 data. We have acquired more intuition on the distribution of galaxies across the two main scores, *elliptical* and *spiral*, in the two types provided within the dataset, as well as across the total number of votes received by each galaxy. These two studies have shown a more precise knowledge about the debiasing process conducted by the GZ1 team, as detailed in Bamford et al. (2009), and the great variability in the total number of votes, which confirms a potential useful usage of this information not previously considered. On the other hand, the joint expert catalogue has allowed us to perform a first evaluation of the GZ1 flags, and other classifications extracted from different thresholds imposed over the original and debiased scores.

Appendix B

Fast and Robust Fuzzy-kNN Classification with CzS Data

B.1 Introduction

CzS is becoming mainstream in a wide variety of real-world applications across different research domains in which, for example, classification tasks by experts are very time-consuming. These projects engage amateur volunteers that are tasked to manually classify unlabelled images. As a result, we obtain a larger volume of labelled data that, however, contains a great level of uncertainty due to the wide range of expertise of the volunteers. Handling that inherent uncertainty is key to building robust and fast ML models that maximise the outcome of CzS projects. In this research, we propose a novel approach that first leverage the transformation of the original results previously developed (Chapter 3), pursuing to handle the prevalent uncertainty prevalent in CzS data, and then uses this transformed data as input to the fuzzy k-nearest neighbours (Fuzzy-kNN) classifier. We propose the utilisation of CzS results in such a way that it naturally speeds up the learning and classification phases of the mentioned fuzzy classifier and improves the classification performance. The experimental results with the GZ1 data show that an appropriate use of CzS data enables a faster and more robust classification with a fuzzy-based classifier, also confirming the advantages of using the transformed data versus the original CzS results.

The appendix is structured as follows. First, in Section B.2 we provide some preliminaries about the management of uncertainty in CzS data and the selected fuzzy classifier used in the experiments, the Fuzzy-kNN algorithm. Then, Section B.3 introduces the proposed methodology for the exploitation of CzS data in conjunction with the Fuzzy-kNN classifier. In Section B.4, we describe

the experiments carried out and discuss the obtained results. Finally, we summarise the achievements of this research in Section B.5.

B.2 Preliminaries

In this introductory section, we cover the necessary background to understand the proposed classification approach. First, Section B.2.1 refreshes some related work about the uncertainty within CzS projects. Then, Section B.2.2 briefly introduces the Fuzzy-kNN algorithm, which is employed through the proposed methodology.

B.2.1 Uncertainty in CzS Data

As it is referred within Section 2.2, CzS projects usually involve a great number of volunteers to help with diverse tasks that are typically tedious and very time-consuming for experts to be conducted. Many of these projects have consisted in the classification of large collections of images. Amateurs are asked to classify the images displayed in the project website by choosing amongst a set of pre-defined categories that involve some classes of interest in the research covered by the project. When the project is finalised, these amateur classifications are recorded in the form of a count of votes for each of the classes (e.g. image 1 has got 24 votes for class x, and 3 votes for class y). An example of such CzS data is shown in Table B.1.

Image ID	*Votes*	*Class 1*	*Class 2*	$\cdots$	*Don't Know*
0152948451	58	0.310	0.414	$\cdots$	0.052
0152863349	14	0.643	0.214	$\cdots$	0.071
0152878152	33	0.000	1.000	$\cdots$	0.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0152721030	19	0.316	0.263	$\cdots$	0.263

TABLE B.1: Typical look of the data recorded in the course of a CzS project involving image classification.

ML techniques have been used in the context of CzS, taking the form of either on-line and off-line approaches. For instance, on-line approaches deal with the training of participants as they gain expertise in the task and the synergy between amateurs, experts and automated classifiers (Crowston, Osterlund, and Lee, 2017; Zevin et al., 2017). However, most efforts have been devoted to off-line approaches more focused on analysing the data resulting from CzS projects. Various works have aimed to replicating amateur classification skills.

This has been normally achieved by firstly applying a FE algorithm from the images (Banerji et al., 2010; Kuminski et al., 2014), or using the image directly by means of a DL-based approach (Dieleman, Willett, and Dambre, 2015). The current literature has shown that ML classifiers can achieve similar results to those obtained by amateurs, when these algorithms are trained using CzS data (Beaumont et al., 2014). However, those approaches do not tackle the uncertainty prevalent in the results, which considering the varied amateurs' skills may be very diverse. Consequently, they usually end up replicating the biases and various forms of uncertainty present in the data.

In Chapter 3, we propose a sequence of fuzzy-based data transformations that take the original amateurs' count of votes and mitigate this uncertainty. Through various aggregation mechanisms, the proposed approach is able to deal with different types of uncertainty present in CzS data such as the IU, due to the intrinsic lack of agreement between participants, MU, which is derived from the clicks on the DK option, etc. As a result, we obtain an enriched version of the data that is evaluated by considering available expert classifications for the data, showing an improved classification Acc and enabling the classification of a vast number of examples previously left as *uncertain*.

B.2.2 Fuzzy k-Nearest Neighbour Classification

In this research, we use the Fuzzy-kNN algorithm (Keller and Givens, 1985) as a simple yet effective fuzzy-based classifier. The Fuzzy-kNN algorithm was devised to improve upon the standard kNN classifier (Section 4.3.1). The main weakness of kNN resides in considering all neighbours as equally important in the classification, making the kNN algorithm more vulnerable to noise at the class boundaries. This algorithm has shown to perform well in a wide variety of applications, and more important, it is capable of considering uncertainty when predicting the class label of a new example.

The Fuzzy-kNN usually performs a learning phase, in which the training set is modified to compute the class memberships of every single training instance. This stage consists of calculating the k nearest neighbours for each instance of the training set against the training set (using a leave-one-out scheme), and selecting those k instances holding the shortest distance. Class memberships for each training instance x is then calculated following Equation (B.1), where n_j is the number of examples of class j within the k_{memb} nearest neighbours computed.

$$u_j(x) = \begin{cases} 0.51 + (n_j/k_{memb}) \cdot 0.49 & if \quad j = i \\ \\ (n_j/k_{memb}) \cdot 0.49 & if \quad j \neq i \end{cases} \tag{B.1}$$

The result of this first stage will be a modified training set in which the class label ω of each sample is replaced by a membership vector (ω_1, ω_2, ..., ω_l), where l is the number of classes. For each instance of the test set, the classification stage calculates its k nearest neighbours in the modified training data (with class memberships). Thus, it obtains the membership vector of each neighbour and aggregates this vector by applying the Equation (B.2), where the m parameter modulates how heavily the distances are weighted when the contributions of each neighbour are added up. Finally, the prediction will be based on the class with the higher membership.

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}(1/||x - x_j||^{2/(m-1)})}{\sum_{j=1}^{k}(1/||x - x_j||^{2/(m-1)})} \tag{B.2}$$

It is well-known that the Fuzzy-kNN algorithm does not cope well with very large datasets, especially because the class membership computation involves a training versus training comparison, which is very time-consuming. Recent approaches based on big data technologies proposed parallel alternatives to speed up the fuzzification process (Maillo et al., 2018). Nevertheless, the use of CzS data allows for skipping this fuzzification phase, considerably speeding up the total Fuzzy-kNN execution time, as we detail in the following section.

B.3 Methodology

In this section, we explain the main motivations for the approach proposed here. First, we introduce the data scores that enable an immediate fuzzification for the training examples in Section B.3.1. After this, in Section B.3.2 we explain how our method also leverages the amateur labels obtained within the course of CzS projects for a faster classification with the Fuzzy-kNN classifier.

B.3.1 Human-based Fuzzification

In off-line approaches, each example shown to participants throughout the project holds a particular count of votes once the project has finished. These votes are then converted to scores, simply dividing the votes assigned to each

category by the total number of votes received by the example. Let $\mathbf{N} = (n_1, n_2, ..., n_{|N|})$ be the vote vector, with $|N|$ the number of different options offered to participants, we get the score vector $\mathbf{X} = (x_1, x_2, ..., x_{|N|})$ by computing $x_i = \frac{n_i}{M}$, with $M = \sum n_i$ and $i \in \{1, 2, ..., |N|\}$. Hence, the amateur labelling does not result in final classifications as such. Conversely, there is available a set of scores for each example in the resulting data, as it is shown in Table 2.1. This set of options for amateurs' clicks generally involve a set of main classes, which are actually the target of the classification problem. Nonetheless, amongst them, one DK option is usually offered, assuring that every time the example is shown it gets a click from the person deciding at the moment. Unlike the classification target classes, this DK score provides a direct quantification of the example level of confidence: the more DK votes it gets, the higher is the expected noise introduced by the example to the final data.

Through the set of scores, each one accounts for the example's degree of membership to the $i - th$ option displayed on the web. Here, we propose to exploit these scores as a human-based fuzzification for the training of the Fuzzy-kNN classifier. However, the *original* scores directly obtained from amateur votes does not differentiate between the target classes, which tend to accumulate the major part of the votes, and the DK clicks and other minority classes, nor does the variability in the total number of votes received by the example, M. To prevent this, we first refine the original scores by the application of the data transformations that aggregate information about the uncertainty in the amateur votes (Chapter 3). Then, we directly use this refined scores obtained for the target classes as the fuzzy labels needed for the training phase of the Fuzzy-kNN algorithm. By this, we are skipping the learning phase that involves the computation of the fuzzy memberships for the training examples solely based on the distances rather than actually taking into account the uncertainty in amateur classifications (Equation B.1). Although here we use the Fuzzy-kNN for carrying out the different experimental trials, these transformed scores could be used with any other fuzzy classifier.

B.3.2 Instance Selection Based on Confidence

The score vector available for each of the examples in the data does not only tell us about the example degree of membership to each class of the classification

problem. From a ML point of view, the target classes' scores also hold information about the quality of the instance for the training of the automated classifier. Uneven score vectors will come from examples for which there has been a higher consensus amongst the people who voted that example. Thus, they will represent valuable examples for the learning algorithm. On the contrary, examples for which the votes are equally spread through the classes convey a poor agreement, thus providing a low confidence.

As part of our proposal, we apply a set of filters over the training set. This results in a straightforward selection of instances (García et al., 2012) based on the examples' level of confidence. We investigate the application of a set of thresholds over the target classes scores: examples with scores that do not reach the threshold are removed from the training set. By this, we carry out a cleaning of the training data based on the samples' confidence, which is given by the transformed scores that, in turn, incorporate information about the uncertainty within the CzS data.

B.4 Experiments and Analysis

In this section, we investigate the behaviour of the proposed approach in one of the most well-known CzS projects, the GZ1 (Lintott et al., 2008). First, Section B.4.1 defines the experimental setup, describing the main features of the GZ1 data and the methodology followed through the experiments. Then, Section B.4.2 presents and discusses the obtained results.

B.4.1 Experimental Setup

The GZ1 project focused on the classification of galaxy images in terms of their morphology. Later editions of this project were launched, providing more specific information about the morphological features of the galaxies, but they involved a considerably reduced number of images (Willett et al., 2013). The main focus was on the distinction between elliptical and spiral classes (morphologies), turning it into a true binary classification problem. Nonetheless, various aspects complicated the classification for the amateurs taking part, which explains the uncertainty encountered within the GZ1 data. These included the varied quality of the images, depending on the distance to the galaxy, its physical size and brightness, or the presence of artefacts that hinder the class identification (Lintott et al., 2008).

All images displayed on the GZ1 website held a common size of 423×423 pixels in order to provide a similar basis for all classifications, as described in Lintott et al. (2011). However, the Fuzzy-kNN algorithm does not directly work on the images but on a feature set extracted from them. As the original images generally leave the galaxy in the centre of the image surrounded by a big amount of background pixels, we first crop the images to the half of their size (112×112 pixels), and then shrink them to a final size of 52×52 pixels. After some additional experiments that are outside of the scope of this contribution, we opted for a FE based on AEs (Charte et al., 2018), employing the CAE thoroughly employed in the experiments completed in the main body of this research. This CAE is composed of three consecutive convolutional – pooling pairs of layers, from input to the encoding (Section 4.2.1), and generates a feature set of 392 features extracted from the 52×52 images of the GZ1-E sample. Acknowledging that some of these features may be redundant or noisy, we decided to apply a randomised decision tree extraction of features (50 trees) to choose the most relevant features. Figure B.1 shows the general workflow of the proposed approach for the GZ1 classification problem.

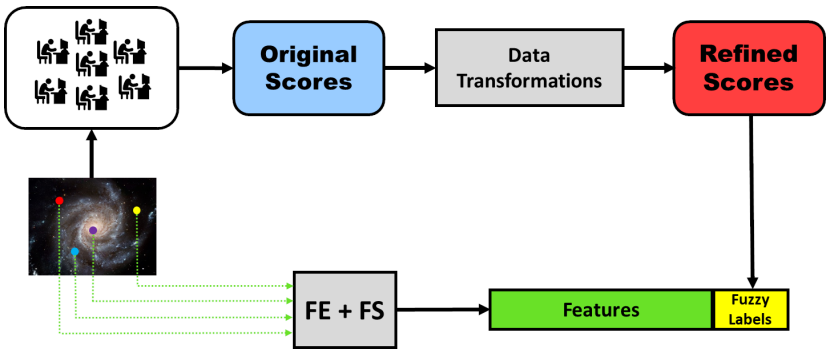


FIGURE B.1: Workflow of the proposed approach with the Fuzzy-kNN classifier.

Similarly as it is done in the comparative study of ML approaches for this classification problem previously developed (Chapter 4), we first make use of the GZ1-E subset of the GZ1-T2 data, which is composed of 41,424 galaxies (16,375 elliptical, 25,049 spiral) defined for the availability of expert classifications for these examples. In our experiments, we train the different k-NN variations on this sample, due to the high runtime required to handle the entire data. Nonetheless, the best performing approach is also tested on the whole GZ1-T2 dataset. To test the adequacy of the data transformations, we make use of the transformation sequence that best performed in a previous research conducted in Chapter 3. This transformation sequence consists of a combination of *Normalisation*, *DK votes shift*, and*Votes boost*, in this order. To assess how well the

uncertainty is managed within an automatic classification process, we carry out all experimental trials using both original and transformed scores, pursuing a first comparison amongst the unmodified data and the transformed data that incorporates the information about the uncertainty.

For the comparative study, we analyse the performance of the proposed approach using the original (raw) and transformed scores as the fuzzy labels needed for completing the Fuzzy-kNN training, which we will refer from now on as Citizen Science Fuzzy-kNN (CSF-kNN). We compare the CSF-kNN approach against the Crisp k-NN (Crisp-kNN) algorithm and the original Fuzzy-kNN (i.e. using the algorithm's original fuzzification phase (Keller and Givens, 1985)). Therefore, for either the Fuzzy-kNN and the Crisp-kNN, we need to turn the scores into crisp labels. This is accomplished by a simple majority criterion: the class holding a greater score is assigned the positive label (1), whereas the other is labelled as negative (0). When a draw is encountered, which occurs for only 225 and 37 examples with original and transformed scores, respectively, the labels are assigned at random.

We also analyse the impact of removing low confidence examples from the training set before running the algorithms detailed above. For this, we apply a set of thresholds ($\boldsymbol{Th_{conf}}$) over both score types: original and transformed. The mechanism is simple: examples with *both* elliptical and spiral scores strictly less than the selected threshold are removed.

The most well-known parameter for the Fuzzy-kNN is the number of neighbours ($\boldsymbol{k}$) to be considered in order to take the final classification decision (Equation B.2). We take $\boldsymbol{k}$ = 3 and $\boldsymbol{m}$ = 2.0 in our experiments. For the learning of the Fuzzy-kNN, we also take $\boldsymbol{k_{memb}}$ = 3 (Equation B.1). The whole experimental setting specifications are summarised in Table B.2. The final dataset is partitioned using a 5-fold cross-validation scheme (Section 2.5.1) to test all algorithms' performance with original and transformed scores. As result metrics, we measure the Acc (Section 2.5.3) and runtime required in each case.

Algorithm	**Parameters**
CAE	3 × convolutional – pooling layers ReLU activation function
Decision Tree FS	$\boldsymbol{n_{trees}}$ = 50
Crisp-kNN	$\boldsymbol{k}$ = 3 , $\boldsymbol{m}$ = 2.0
Fuzzy-kNN	$\boldsymbol{k}$ = $\boldsymbol{k_{memb}}$ = 3, $\boldsymbol{m}$ = 2.0
CSF-kNN	$\boldsymbol{k}$ = 3, $\boldsymbol{m}$ = 2.0, $\boldsymbol{Th_{conf}}$ = {0.8, 0.9, 1.0}

TABLE B.2: Parameter settings of the algorithms involved.

All the experiments have been carried out in a single node with an Intel(R) Xeon(R) CPU E5-1650 v4 processor (12 cores) at 3.60GHz, and 64 GB of RAM. In terms of software, the Keras Python package was used for the training of the AE, and the Scikit-learn library[1] was used for the FS and the different experiments involving the training and testing of the different k-NN versions explained above.

B.4.2 Results and Analysis

In this section, we present the results of all experimental trials performed to test the proposed approach. As detailed above, we first carry out a comparison between the algorithms using both original and transformed scores for the two classes of the problem, elliptical and spiral. After this, we extend the testing of the CSF-kNN algorithm applying three thresholds that filter the training set. Finally, we complete the testing of the CSF-kNN on the GZ1 whole dataset after applying a stringent 1.0 threshold to both score types, original and transformed. This filter allows for selecting, between the whole dataset, the examples holding a perfect consensus amongst the amateurs who voted them.

The first experiment presents the comparative study amongst the Crisp-kNN, the Fuzzy-kNN, and the proposed CSF-kNN algorithms. The averaged Acc and runtime values are summarised in Table B.3.

Algorithm	**Original scores**		**Transformed scores**	
	Accuracy	Runtime (s)	Accuracy	Runtime (s)
Fuzzy-kNN	0.8971	23,184	0.9166	23,192
Crisp-kNN	0.8958	5,607	0.9183	5,597
CSF-kNN	0.9004	3,617	0.9188	3,758

TABLE B.3: Averaged Acc and runtime values for the comparative study of algorithms.

The second experiment introduces the instance selection on the training set before carrying out the training of the different algorithms. In each case, the selection is carried out applying the threshold either over original or transformed scores, in such a way that if the example does not reach the threshold value in either of the elliptical or spiral class, it is rejected. Table B.4 presents the results when applying the thresholds $\boldsymbol{Th_{conf}} = \{0.8, 0.9, 1.0\}$. At this respect, we reflect the proportion of the sample that is removed from the training set in Figure B.2. Percentages are shown on top of the bars.

[1] https://scikit-learn.org

$\mathbf{Th_{conf}}$	**Original scores**		**Transformed scores**	
	Accuracy	Runtime (s)	Accuracy	Runtime (s)
0.8	0.9032	2,354	0.9177	3,168
0.9	0.8997	1,497	0.9186	2,671
1.0	0.8820	476	0.9173	906

TABLE B.4: Averaged Acc and runtime values of CSF-kNN with instance selection.

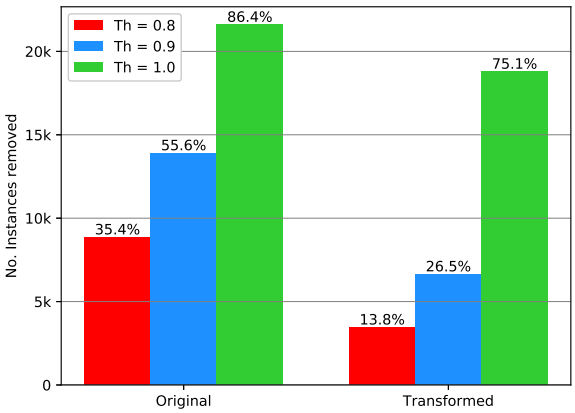


FIGURE B.2: Number of instances removed by the different thresholds (Th).

For a graphical comparison of the execution times, we represent the values corresponding to the first comparative study along with the CSF-kNN with $\boldsymbol{Th_{conf}}$ = 1.0 threshold in a bar chart. It is shown in Figure B.3.

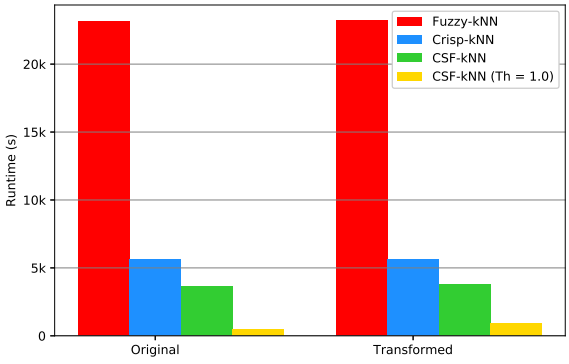


FIGURE B.3: Comparison of algorithms' runtime.

Finally, we present the values obtained when the CSF-kNN algorithm with 1.0 threshold instance selection is trained using the entire GZ1 dataset. The threshold is applied in order to speed up the execution up to a manageable

runtime for this big dataset (∼893k examples). Table B.5 presents the averaged Acc and execution time values recorded for this last experiment.

$\mathbf{Th_{conf}}$	**Original scores**		**Transformed scores**	
	Accuracy	Runtime (s)	Accuracy	Runtime (s)
1.0	0.7616	23,254	0.8291	68,253

TABLE B.5: Results of CSF-kNN with 1.0 threshold over the whole GZ1 dataset.

According to the results shown above, we can draw the following conclusions:

- In first place, the use of either original or transformed scores is able to equalise the performance obtained by the fuzzy and crisp implementations of the k-NN algorithm. However, the execution times are highly improved, as Figure B.3 shows. The transformed scores also improve the Acc result in around two points with respect to the original scores, confirming the utility of the data transformations presented in Jiménez, Triguero, and John (2019).
- The selection of instances across the training set by the application of thresholds provides a great reduction of the training times (Table B.4), which is greater for the original scores. This happens because the transformations tend to boost the original scores for the major part of the GZ1-V examples. Nevertheless, the Acc is not much affected.
- When the CSF-kNN algorithm is tested against the whole GZ1 dataset, we can observe that it generalises well. The transformed scores provide a better classification, making bigger the difference in Acc with the original scores. However, the training is also completed from a higher number of examples, which is also reflected in the execution time.

B.5 Summary

In this appendix, we have proposed an innovative use of the data recorded in the course of a CzS project to build robust and fast fuzzy-based classifiers. The main novelty lies in the introduction of the classifications performed by amateur participants in both original and refined forms as the fuzzy labels needed for the training of the Fuzzy-KNN algorithm. Throughout two sets of experiments on a real-world application, using the results coming from the GZ1 project, the proposed methodology has shown that a correct exploitation of CzS results may lead to not only a good classification performance in terms

of Acc but also to a natural filter of noisy and/or imprecise training data. One possible extension of this work would consider the use of big data approaches to complete further experiments on larger datasets, and the implementation of other fuzzy-based classifiers that could take advantage of the methodology proposed here.

Bibliography

Ackermann, S. et al. (2018). "Using transfer learning to detect galaxy mergers". In: *Monthly Notices of the Royal Astronomical Society* 479.1, pp. 415–425.

Akila, A.S. and J. Anitha (2019). "Analyzing the effect of optimization strategies in deep convolutional neural network". In: *Intelligent Systems Reference Library* 150, pp. 235–253.

Alabri, A. and J. Hunter (2010). "Enhancing the quality and trust of citizen science data". In: *Proceedings of the 6th IEEE International Conference on e-Science*. eScience'10, pp. 81–88.

Anderson, T.M. et al. (2016). "The spatial distribution of african savannah herbivores: Species associations and habitat occupancy in a landscape context". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1703.

Bahaadini, S. et al. (2018). "Machine learning for Gravity Spy: Glitch classification and dataset". In: *Information Sciences* 444, pp. 172–186.

Ball, N.M. and R.J. Brunner (2010). "Data mining and machine learning in astronomy". In: *International Journal of Modern Physics D* 19.7, pp. 1049–1106.

Bamford, S.P. et al. (2009). "Galaxy Zoo: The dependence of morphology and colour on environment". In: *Monthly Notices of the Royal Astronomical Society* 393.4, pp. 1324–1352.

Banerji, M. et al. (2010). "Galaxy Zoo: Reproducing galaxy morphologies via machine learning". In: *Monthly Notices of the Royal Astronomical Society* 406.1, pp. 342–353.

Barber, S. T. (2018). "The Zooniverse is expanding: Crowdsourced solutions to the hidden collections problem and the rise of the revolutionary cataloging interface". In: *Journal of Library Metadata* 18.2, pp. 85–111.

Barnard, L. et al. (2014). "The solar stormwatch CME catalogue: Results from the first space weather citizen science project". In: *Space Weather* 12.12, pp. 657–674.

Basheer, I.A. and M. Hajmeer (2000). "Artificial neural networks: Fundamentals, computing, design, and application". In: *Journal of Microbiological Methods* 43.1, pp. 3–31.

Beaumont, C.N. et al. (2014). "The milky way project: Leveraging citizen science and machine learning to detect interstellar bubbles". In: *Astrophysical Journal Supplement Series* 214.1, p. 3.

Bengio, Y. (2009). "Learning deep architectures for AI". In: *Foundations and Trends in Machine Learning* 2.1, pp. 1–27.

Bergh, S. van den (1976). "A new classification system for galaxies". In: *Astrophysical Journal* 206, pp. 883–887.

Bird, T.J. et al. (2014). "Statistical solutions for error and bias in global citizen science datasets". In: *Biological Conservation* 173, pp. 144–154.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer. ISBN: 0387310738.

Blickhan, S. et al. (2019). "Individual vs. collaborative methods of crowdsourced transcription". In: *Journal of Data Mining and Digital Humanities* Special Issue on Collecting, Preserving, and Disseminating Endangered Cultural Heritage for New Understandings through Multilingual Approaches.

Bonney, R. et al. (2009). "Citizen science: A developing tool for expanding science knowledge and scientific literacy". In: *BioScience* 59.11, pp. 977–984.

Bonney, R. et al. (2014). "Next steps for citizen science". In: *Science* 343.6178, pp. 1436–1437.

Bonter, D.N. and C.B. Cooper (2012). "Data validation in citizen science: A case study from Project FeederWatch". In: *Frontiers in Ecology and the Environment* 10.6, pp. 305–307.

Borne, K. (2010). "Astroinformatics: Data-oriented astronomy research and education". In: *Earth Science Informatics* 3.1, pp. 5–17.

– (2013). *Virtual observatories, data mining, and astroinformatics*, pp. 403–443.

– (Taylor and Francis, 2009). *Scientific Data Mining in Astronomy*, pp. 91–114.

Bouveyron, C. and S. Girard (2009). "Robust supervised classification with mixture models: Learning from data with uncertain labels". In: *Pattern Recognition* 42.11, pp. 2649–2658.

Brabham, D.C. (2008). "Crowdsourcing as a model for problem solving: An introduction and cases". In: *Convergence* 14.1, pp. 75–90.

Breiman, L. (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32.

Brescia, M. and G. Longo (2013). "Astroinformatics, data mining and the future of astronomical research". In: *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 720, pp. 92–94.

Bu, Y. et al. (2019). "Searching for hot subdwarf stars from the LAMOST spectra. III. Classification of hot subdwarf stars in the fourth data release of LAMOST using a deep learning method". In: *Astrophysical Journal* 886.2.

Burges, C.J.C. (1998). "A tutorial on support vector machines for pattern recognition". In: *Data Mining and Knowledge Discovery* 2.2, pp. 121–167.

Charte, D. et al. (2018). "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines". In: *Information Fusion* 44, pp. 78–96.

Cherkassky, V. and F. Mulier (2006). *Learning from Data: Concepts, Theory, and Methods*. 2nd Edition. John Wiley and Sons. ISBN: 978-047168182-3.

Cohn, J.P. (2008). "Citizen science: Can volunteers do real research?" In: *BioScience* 58.3, pp. 192–197.

Conselice, C.J., M.A. Bershady, and A. Jangren (2000). "The asymmetry of galaxies: Physical morphology for nearby and high-redshift galaxies". In: *Astrophysical Journal* 529.2 I, pp. 886–910.

Cooper, S. et al. (2010). "Predicting protein structures with a multiplayer online game". In: *Nature* 466.7307, pp. 756–760.

Coughlin, S. et al. (2019). "Classifying the unknown: Discovering novel gravitational-wave detector glitches using similarity learning". In: *Physical Review D* 99.8.

Cover, T.M. and P.E. Hart (1967). "Nearest Neighbor Pattern Classification". In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27.

Crowston, K., C. Osterlund, and T. K. Lee (2017). "Blending machine and human learning processes". In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. HICSS'17, pp. 65–73.

Crowston, K. et al. (2020). "Knowledge tracing to model learning in online citizen science projects". In: *IEEE Transactions on Learning Technologies* 13.1, pp. 123–134.

Cui, Yin et al. (2014). "A spatial-color layout feature for representing galaxy images". In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 213–219.

Darg, D.W. et al. (2010). "Galaxy Zoo: The fraction of merging galaxies in the SDSS and their morphologies". In: *Monthly Notices of the Royal Astronomical Society* 401.2, pp. 1043–1056.

De La Calleja, J. and O. Fuentes (2004). "Machine learning and image analysis for morphological galaxy classification". In: *Monthly Notices of the Royal Astronomical Society* 349.1, pp. 87–93.

Deibe, D., M. Amor, and R. Doallo (2020). "Big data geospatial processing for massive aerial LiDAR datasets". In: *Remote Sensing* 12.4.

Deng, J. et al. (2009). "ImageNet: A large-scale hierarchical image database". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Dickinson, J.L., B. Zuckerberg, and D.N. Bonter (2010). "Citizen science as an ecological research tool: Challenges and benefits". In: *Annual Review of Ecology, Evolution, and Systematics* 41, pp. 149–172.

Dieleman, S., K.W. Willett, and J. Dambre (2015). "Rotation-invariant convolutional neural networks for galaxy morphology prediction". In: *Monthly Notices of the Royal Astronomical Society* 450.2, pp. 1441–1459.

Douzas, G. and F. Bacao (2018). "Effective data generation for imbalanced learning using conditional generative adversarial networks". In: *Expert Systems with Applications* 91, pp. 464–471.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification*. 2nd Edition. New York: Wiley. ISBN: 978-0-471-05669-0.

Eckstein, L.N. (2019). "Of scribes and scripts: Citizen science and the Cairo genizah". In: *Manuscript Studies* 3.1.

Fedrizzi, M. and G. Pasi (2008). "Fuzzy logic approaches to consensus modelling in group decision making". In: *Studies in Computational Intelligence* 117, pp. 19–37.

Fortson, L. et al. (2012). "Galaxy Zoo: Morphological classification and citizen science". In: *Machine Learning and Data Mining for Astronomy* 11, pp. 118–125.

Galar, M. et al. (2011). "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes". In: *Pattern Recognition* 44.8, pp. 1761–1776.

Galar, M. et al. (2012). "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches". In: *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 42.4, pp. 463–484.

García, S. et al. (2012). "Prototype selection for nearest neighbor classification: Taxonomy and empirical study". In: *IEEE Transactions Pattern Analysis and Machine Intelligence* 34.3, pp. 417–435.

Garibaldi, J.M. and T. Ozen (2007). "Uncertain fuzzy reasoning: A case study in modelling expert decision making". In: *IEEE Transactions on Fuzzy Systems* 15.1, pp. 16–30.

González, R.E., R.P. Muñoz, and C.A. Hernández (2018). "Galaxy detection and identification using deep learning and data augmentation". In: *Astronomy and Computing* 25, pp. 103–109.

González Delgado, R.M. et al. (2017). "Spatially-resolved star formation histories of CALIFA galaxies: Implications for galaxy formation". In: *Astronomy and Astrophysics* 607.128, pp. 1–21.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. ISBN: 9780262035613.

Grayson, R. (2016). "A life in the trenches? The use of operation war diary and crowdsourcing methods to provide an understanding of the British Army's day-to-day life on the western front". In: *British Journal for Military History* 2.2.

Gu, J. et al. (2018). "Recent advances in convolutional neural networks". In: *Pattern Recognition* 77, pp. 354–377.

Guyon, I. and A. Elisseefl (2006). "An introduction to feature extraction". In: *Studies in Fuzziness and Soft Computing* 207, pp. 1–25.

Hart, R.E. et al. (2018). "Galaxy Zoo: Constraining the origin of spiral arms". In: *Monthly Notices of the Royal Astronomical Society* 478.1, pp. 932–949.

He, K. et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hennon, C.C. et al. (2015). "Cyclone center can citizen scientists improve tropical cyclone intensity records?" In: *Bulletin of the American Meteorological Society* 96.4, pp. 591–607.

Herodotou C., Aristeidou M. Miller G. Ballard H. and L. Robinson (2020). "What do we know about young volunteers? An exploratory study of participation in Zooniverse". In: *Citizen Science: Theory and Practice* 5.1, pp. 1–14.

Holincheck, A.J. et al. (2016). "Galaxy zoo: Mergers - Dynamical models of interacting galaxies". In: *Monthly Notices of the Royal Astronomical Society* 459.1, pp. 720–745.

Hubble, E. (1926). "Extra-galactic nebulae". In: *The Astrophysical Journal* 64, pp. 321–373.

– (Yale University Press, 1936). *The Realm of the Nebulae*, pp. 1–207.

Huertas-Company, M. (2015). "Galaxy morphologies in the era of big-data surveys". In: *Proceedings of the International Astronomical Union* 11, pp. 118–125.

Huertas-Company, M. et al. (2008). "A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images: I. Method description". In: *Astronomy and Astrophysics* 478.3, pp. 971–980.

Huertas-Company, M. et al. (2015). "A catalog of visual-like morphologies in the 5 CANDELS fields using deep learning". In: *Astrophysical Journal Supplement Series* 221.1, pp. 1–23.

Ivezić, Z. et al. (2019). "LSST: From science drivers to reference design and anticipated data products". In: *Astrophysical Journal* 873.2, pp. 1–44.

Jackson, C. B. et al. (2015). "Motivations for sustained participation in crowdsourcing: Case studies of citizen science on the role of talk". In: *2015 48th Hawaii International Conference on System Sciences*, pp. 1624–1634.

Jackson, C. B. et al. (2020). "Teaching citizen scientists to categorize glitches using machine learning guided training". In: *Computers in Human Behavior* 105.

Jain, A.K., M.N. Murty, and P.J. Flynn (1999). "Data clustering: A review". In: *ACM Computing Surveys* 31.3, pp. 264–323.

Jain, D.K., P. Shamsolmoali, and P. Sehdev (2019). "Extended deep neural network for facial emotion recognition". In: *Pattern Recognition Letters* 120, pp. 69–74.

Jiménez, M., I. Triguero, and R. John (2019). "Handling uncertainty in citizen science data: Towards an improved amateur-based large-scale classification". In: *Information Sciences* 479, pp. 301–320.

Jiménez, R.Z., L.F. Gallardo, and S. Moller (2018). "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing". In: *Proceedings of the 10th International Conference on Quality of Multimedia Experience*. QoMEX 2018.

Jiménez, M., M. Torres Torres, R. John, and I. Triguero (2020). "Galaxy image classification based on citizen science data: A comparative study". In: *IEEE Access* 8, pp. 47232–47246.

Jungbluth, A., J.L. Yeng, and L. Vives (2018). "Quality data extraction methodology based on the labeling of coffee leaves with nutritional deficiencies". In: *ACM International Conference Proceeding Series*, pp. 59–64.

Kamar, E., S. Hacker, and E. Horvitz (2012). "Combining human and machine intelligence in large-scale crowdsourcing". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*. AAMAS'12, pp. 467–474.

Karras, T. et al. (2018). "Progressive growing of GANs for improved quality, stability, and variation". In: *6th International Conference on Learning Representations, ICLR 2018*.

Keller, J.M. and Jr. Givens J.A. (1985). "A fuzzy K-nearest neighbor algorithm". In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-15.4, pp. 580–585.

Kieffer, B. et al. (2018). "Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks". In: *Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications, IPTA 2017*. Vol. 2018-January, pp. 1–6.

Kim, J.S. et al. (2014). "Space-time wiring specificity supports direction selectivity in the retina". In: *Nature* 509.7500, pp. 331–336.

Klir, G.J. (1987). "Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like?" In: *Fuzzy Sets and Systems* 24.2, pp. 141–160.

Knapp, K. R. et al. (2016). "Identification of tropical cyclone storm types using crowdsourcing". In: *Monthly Weather Review* 144.10, pp. 3783–3798.

Kosmala, M. et al. (2016). "Assessing data quality in citizen science". In: *Frontiers in Ecology and the Environment* 14.10, pp. 551–560.

Kotsiantis, S.B. (2007). "Supervised machine learning: A review of classification techniques". In: *Informatica (Ljubljana)* 31.3, pp. 249–268.

Kuminski, E. and L. Shamir (2016). "A computer-generated visual morphology catalog of 3,000,000 SDSS galaxies". In: *Astrophysical Journal Supplement Series* 223.2, pp. 1–10.

Kuminski, E. et al. (2014). "Combining human and machine learning for morphological analysis of galaxy images". In: *Publications of the Astronomical Society of the Pacific* 126.944, pp. 959–967.

Lahav, O. et al. (1995). "Galaxies, human eyes, and artificial neural networks". In: *Science* 267.5199, pp. 859–862.

Lecun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

LeCun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2323.

Lee, J. et al. (2014). "RNA design rules from a massive open laboratory". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.6, pp. 2122–2127.

Lee, T. K. et al. (2018). "Appealing to different motivations in a mesage to recruit citizen scientists: Results of a field study". In: *Journal of Science Communication* 17.1, pp. 1–22.

Li, W. et al. (2012). "Co-labeling: A new multi-view learning approach for ambiguous problems". In: *Proceedings of the IEEE International Conference on Data Mining*. ICDM'12, pp. 419–428.

Lintott, C. et al. (2008). "Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 389.3, pp. 1179–1189.

Lintott, C. et al. (2011). "Galaxy Zoo 1: Data release of morphological classifications for nearly 900 000 galaxies". In: *Monthly Notices of the Royal Astronomical Society* 410.1, pp. 166–178.

Longo, M.J. (2011). "Detection of a dipole in the handedness of spiral galaxies with redshifts z 0.04". In: *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics* 699.4, pp. 224–229.

Lopes, U.K. and J.F. Valiati (2017). "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection". In: *Computers in Biology and Medicine* 89, pp. 135–143.

Lotz, J.M., J. Primack, and P. Madau (2004). "A new nonparametric approach to galaxy morphological classification". In: *Astronomical Journal* 128.1, pp. 163–182.

López, V. et al. (2013). "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". In: *Information Sciences* 250, 113–141.

Lu, D. et al. (2019). "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network". In: *Medical Image Analysis* 54, pp. 100–110.

Mahdianpari, M. et al. (2018). "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery". In: *Remote Sensing* 10.7.

Maillo, J. et al. (2018). "A preliminary study on Hybrid Spill-Tree Fuzzy k-Nearest Neighbors for big data classification". In: *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8.

Mansourifar, H., L. Chen, and W. Shi (2019). "Virtual big data for GAN based data augmentation". In: *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 1478–1487.

Marmanis, D. et al. (2016). "Deep learning earth observation classification using ImageNet pretrained networks". In: *IEEE Geoscience and Remote Sensing Letters* 13.1, pp. 105–109.

Masters, K.L. (2019). "Twelve years of Galaxy Zoo". In: *Proceedings of the International Astronomical Union* 14.S353, 205–212.

Miller-Rushing, A., R. Primack, and R. Bonney (2012). "The history of public participation in ecological research". In: *Frontiers in Ecology and the Environment* 10.6, pp. 285–290.

Mjolsness, E. and D. DeCoste (2001). "Machine learning for science: State of the art and future prospects". In: *Science* 293.5537, pp. 2051–2055.

Nair, V. and G.E. Hinton (2010). "Rectified linear units improve Restricted Boltzmann machines". In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814.

Nisbet, R., G. Miner, and J. Elder (2009). *Handbook of Statistical Analysis and Data Mining Applications*. ISBN: 978-0-12-374765-5.

Nixon, M. and A.S. Aguado (2008). *Feature Extraction & Image Processing*. 2nd Edition. Orlando: Academic Press. ISBN: 780123725387.

Orlov, N. et al. (2008). "WND-CHARM: Multi-purpose image classification using compound image transforms". In: *Pattern Recognition Letters* 29.11, pp. 1684–1693.

Oswalt, T.D. and H.E. Bond (Springer, 2013). *Planets, stars and stellar systems: Volume 2: Astronomical techniques, software, and data*, pp. 1–508.

Ou, G. and Y.L. Murphey (2007). "Multi-class pattern classification using neural networks". In: *Pattern Recognition* 40.1, pp. 4–18.

Peralta, D. et al. (2018). "On the use of convolutional neural networks for robust classification of multiple fingerprint captures". In: *International Journal of Intelligent Systems* 33.1, pp. 213–230.

Petrillo, C.E. et al. (2019). "Testing convolutional neural networks for finding strong gravitational lenses in KiDS". In: *Monthly Notices of the Royal Astronomical Society* 482.1, pp. 807–820.

Pulgar, F.J. et al. (2018). "AEkNN: An AutoEncoder kNN-based classifier with built-in dimensionality reduction". In: *International Journal of Computational Intelligence Systems* 12.1, pp. 436–452.

Ranzato, M. et al. (2007). "A unified energy-based framework for unsupervised learning". In: *Journal of Machine Learning Research* 2, pp. 371–379.

Rawat, W. and Z. Wang (2017). "Deep convolutional neural networks for image classification: A comprehensive review". In: *Neural Computation* 29.9, pp. 2352–2449.

Reis, F.J. Candido dos et al. (2015). "Crowdsourcing the general public for large scale molecular pathology studies in cancer". In: *EBioMedicine* 2.7, pp. 681–689.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536.

Saeys, Y., I. Inza, and P. Larrañaga (2007). "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23.19, pp. 2507–2517.

Sandage, A. (2005). "The classification of galaxies: Early history and ongoing developments". In: *Annual Review of Astronomy and Astrophysics* 43, pp. 581–624.

Sandage, A.R. (Carnegie Institute of Washington, 1961). *The Hubble Atlas of Galaxies*, pp. 1–621.

Schawinski, K. et al. (2007). "Observational evidence for AGN feedback in early-type galaxies". In: *Monthly Notices of the Royal Astronomical Society* 382.4, pp. 1415–1431.

Schwamb, M.E. et al. (2012). "Planet hunters: Assessing the kepler inventory of short-period planets". In: *Astrophysical Journal* 754.2, p. 129.

Shamir, L. (2012). "Automatic detection of peculiar galaxies in large datasets of galaxy images". In: *Journal of Computational Science* 3.3, pp. 181–189.

Shamir, L., A. Holincheck, and J. Wallin (2013). "Automatic quantitative morphological analysis of interacting galaxies". In: *Astronomy and Computing* 2, pp. 67–73. ISSN: 2213-1337.

Shamir, L. et al. (2008). "Wndchrm - An open source utility for biological image analysis". In: *Source Code for Biology and Medicine* 3.

Shamir, L. et al. (2014). "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls". In: *Journal of the Acoustical Society of America* 135.2, pp. 953–962.

Shin, H.-C. et al. (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE Transactions on Medical Imaging* 35.5, pp. 1285–1298.

Show, H. (2015). "Rise of the citizen scientist". In: *Nature* 524.7565, p. 265.

Silverberg, S. M. et al. (2018). "Follow-up imaging of disk candidates from the Disk Detective citizen science project: New discoveries and false positives in WISE circumstellar disk surveys". In: *The Astrophysical Journal* 868.1, pp. 1–43.

Silvertown, J. (2009). "A new dawn for citizen science". In: *Trends in Ecology and Evolution* 24.9, pp. 467–471.

Simonyan, K. and A. Zisserman (2015). "Very deep convolutional networks for large-scale image recognition". In: *International Conference on Learning Representations*.

Simpson, R., K.R. Page, and D. De Roure (2014). "Zooniverse: Observing the world's largest citizen science platform". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW'14, pp. 1049–1054.

Skibba, R.A. et al. (2009). "Galaxy Zoo: Disentangling the environmental dependence of morphology and colour". In: *Monthly Notices of the Royal Astronomical Society* 399.2, pp. 966–982.

Skibba, R.A. et al. (2012). "Galaxy Zoo: The environmental dependence of bars and bulges in disc galaxies". In: *Monthly Notices of the Royal Astronomical Society* 423.2, pp. 1485–1502.

Spiers, Helen et al. (2019). "Everyone counts? Design considerations in online citizen science". In: *Journal of Science Communication* 18.1, pp. 1–32.

Sprinks, J. et al. (2017). "Task workflow design and its impact on performance and volunteers' subjective preference in virtual citizen science". In: *International Journal of Human Computer Studies* 104, pp. 50–63.

Sprinks, J. et al. (2019). "Planet Four: Craters—Optimizing task workflow to improve volunteer engagement and crater counting performance". In: *Meteoritics and Planetary Science* 54.6, pp. 1325–1346.

Starkey, E. et al. (2017). "Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation". In: *Journal of Hydrology* 548, pp. 801–817. ISSN: 0022-1694.

Szegedy, C. et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Taherkhani, F. et al. (2019). "A weakly supervised fine label classifier enhanced by coarse supervision". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Thabtah, F. et al. (2020). "Data imbalance in classification: Experimental evaluation". In: *Information Sciences* 513, pp. 429–441.

Thonnat, M. and M. Berthod (1984). "Automatic classification of galaxies into morphological types". In: *Proceedings of the International Conference on Pattern Recognition*. Vol. 2, pp. 844–846.

Triguero, I., S. García, and F. Herrera (2015). "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study". In: *Knowledge and Information Systems* 42, pp. 245–284.

Triguero, I. et al. (2015). "ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem". In: *Knowledge-Based Systems* 87, pp. 69–79.

Triguero, I. et al. (2017). "A first attempt on global evolutionary undersampling for imbalanced big data". In: *Proceedings of the 2017 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2054–2061.

Trouille, L., C.J. Lintott, and L.F. Fortson (2019). "Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human-machine systems". In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 116. 6, pp. 1902–1909.

Tsiporkova, E. and V. Boeva (2006). "Multi-step ranking of alternatives in a multi-criteria and multi-expert decision making environment". In: *Information Sciences* 176.18, pp. 2673–2697.

Tyson, J.A. (2010). "Optical synoptic telescopes: New science frontiers". In: *Proceedings of The International Society for Optical Engineering*. Vol. 7733. PART 1.

– (2019). "Cosmology data analysis challenges and opportunities in the LSST sky survey". In: vol. 1290. 1.

Umehara, M. et al. (2019). "Analyzing machine learning models to accelerate generation of fundamental materials insights". In: *npj Computational Materials* 5, pp. 1–9.

Varma, S. and R. Simon (2006). "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7, pp. 1–8.

Vaucouleurs, G. de (1959). *Classification and morphology of external galaxies*.

Vaucouleurs, G. de et al. (Springer-Verlag, 1991). *Third Reference Catalogue of Bright Galaxies*, pp. 1–621.

Walmsley, M. et al. (2020). "Galaxy Zoo: Probabilistic morphology through Bayesian CNNs and active learning". In: *Monthly Notices of the Royal Astronomical Society* 491.2, pp. 1554–1574.

Wang, J. et al. (2019a). *Deep High-Resolution Representation Learning for Visual Recognition*. arXiv: 1908.07919 [cs.CV].

Wang, Q. et al. (2019b). "WGAN-Based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in CT images". In: *IEEE Access* 7, pp. 18450–18463.

Wardlaw, J. et al. (2018). "Comparing experts and novices in Martian surface feature change detection and identification". In: *International Journal of Applied Earth Observation and Geoinformation* 64, pp. 354–364.

Watson, C. et al. (2019). "Galaxy merger fractions in two clusters at Z 2 using the Hubble Space Telescope". In: *Astrophysical Journal* 874.63, pp. 1–20.

Watson, D. and L. Floridi (2018). "Crowdsourced science: sociotechnical epistemology in the e-research paradigm". In: *Synthese* 195.2, pp. 741–764.

Wijnhoven, R.G.J. and P.H.N. De With (2010). "Fast training of object detection using stochastic gradient descent". In: *Proceedings of the International Conference on Pattern Recognition*, pp. 424–427.

Willett, K.W. et al. (2013). "Galaxy zoo 2: Detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 435.4, pp. 2835–2860.

Willett, K.W. et al. (2017). "Galaxy Zoo: Morphological classifications for 120,000 galaxies in HST legacy imaging". In: *Monthly Notices of the Royal Astronomical Society* 464.4, pp. 4176–4203.

Witten, Ian H. and Eibe Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition. San Francisco: Morgan Kaufmann. ISBN: 0120884070.

Wright, D. E. et al. (2017). "A transient search using combined human and machine classifications". In: *Monthly Notices of the Royal Astronomical Society* 472.2, pp. 1315–1323.

Wu, X. et al. (2008). "Top 10 algorithms in data mining". In: *Knowledge and Information Systems* 14.1, pp. 1–37.

Xie, S. et al. (2017). "Aggregated residual transformations for deep neural networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995.

Yarowsky, David (n.d.). "Unsupervised word sense disambiguation rivaling supervised methods". In: *33rd Annual Meeting of the Association for Computational Linguistics*.

Yu, D. and M.L. Seltzer (2011). "Improved bottleneck features using pretrained deep neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 237–240.

Yu, J., W.K. Wong, and R.A. Hutchinson (2010). "Modeling experts and novices in citizen science data for species distribution modeling". In: *Proceedings of the IEEE International Conference on Data Mining*. ICDM'10, pp. 1157–1162.

Zadeh, L.A. (1965). "Fuzzy sets". In: *Information and Control* 8.3, pp. 338–353.

Zeng, Z. et al. (2020). "Integrating Internet media into urban flooding susceptibility assessment: A case study in China". In: *Cities* 101, p. 102697. ISSN: 0264-2751.

Zevin, M. et al. (2017). "Gravity Spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science". In: *Classical and Quantum Gravity* 34.6, pp. 64003–64025.

Zhang, Y. and Y. Zhao (2015). "Astronomy in the big data era". In: *Data Science Journal* 14.

Zhu, X. and A.B. Goldberg (2009). "Introduction to semi-supervised learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3.1, pp. 1–130.

Zhu, X.P. et al. (2019). "Galaxy morphology classification with deep convolutional neural networks". In: *Astrophysics and Space Science* 364.55, pp. 1–15.

Zitzler, E. et al. (2003). "Performance assessment of multiobjective optimizers: An analysis and review". In: *IEEE Transactions on Evolutionary Computation* 7.2, pp. 117–132.

Zorich, L., K. Pichara, and P. Protopapas (2020). "Streaming classification of variable stars". In: *Monthly Notices of the Royal Astronomical Society* 492.2, pp. 2897–2909.