University of Nottingham
UK | CHINA | MALAYSIA

THE NARRATIVE OF GALAXY MORPHOLOGICAL CLASSIFICATION TOLD THROUGH MACHINE LEARNING

Thesis submitted to the University of Nottingham
for the degree **Doctor of Philosophy.**

Ting-Yun Cheng (鄭婷筠)

4296811

Supervised by
Christopher J. Conselice
Alfonso Aragón-Salamanca

School of Physics and Astronomy
University of Nottingham

October 2020

Abstract

In this thesis, we present a complete study of machine learning applications, including both supervised and unsupervised, for galaxy morphological classification using calibrated imaging data. Two main topics are approached: (1) classification - we discuss optimal machine learning technique in terms of accuracy, efficiency, and inclusiveness using imaging data for large-scale surveys; (2) exploration - we explore galaxy morphology without human bias and discuss a novel morphological classification scheme defined by machine learning.

In the classification task, we first carry out a thorough comparison in accuracy and efficiency between several common supervised methods using the Dark Energy Survey (DES) imaging data (Chapter 2). The morphology labels from the Galaxy Zoo 1 (GZ1) catalogue (Lintott et al., 2008, 2011) are used to train the supervised methods. We conclude that using a combination of linear and gradient images (with the Histogram of Oriented Gradient technique) to train our convolutional neural networks (CNN) shows the most optimal performance in terms of accuracy and efficiency amongst the supervised methods tested using imaging data. Due to the better resolution ($0.''263$ per pixel) and greater depth ($i = 22.51$) of DES data than the Sloan Digital Sky survey (SDSS) imaging data used in the GZ1 project, we reveal that $\sim 2.5\%$ galaxies in our dataset are mislabeled by the GZ1. After correcting these galaxies' labels based on the DES imaging data, we reach a final accuracy of over 0.99 for binary classification (ellipticals and spirals) with the CNN (Chapter 3). We then use the CNN to build one of the largest galaxy morphological classification catalogues which includes over 20 million galaxies from the DES Year 3 data (Chapter 4). However, supervised machine learning techniques are biased towards the training set and the human-defined labels. Therefore, we test the possibility of a classification task using unsupervised machine learning techniques (Chapter 5 and Chapter 6). In Chapter 5, the combination of a convolutional autoencoder and a Bayesian Gaussian mixture model successfully distinguishes a variety of lensing features such as different Einstein ring sizes and arcs from galaxy-galaxy strong lensing systems (GGSL). This unsupervised method categorises simulated images from Metcalf et al. (2019a) into 24 classes without human involvement and picks up ~ 63 percent of lensing images from all lenses in the training set. Additionally, with fewer human judgements involved to classify 24 machine classes, we reach an accuracy of $77.3 \pm 0.5\%$ in the binary classification of lensing and non-lensing systems.

On the other hand, unsupervised machine learning techniques are used to objectively explore galaxy morphology using the SDSS imaging data in Chapter 6. We improve the efficiency of the unsupervised method used in Chapter 5 by applying a vector quantisation process in the feature learning phase, and achieve a better 'accuracy' compared to the current knowledge towards galaxy morphology using an uneven iterative hierarchical clustering (Chapter 6). This unsupervised method can categorise the galaxies in the dataset, which includes 23% early-type galaxies (ETGs) and 77% late-type galaxies (LTGs), into two preliminary classes and reach an accuracy of ~ 0.87 for binary classification of ETGs and LTGs. To

explore galaxy morphology, our method provides 27 classes based on the galaxy shape and structure. We further confirm that regardless of the galaxy morphological mix that existed in the dataset, this unsupervised machine captures consistent features. The 27 machine-defined morphological classes show a solid division on stellar properties such as colour, absolute magnitude, stellar mass, and physical size of the galaxies. Each class has distinctive galaxy features which distinguish each class uniquely from other classes. Moreover, when comparing the machine classes with visual Hubble types, it is clear that a mix of different galaxy structures can exist in one visual morphological Hubble type. This reveals that an intrinsic uncertainty exists in visual classification schemes such as the Hubble sequence in precisely classifying galaxies. With the investigation in Chapter 6, we propose to rethink the current visual morphological classification scheme, and consider the possibility of using a novel classification scheme defined by machine learning to re-approach studies of galaxy evolution and formation from a different perspective.

"Home is now behind you, the world is ahead."
– J.R.R. Tolkien (1892-1973)

This thesis is dedicated to the heroes in my life:
my stoic mother (林玎憶),
my supportive aunt (鄭金蓮),
and my wonderful grandma (鄭楊秀鳳).

Acknowledgements

This three year PhD journey was one of the most rough trips in my life. Such a positive person like me, there was a period that I felt myself living in a dream, a dream that I would have to tie myself tightly to keep me on the ground, to prevent me from unstopping tears. Now I finally could say, all the bad dreams would pass. I woke up with great supports from my partner, from family, from people who love me so much, and from my supervisors. Then I realised that I might just achieve a goal that was not on the list.

First, I would like to express my sincere gratitude to my supervisors, Chris Conselice and Alfonso Aragón-Salamanca. I feel so grateful for the helps, supports, and patience that they have provided in every aspects during my PhD. They taught me not only a great amount of knowledge, but also an important attitude to science, to research, and to the balance between life and work.

Second, I would like to thank my partner, Bobby Clement. All these three years, he supported and took care of me every time when I had a difficult breakdown. I sincerely appreciate his great company. Third, the great thanks are given to my beloved family who have been always supporting me since I was born. As the youngest kid in the house, I have always been a trouble maker. Thank you all for giving me such a great patience and tolerance, and allowing me to have my wild adventures even though some of them are so peculiar.

Finally, I would like to give my highest gratitude to my grandma who had been so eager to see me as the Dr Cheng in the family. She passed away in the beginning of my second year PhD. I remembered that she always concentrated on listening to me talking about my studies, although she barely understood a word I said. She was just so concentrated. It seemed like that whatever I said about a star millions millions miles away, she just simply believed I would make it to wherever I wanted, even it is a star millions millions miles away. Thank you for having so much faith in me.

I understand that the accomplishment of a PhD is not the end of my academic journey but the commencement. As an astrophysicist from an island nation, Taiwan, home is now behind me, the world is ahead. A saying said “a ship in harbor is safe, but that is not what ships are built for”.

The ship is now ready to sail.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Big Data Era in Astronomy

The story is set in the ages of the data explosion, which started since the 1950s due to the fast development of all kinds of technology. In the late 1990s, another exponential growth of information happened when digital storage replaced analog storage. Since then, the amount of data generated per day in the world has reached over a trillion gigabytes (Sivarajah et al., 2017). This revolution of the storage capacity also boosted the development of large astronomical surveys and resulted in the Big Data era in Astronomy.

The Big Data era in Astronomy is presented in four different aspects: *Volume*, *Velocity*, *Variety*, and *Value/Veracity* (Zhang and Zhao, 2015). The *Volume* and *Velocity* represent the size of the data and the speed of data production as well as analysis, respectively. For the past decades, the remarkable development in computational capability and storage capacity enables the success of large astronomical surveys such as the Galaxy Evolution Explorer (GALEX; Martin et al., 2005)[1], the Sloan Digital Sky Survey (SDSS; York et al., 2000)[2], the Dark Energy Survey (DES; DES Collaboration, 2005; DES Collaboration et al., 2016)[3], and the future surveys such as the Large Synoptic Survey Telescope (LSST; Ivezić et al., 2019)[4], the Euclid Space Telescope[5], etc. The size of astronomical data is exponentially increasing so that more than hundreds of millions of galaxies are imaged in one survey. Meanwhile, the data production also significantly accelerates, for instance, the LSST will generate the size of the SDSS data for ten years in one night. The *Variety* indicates the complexity of astronomical data which can be reflected on different data types such as photometric, spectroscopic, and simulated data, or different storage formats in different surveys, etc. Finally, the *Value/Veracity* points to the quality in the data, for example, better resolution yields more information.

Since the scale of astronomical data has officially stepped into the so-called 'Big Data Era' in all four aspects, many conventional astronomical analyses be-

[1]https://archive.stsci.edu/missions-and-data/galex-1
[2]https://www.sdss.org
[3]https://www.darkenergysurvey.org/
[4]https://www.lsst.org
[5]https://sci.esa.int/web/euclid/

come challenging, particularly in terms of *Volume*, and many new approaches are carried out. One of the most successful example is the series of the Galaxy Zoo projects (Lintott et al., 2008, 2011; Willett et al., 2013) on galaxy morphological classifications for the SDSS which is also the flagship of citizen science. It allows the general public to classify galaxies by answering a series of questions based on galaxy images. The statistical analysis based upon the volunteers' votes have shown a great achievement in accelerating the time taken to classification, and providing a reliable catalogue of a large set of galaxies. Similar approaches to this are carried out in a plurality of astronomical topics, e.g., Planet Hunters[6].

Although citizen science such as the Galaxy Zoo projects accelerates the analysis that could be done by single individuals, it has an upper limit of time-saving. For example, the Galaxy Zoo projects spent around 3 years obtaining the classifications of ∼300,000 galaxies, due to the need for so many individual classifications per object. DES and LSST, for instance, would take of the order of > 100 years to classify with the Galaxy-Zoo-type projects. Therefore, machine learning techniques are introduced to astronomical studies in particular to deal with the large scale of astronomical data generated in the current and future surveys (Ball and Brunner, 2010; Baron, 2019).

1.2 Machine Learning in Astronomy

With machine learning as the main storyline of this thesis, its concept can be traced to the Turing Test introduced by Turing (1950): the programmed machine tries to convince humans that it is a human rather than a computer. In 1952, the 'Machine Learning' phrase was firstly used by Arthur Samuel who created a computer program for playing checkers which could memorise the game it had played and became better at it (Samuel, 1959). After this, the first artificial neural network - *Mark I Perceptron* (Rosenblatt, 1958) was proposed. However, as the first successful neural network, it failed to fulfill the expectation of recognising a variety of visual patterns, e.g., face recognition. The breakthrough of visual pattern recognition was not achieved until Fukushima (1980) and Fukushima et al. (1983) when a hierarchical and multilayered neural network, *Neocognitron*, was proposed. Since then, machine learning applications in visual pattern recognition have been rapidly developed, for instance, Self-organizing Maps (Kohonen, 1997), Boltzmann machines (Smolensky, 1986; Ackley et al., 1988; Hinton, 2002; Salakhutdinov et al., 2007; Salakhutdinov and Hinton, 2009), recurrent neural networks (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997), convolutional neural networks (Lecun et al., 1998; Krizhevsky et al., 2012), etc.

Machine learning techniques can be categorised through a variety of perspectives such as classification or regression problem, decision trees or Bayesian algorithms, instance-based or density-based, etc. The simplest and most general category is defined based upon whether a prior knowledge (e.g., labels) is input to the machine: 'supervised' if yes and 'unsupervised' if no. For example, a supervised machine is trained with the data represented by a group of features and

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

pre-labelled by the class assigned based on these features. Therefore, we have told the machine a kind of correlation between the input features and the classes that we intend to predict. Conversely, to train an unsupervised machine, we provide only the data presented through a set of features. In this way, the machine figures out the possible correlation between features themselves, which is unnecessary to be corresponding with any prior knowledge concluded by humans. In this thesis, we approach our astronomical studies using both types of machine learning.

Machine learning techniques started to be introduced to astronomical studies in the 1990s. They have been widely discussed in three different stages of astronomical data: (1) before observation, (2) raw data, and (3) after calibration. First, before observations are carried out, machine learning techniques are considered to help the complicated calculations in simulations such as the three-body problem (Breen et al., 2020) or to build data driven simulations such as high-fidelity synthetic data emulators (e.g., Rodríguez et al., 2018; He et al., 2019; Mustafa et al., 2019; Perraudin et al., 2019; Kodi Ramanah et al., 2020, etc). Second, machine learning techniques are applied in the pipeline for preprocessing the raw data such as object detection (e.g., Vafaei Sadr et al., 2019), signal reconstruction (e.g., Higson et al., 2019), deblending (e.g., Reiman and Göhre, 2019; Burke et al., 2019; Arcelin et al., 2020), etc. Finally, the most common machine learning applications in astronomy are used for the analysis of calibrated data such as star-galaxy separation (e.g., Odewahn et al., 1992; Weir et al., 1995; Ball et al., 2006; Kim and Brunner, 2017), anomaly detection (e.g., Baron and Poznanski, 2017; Giles and Walkowicz, 2019; Margalef-Bentabol et al., 2020), strong-lensing identification (e.g., Jacobs et al., 2017; Petrillo et al., 2017; Lanusse et al., 2018; Jacobs et al., 2019; Cheng et al., 2020b), studies of galaxy mergers (e.g., Bottrell et al., 2019; Ferreira et al., 2020), the measurement of different physical properties (e.g., Ball et al., 2007; CarrascoKind and Brunner, 2014; Ntampaka et al., 2015; D'Isanto and Polsterer, 2018; Tuccillo et al., 2017, 2018; Bonjean et al., 2019; Calderon and Berlind, 2019), etc.

In this thesis, we work on reduced and calibrated data, and the story focuses on the morphological classification of galaxies. The machine learning applications on this topic can be traced to Storrie-Lombardi et al. (1992). They applied a neural network with an input of 13 parameters indicating different galaxy features, such as stellar properties and brightness profiles, to predict the galaxy morphological types. Since then, a multitude of approaches have appeared utilising the technology of machine learning (e.g., Huertas-Company et al., 2008, 2009, 2011; Shamir, 2009; Polsterer et al., 2012; Sreejith et al., 2018), neural networks (e.g., Maehoenen and Hakala, 1995; Naim et al., 1995; Lahav et al., 1996; Goderya and Lolling, 2002; Ball et al., 2004; de la Calleja and Fuentes, 2004; Banerji et al., 2010), and convolutional neural networks (e.g., Dieleman et al., 2015; Huertas-Company et al., 2015, 2018; Domínguez Sánchez et al., 2018; Huertas-Company et al., 2019; Cheng et al., 2020a; Walmsley et al., 2020) for the morphological classification of galaxies. Amongst them, only two studies explore this topic using the concept of unsupervised machine learning (Hocking et al., 2018; Martin et al., 2020). They applied Self-organizing Maps (Kohonen, 1997) to extract representative features, and grouped the data with similar features together using

Hierarchical Clustering. In this thesis, different unsupervised machine learning approaches than the previous two studies are explored in Chapter 5 and Chapter 6 for galaxy-galaxy strong lensing systems and galaxy morphological classifications, respectively.

1.3 Galaxies and Morphological Classification

The main characters in this story, galaxies, are gravitationally bound systems of stars, gas, dust, and dark matter. Each galaxy can harbour more than hundreds of billions of stars, and have stellar masses in the range from $10^8 M_\odot$ to $10^{12} M_\odot$ and total masses from $10^{10} M_\odot$ to $10^{13} M_\odot$. Galaxies are critically important probes for the formation of stars and metals as well as the structure of the Universe. However, galaxies were thought to be nebulae belonging to the Milky Way until the 1920s, when Cepheid variables were identified in the Andromeda galaxy by Edwin Hubble. This discovery indicated that Andromeda was beyond our Milky Way.

Galaxy structure and morphology are connected with the stellar properties and formation mechanism of galaxies (Holmberg, 1958; Dressler, 1980). Visual classification of galaxy morphology has been approached since the pioneer works by Hubble (1926); galaxies are elegantly categorised into two main types: early-type galaxies and late-type galaxies. The former is composed of stars with older population, have redder colours, and shows little star formation, while the latter generally has younger populations of stars which reflected is a bluer colour and ongoing star formation, and often shows spiral structures. This system is then further revised in Hubble (1936) and Sandage (1961) to the well-known classification system called 'Hubble Tuning Fork' (Fig. 1.1).

Since then, several detailed classification systems were suggested in later works. For example, de Vaucouleurs (1959) revised the Hubble sequence with extra descriptions of structures such as bars, rings, etc using observations of the Southern sky. A number of catalogues were then published based upon the 'de Vancouleurs revised Hubble-Sandage system (VRHS)' (de Vaucouleurs, 1964; de Vaucouleurs et al., 1995a,b). Meanwhile, a luminosity class system was proposed by van den Bergh (1960), and systems focusing on the spiral structures were developed (van den Bergh, 1976; Elmegreen and Elmegreen, 1982, 1987). While many detailed classification systems are proposed, the two main morphological types of galaxies, early-type and late-type galaxies, are fundamental and valid in separating a variety of galaxy properties and formation histories.

In addition to the visual classification systems which can be subjective and intrinsically bias, more objective and quantitative relations between physical parameters, shape measurements, and galaxy morphology are carried out such as color and magnitude (de Vaucouleurs, 1961; Chester and Roberts, 1964; Aaronson, 1978; Strateva et al., 2001; Baldry et al., 2004), spectrum (Morgan and

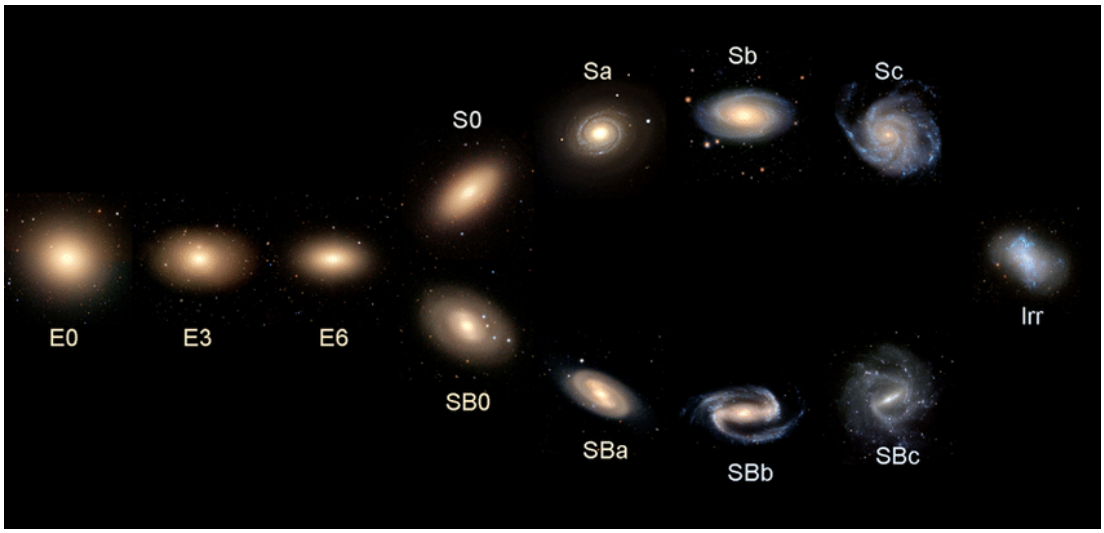

Figure 1.1: Hubble sequence classification scheme (credit: Department of Physics and Astronomy, University of Iowa)

Mayall, 1957; Madgwick, 2003), de Vaucouleurs 1/4 profile (de Vaucouleurs, 1948), sérsic profile (Sérsic, 1963, 1968), *CAS systems* (Concentration, Asymmetry, Smoothness/Clumpiness), Gini coefficient, M20, etc (Morgan, 1962; Rix and Zaritsky, 1995; Bershady et al., 2000; Conselice et al., 2000; Abraham et al., 2003; Conselice, 2003; Lotz et al., 2004; Conselice, 2006; Law et al., 2007). In this thesis, we mostly based upon the visual classifications of the two fundamental morphological types, and some of the systems listed above are used for cross-validations. Additionally, in Chapter 6, we apply an unsupervised machine learning technique to propose an objective classification and analysis without human involvement for galaxy morphology.

1.4 Thesis Overview

In this thesis we present the story of galaxy morphological classifications approached through a variety of machine learning techniques using data from several different large surveys such as the Dark Energy Surveys (DES; Chapter 2, 3, and 4) and the Sloan Digital Sky Survey (SDSS; Chapter 6). Through this, we provide a novel analysis of the capabilities of different machine learning techniques applied to the galaxy morphological classification problems as we step into the Big Data era of Astronomy.

The machine learning techniques applied in this thesis can be simply categorised into two kinds: supervised and unsupervised machine learning. Starting with the supervised machine learning techniques, we present a thorough comparison of several common supervised machine learning techniques in chapter 2 in order to determine the most optimal method for analysing the data from DES. After this analysis, in Chapter 3 a further investigation is carried out on the capabilities and performance of convolutional neural networks (CNN) applied to galaxy morphological classification. Using these CNN, the largest galaxy morpho-

logical classification catalogue of DES Year three data is presented in Chapter 4.

A new unsupervised machine learning application, combining a convolutional autoencoder and a Bayesian Gaussian mixture model, is introduced to astronomical studies in Chapter 5. This new method is applied to galaxy-galaxy strong lensing identification using simulated data for the Euclid Space Telescope (Metcalf et al., 2019a). In Chapter 6 we improve the unsupervised method proposed in Chapter 5, and extend this research to galaxy morphological classification. Specifically, we combine a recently developed technique - Vector-Quantised Variational Autoencoder by Google DeepMind (van den Oord et al., 2017; Razavi et al., 2019) - with iterative Hierarchical Clustering. The resulting new method is used to further explore galaxy morphology from a machine's perspective. A summary of this thesis and a discussion of future work are shown in Chapter 7.

Chapter 2

Finding the Optimal Supervised Machine Learning for Categorising Galaxies in the Dark Energy Survey

This chapter is based on published material by ***Ting-Yun Cheng****, Christopher J. Conselice, Alfonso Aragón-Salamanca, Nan Li, Asa F. L. Bluck, Will G. Hartley, et al. Monthly Notices of the Royal Astronomical Society, Volume 493, Issue 3, April 2020, Pages 4209–4228.*

Abstract

There are several supervised machine learning methods used for the application of automated morphological classification of galaxies; however, there has not yet been a clear comparison of these different methods using imaging data, or an investigation to maximise their effectiveness. In this chapter, we carry out a comparison between several common machine learning methods for galaxy classification [Convolutional Neural Networks (CNN), K-nearest neighbour, Logistic Regression, Support Vector Machine, Random Forest, and Neural Networks] using Dark Energy Survey (DES) data combined with visual classifications from the Galaxy Zoo 1 project (GZ1). Our goal is to determine the optimal machine learning methods when using imaging data for galaxy classification. We show that CNN is the most successful method of the ten methods in our study. Using a sample of $\sim$2,800 galaxies with visual classification from GZ1, we reach a preliminary accuracy of $\sim$0.95 for the morphological classification of Ellipticals and Spirals.

2.1 Introduction

The morphological classification of galaxies is a very important tool for understanding the history of galaxy assembly. It not only tells us about the evolution of galaxies, but it can also reveal their stellar properties, and thus their histories (see Section 1.3). Conventionally, visual assessment is the main method of galaxy morphological classification (e.g., de Vaucouleurs, 1959, 1964; Sandage, 1961; Fukugita et al., 2007; Nair and Abraham, 2010; Baillard et al., 2011). However, in recent decades, the amount of observed astronomical data has been grown at an exponential rate due to the fast development in computational capacity and observing capability. Studies in Astronomy have officially stepped into the age of the big data, and conventional methods for analysing astronomical data become challenging. In Section 1.1, we mentioned that the Galaxy Zoo projects (Lintott et al., 2008, 2011; Willett et al., 2013) achieved large scale morphological classification of galaxies by involving amateurs in the classification process. Citizen science projects such as Galaxy Zoo have thus provides an impressive number of galaxy morphological classifications which have been used in a variety of scientific studies. Nevertheless, the enormous amount of data generated in a larger astronomical survey such as the Dark Energy Survey (DES)[1] (Abbott et al., 2018) is too large to be analysed by visual assessment after all (Section 1.1). Therefore, an efficient automated classification is sought, and machine learning techniques from computational science are therefore introduced to tackle the problem efficiently.

Machine learning applications on large scale datasets have been widely discussed for the past decades, particularly, the last few years in Astronomy (Section 1.2). However, we are still learning the best ways to apply this to galaxy morphology and other areas of astronomy. Several astronomical 'challenges' were carried out to find the best solution to a specific astronomical problem which can be applying machine learning techniques or other approaches, e.g., the Galaxy Zoo challenge (e.g., Dieleman et al., 2015), the strong gravitational lensing (Metcalf et al., 2019a), the weak gravitational lensing (Mandelbaum et al., 2014), etc.

For galaxy morphology, there are now several different machine learning methods used to carry out supervised classifications; however, there is not a clear quantitative comparison between these different methods yet, especially concerning imaging data. In previous works, except for the application of CNN, there has been very few studies which directly exploited imaging data when using other machine learning algorithms, such as neural networks or support vector machine. Therefore, in this chapter, we carry out a comparison of the simplest classification – binary morphological classification of 'Ellipticals' and 'Spirals' (follows the classification of the Galaxy Zoo 1 project) – between several common methods in machine learning (listed in Table 2.1) using imaging data. We emulate the application of face and hand-writing recognition in computational science (Bishop, 2006) that directly input image pixels as features to all the methods we compared so that a fair comparison can be achieved. Additionally, this comparison also provides a decision of the most optimal technique for DES imaging data on galaxy

[1] https://www.darkenergysurvey.org/

Labels	Machine Learning Algorithms
1	K-Nearest Neighbour (KNN)
2	KNN + Restricted Boltzmann Machine (KNN+RBM)
3	Support Vector Machine (SVM)
4	SVM + Restricted Boltzmann Machine (SVM+RBM)
5	Logistic Regression (LR)
6	LR + Restricted Boltzmann Machine (LR+RBM)
7	Random Forest (RF)
8	RF + Restricted Boltzmann Machine (RF+RBM)
9	Multi-Layer Perceptron Classifier (MLPC)
10	Convolutional Neural Networks (CNN)

Table 2.1: The list of machine learning methods tested in this chapter.

morphological classification that will be discussed with more details in Chapter 3 and used for building the largest galaxy morphological classification catalogue of the DES year three data in Chapter 4.

The arrangement for this chapter is as follows. Section 2.2 describes the data resources, the procedure of pre-processing, and the datasets we use in this chapter. Each supervised machine learning method used is introduced in Section 2.3. We present the main results in Section 2.4 and the conclusion is shown in Section 2.5.

2.2 Data Sets

For the images in this analysis we use the subset of Dark Energy Survey (DES) first year (Y1) GOLD data - DES observation of the Sloan Digital Sky Surveys (SDSS) stripe 82, selected at magnitude i<22.5 and redshift z<0.7 (Drlica-Wagner et al., 2018). DES data covers 5000 square degrees ($\sim$ 1/8 sky) and partially overlaps with the survey area of the SDSS, but has a better seeing than the SDSS images from Galaxy Zoo. Dark Energy Camera (DECam) (Flaugher et al., 2015), the new installed camera used in DES, which is mounted on the Victor M. Blanco 4-meter Telescope at the Cerro Tololo Inter-American Observatory (CTIO) in the Chilean Andes, improved the quantum efficiency in the infrared wavebands (>90% from ∼650 nm to ∼900 nm), and gives a better quality images for the observation of very distant objects than previous surveys with the spatial resolution of 0.″263 per pixel and the depth of $i = 22.51$ (Abbott et al., 2018).

A DES survey image has more than 500M pixels. Each tile is 1/2 sq.-deg. The coadd (tile) images are 10000 by 10000 pixels in size with a pixel scale 0.″263. The total number of the data in this subset is around 1.87 million galaxy stamps with

photometric redshift, and photometry information in 308 i-band coadd images.

In order to train our machine learning algorithm, we match the DES data with the visual morphological classifications from the Galaxy Zoo 1 project (GZ1, hereafter)[2] (Lintott et al., 2008, 2011). We only exploit the visual classifications which have agreements (votes rates) over 80 percent and have been bias corrected by Bamford et al. (2009) for both Ellipticals and Spirals in GZ1. However, the matching of DES data with visual classifications from GZ1 only gives 2,862 objects in total, with the number ratio between Ellipticals and Spirals being 1 to 3. Their magnitude ranges from $\sim$12.5 to 18 in i-band, and the redshift z$\leq$0.25 (peak at z$\sim$0.1). To avoid overfitting while carrying out the ML training, we apply data augmentation in the pre-processing procedure in our study (Section 2.2.1.1). To improve the performance of our machine learning methods, we apply other techniques including feature extraction, i.e. Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) to extract other informative features from galaxy stamps (Section 2.2.1.3).

2.2.1 Pre-Processing

Before data pre-processing, we separate our 2,862 galaxies with DES data and the GZ1 classification randomly into training sets, and testing set, to prevent repeated galaxies in both sets. Our data pre-processing has four main steps: (1) data augmentation; (2) stamps creation; (3) feature extraction; (4) rescaling. The details are shown below.

2.2.1.1 Data augmentation

Data augmentation is of great importance while using pixel inputs in machine learning. Since Dieleman et al. (2015), data augmentation by rotating images has been widely used within CNN for the morphological classification of galaxies. In this study, we have 2,862 galaxies with visual classifications from GZ1, 759 Ellipticals and 2,103 Spirals, respectively, to train and test our methods. In order to prevent over-fitting during training, we rotate each galaxy image by 10 degrees differences from 0 to 350 degrees to increase the number of training samples. Hence, the available number of training samples increases to $\sim$100,000. After rotation, we add Gaussian noise to the rotated images (Huertas-Company et al., 2015). This noise is small enough to not to influence the visual appearance and structures of the galaxies (namely, remain the same visual classification), but it is big enough to make a detectable change of pixel values.

Although data augmentation through rotating images is a well known method used in machine learning application (e.g., Dieleman et al., 2015; Huertas-Company et al., 2015), the effect of these rotated images is unexplored. Therefore, we investigate the difference of performance between partially and fully using rotated images in the datasets in Section 2.4.2.

[2]https://data.galaxyzoo.org/

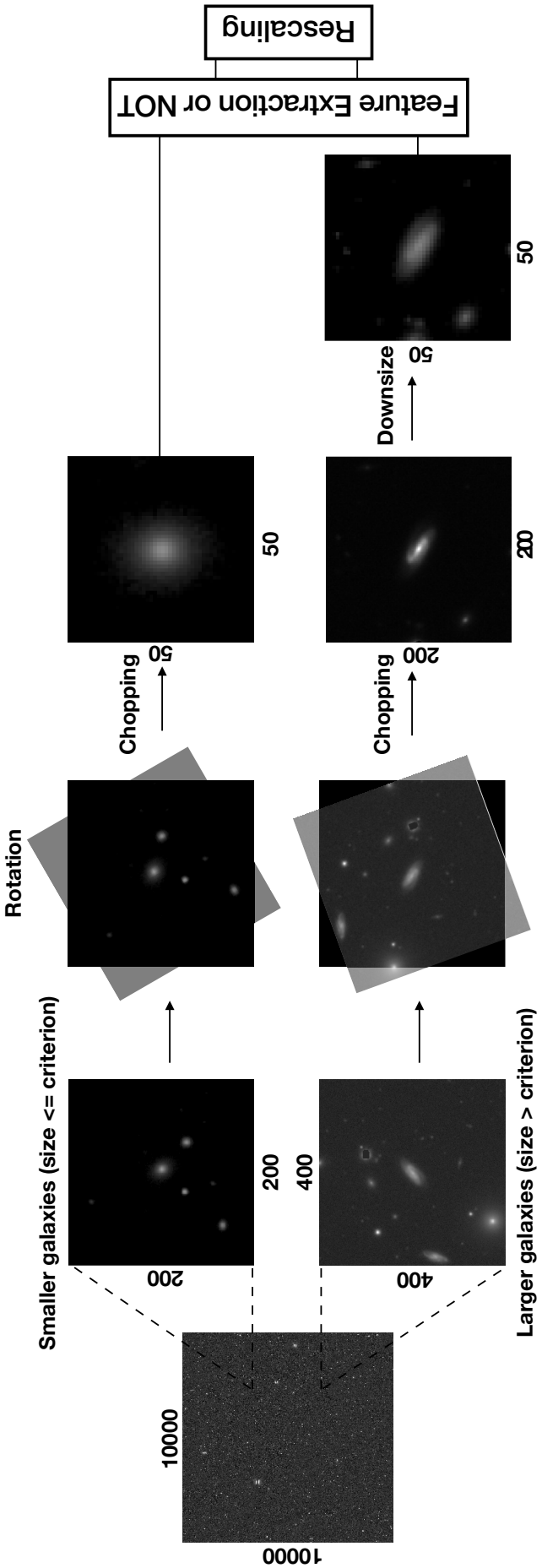

Figure 2.1: Pre-processing procedure pipeline. The pipeline starts from the initial coadd images, then we chop the coadd images into different sizes according to the size of galaxies. After rotation, we chop and downsize the images to the required sizes: 50 by 50 pixels. The details of the procedure is in Section 2.2.1

2.2.1.2 Creation of the galaxy stamps

Fig. 2.1 shows the pre-processing procedure used in our study. Using the galaxy catalogue from DES, we cut the coadd images with units of size 10000 by 10000 pixels into millions of galaxy stamps with sizes of 50 by 50 pixels. The size of galaxy stamp is based on the size distribution of galaxies in the DES Y1 GOLD data (stripe 82), where over 99% of galaxies are smaller than a threshold of 25 by 25 pixels. Therefore, the size of our stamp is 50 by 50 pixels, which is twice as large as the threshold in the size distribution of galaxies.

Fig. 2.1 shows that before chopping the stamp to the size of 50 by 50 pixels, we create the galaxy stamps with an initial size of 200 by 200 pixels when the galaxy size is smaller than 30 by 30 pixels, and 400 by 400 pixels when the galaxy size is larger than 30 by 30 pixels. For smaller galaxies, we rotate the 200 by 200 pixels stamps first, then reduce them in size to 50 by 50 pixels; for larger galaxies, we rotate 400 by 400 pixel stamps, reduce them in size to 200 by 200 pixels, then downsize them to 50 by 50 pixels by calculating the mean value of pixels in a size of 4 by 4 pixel cell. This procedure is designed to prevent empty pixel values showing up at the corner of stamps when we rotate images with non-90 degrees rotations.

2.2.1.3 Feature Extraction

In our study, we apply the Histogram of Oriented Gradients (HOG) on both our original and rotated stamps to investigate the impact of this feature extractor on supervised machine learning. The HOG is a feature extractor which is able to extract the distribution of gradients with their direction from each pixel value. It is useful for characterising the appearance and the shape of objects (Dalal and Triggs, 2005). It calculates the gradients of the horizontal (x) and vertical (y) direction of stamps. The magnitude and orientation of the gradient are calculated as below,

$$|G| = \sqrt{G_x^2 + G_y^2}, \theta = \arctan\left(\frac{G_y}{G_x}\right) \tag{2.1}$$

where $|G|$ is the gradient magnitude of each pixel, G_x is the gradient magnitude measured in x-direction, G_y is the gradient magnitude measured in y-direction, and θ is the orientation of the gradient for each pixel in the images. It then measures the contribution of gradients from each pixel in the cell with the size of 2 by 2 pixels, and uses a histogram to describe the contribution of gradient magnitude to each orientation of gradient. The input of HOG image is the direct output of this feature extraction process, and we rescale the pixel value to the range between 0 and 1 (Section 2.2.1.4). Examples of HOG images are shown in Fig. 2.2.

HOG is very popular within pattern recognition studies, e.g., human detection, face recognition, and handwriting recognition (e.g., Dalal and Triggs, 2005; Shu et al., 2011; Kamble and Hegadi, 2015, etc); however, it is not popular yet in astronomy studies for the usage of machine learning algorithms. One of the applications is the detection of gravitational lensing images (Avestruz et al., 2019a),

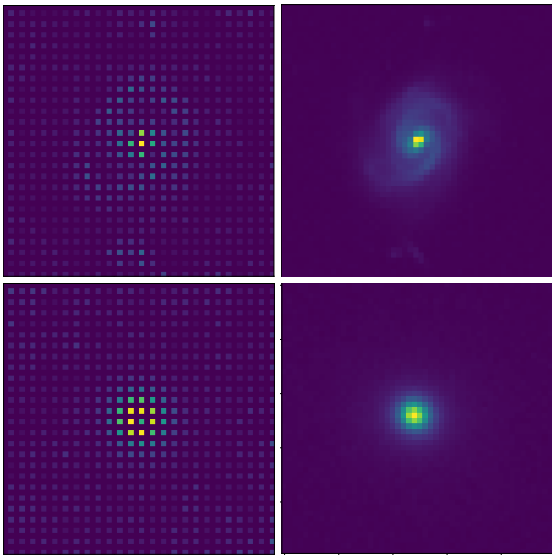

Figure 2.2: Examples of images from Histogram Oriented Gradient (HOG) with the cell size of 2 by 2 pixels. *Left*: HOG images. *Right*: original images in linear scale. *Top*: Spirals. *Bottom*: Ellipticals.

and a few previous works on the galaxy morphology (e.g., The Galaxy Zoo challenge Chou, 2014). However, none of these studies have examined the influence of HOG on the performance of machine learning algorithms. In this study, we apply HOG on our images to investigate not only the effect of it on automated morphological classification of galaxies, but also the impact of it on the performance of different machine learning algorithms (Section 2.4.4).

2.2.1.4 Rescaling

Rescaling is a very important process in the application of machine learning. Different galaxies have different brightness due to their different properties and their distances, so the pixel values of each image have significant variation between galaxies. This would cause difficulties for machine learning algorithms when defining the boundaries between different classes. Therefore, we rescale the pixel values of each image (raw and HOG images) to the range between 0 and 1 through normalising by the maximum and minimum pixel value of each image. We are aware that intrinsic brightness can be a classification criteria, including surface brightness. However, in this study we are interested in the structure only and not on other properties that might correlate with a class of galaxy such as surface brightness.

2.2.2 The datasets

In this study, we create 4 different datasets (see Table 2.2). The first two datasets (1 & 2) contain both the original images and the rotated images, and the last two (3 & 4) contain only the rotated images. This setting is used for investigating the influence of rotated images on the performance (Section 2.4.2).

labels	i (raw), ii (HOG),	iii (combination, for CNN)
1	original images+rotated images	E:S∼1:3, Training=10,448
2	original images+rotated images	E:S∼1:1, Training=11,381
3	only rotated images	E:S∼1:3, Training=11,448
4	only rotated images	E:S∼1:1, Training=12,381

Table 2.2: The arrangement of training datasets in this chapter. The content included in the datasets are shown in the second column, and the third column shows that the ratio between Ellipticals and Spirals and the total number of training data in each dataset.

On the other hand, the datasets 1 & 3 are unbalanced which contain more spiral galaxies than elliptical galaxies in the datasets while the datasets 2 & 4 have an equal number of spiral galaxies and elliptical galaxies in each dataset. We balance the number of each type by adding different numbers of rotated images to each type. For example, we rotate images of the Ellipticals 7 times, but only 2 times for the images of Spirals in the dataset 2, and 3 times for both types in the dataset 1. We use this setting to investigate the effect of the balance between the number of each type in training samples (Section 2.4.3). In addition, we also reduce the differences in the number of total training samples between each dataset to reduce the probable bias from this.

On the other hand, we have 2 (or 3 in CNN) different types of input data (i, ii, iii). The first type (i) is the raw image with linear scale, and the second type (ii) is the HOG image from feature extraction. The third type, 'combination input (iii)', is special for CNN due to the characteristic structure of CNN that we can combine both the raw images (i) and HOG images (ii) as input without increasing the number of features. This is an new way to combine data using CNN whereas people used to restore the images with different colours in the third dimension of CNN in previous studies. We then also investigate the effect of this combination input (iii) and compare it with the other two types (i & ii) (Section 2.4.4).

For the testing set, we randomly pick 500 galaxies from 2,862 galaxies for each type (Ellipticals and Spirals). The rest of unselected galaxies are training set. Therefore, we have 1,000 galaxies in total for testing and the ratio between Ellipticals and Spiral is 1:1.

2.3 Models of Machine Learning

The concept of machine learning can connect with the invention of calculators (Turing, 1950) that we program machine to obtain the information we want through the input numbers or characters (features). More introduction for the background history of machine learning is shown in Section 1.2. The breakthrough of machine learning applications in visual pattern recognition started from Fukushima (1980) and Fukushima et al. (1983). However, it was not until the 1990s, the machine learning stood on the stage of astronomical applications (e.g., Odewahn et al., 1992; Storrie-Lombardi et al., 1992; Weir et al., 1995, etc).

There are two main types of features, 'parameter input' and 'pixel input', that can be fed into machine. In the studies of galaxy morphological classification, the 'parameter input' is where we use parameters, which have clear correlations with galaxy types (e.g., Storrie-Lombardi et al., 1992; Naim et al., 1995; Lahav et al., 1996; Ball et al., 2004; Huertas-Company et al., 2008, 2009; Banerji et al., 2010; Huertas-Company et al., 2011; Sreejith et al., 2018). For example, the 'parameter' input can be surface brightness profile, colour, *CAS systems* (Concentration, Asymmetry, Smoothness/Clumpiness), Gini coefficient, M20, etc (e.g., Abraham et al., 2003; Conselice, 2003; Lotz et al., 2004; Law et al., 2007).

On the other hand, the 'pixel input' means that we treat each pixel of an image as a feature to feed machine learning algorithms. The 'pixel input' is the most straightforward feature used between the two for machine to learn, although it significantly increases the number of features for computation. Nonetheless, it is uncommon in previous studies of automated classification of galaxy morphology to use 'pixel input' (e.g., Maehoenen and Hakala, 1995; Goderya and Lolling, 2002; de la Calleja and Fuentes, 2004; Polsterer et al., 2012) until the application of CNN become popular in recent years (e.g., Dieleman et al., 2015; Huertas-Company et al., 2015; Domínguez Sánchez et al., 2018; Walmsley et al., 2020, etc).

We use 'pixel input' for each method in this study to investigate the effect of 'pixel input' on different machine learning algorithms (Table 2.1). The Restricted Boltzmann machine (RBM) (Smolensky, 1986; Hinton, 2002; Salakhutdinov et al., 2007), shown in Table 2.1, is the simplest neural network with one hidden layer, which we treat as a feature extractor for some methods in this study (Section 2.3.1).

All of the codes in this study are built on PYTHON. The main packages we use in this study are SCIKIT-LEARN[3] (Pedregosa et al., 2011) for most of methods; THEANO[4] (Al-Rfou et al., 2016), LASAGNE[5] (Dieleman et al., 2015), and NOLEARN[6] (Nouri, 2014) for CNN.

2.3.1 Restricted Boltzmann Machine (RBM)

Restricted Boltzmann Machine (RBM) (e.g., Smolensky, 1986; Hinton, 2002; Salakhutdinov et al., 2007) contains one hidden layer which is the simplest neural network architecture (more explanation for the architecutre of neural network in section 2.3.6). This is a useful algorithm for dimensionality reduction and feature learning; therefore, in this chapter, the RBM is used as a feature extractor to connect each feature for some machine learning methods (Table 2.1). It extracts the features which are more interlinked with each other before we feed them to other machine learning algorithms. The combination of machine learning algo-

[3]http://scikit-learn.org/stable/
[4]http://deeplearning.net/software/theano/
[5]http://lasagne.readthedocs.io/en/latest/
[6]https://pythonhosted.org/nolearn/

rithms such as logistic regression (Chopra and Yadav, 2017) and RBM is actually widely used in face and handwriting recognition.

In this study, the setting of RBM is identical amongst all methods that we apply a fixed learning rate (=0.001), 1,024 numbers of hidden units, and 500 iterations for RBM in training, where the learning rate determines how far to move the weights each time towards the local minimum of loss function. The number of iteration is approximately determined by where the maximum of log-likelihood is shown.

2.3.2 k-Nearest Neighbours (KNN)

K-nearest neighbours (KNN) is the simplest non-parametric machine learning algorithm (e.g., Fix and Hodges, 1989; Cover and Hart, 1967; Short and Fukunaga, 1981; Cunningham and Delany, 2007). This is one of the most common methods in pattern recognition and has several applications in clustering and classification problems (in astronomy e.g., Kügler et al., 2015). The concept of KNN is to find highly similar data, where similarity is defined by the 'distance' in the feature space between data. Parameter k is the number of nearest neighbours counted in the same group. This factor controls the shape of the decision boundary for the distribution of data.

Increasing the value of k decreases the variance in the classification but also increases the bias of the classification. We chose the value of k by plotting the accuracy (Equation 2.7) versus different values of k, and the value we ultimately use is k=5. The distance metric for calculating the distance between each data is defined by the *Minkowski metric*,

$$d = \left(\sum_{i=1}^{m} \left(\left| x_i - x_i^{'} \right| \right)^q \right)^{1/q} . \tag{2.2}$$

The x and $x^{'}$ represent the input data, and the x_i and $x_i^{'}$ values here are the features of input data. The m is the number of features. The value of q is equal to 2 in this study, namely, the metric we use is the *Euclidean metric*.

2.3.3 Logistic Regression (LR)

Logistic regression (LR) is a generalised linear model (McCullagh and Nelder, 1989) which uses the sigmoid function $\frac{1}{1+e^{-x}}$ (or logistic function) to output the probability of classification. The application in astronomy such as Huppenkothen et al. (2017) studies the variability of galactic black hole binary. The combination of LR and RBM is commonly used in face and handwriting recognition (Chopra and Yadav, 2017). The improvement of this combination is rather significant in LR while using 'pixel input' because of the characteristics of neural networks (See section 2.4).

2.3.4 Support Vector Machine (SVM)

The concept of support vector machine (SVM) algorithm is to find a hyperplane defined as below,

$$\vec{w} \cdot \vec{x} - b = 0, \tag{2.3}$$

where $\vec{w}$ is a weighted vector, $\vec{x}$ is the input data, and b is the bias, with the maximum distance to the nearest data for each type (*support vector*): $|\vec{w} \cdot \vec{x} - b| = 1$ (Vapnik, 1995; Cortes and Vapnik, 1995). For example, see the top of Fig. 2.3, where in a 2-class classification, $\{\vec{x_j}, y_j\}$, $\vec{x_j}$ is a vector which represents input data, and y_j represents the classification. The j means the j-th data. $y_j \in \{1(\text{circle}), -1(\text{square})\}$. While the parameter $\frac{b}{\|\vec{w}\|}$ determines the distance between the hyperplane to the support vectors, finding the maximum of this parameter is finding the minimum $\|\vec{w}\|$. After determining the decision boundary, data above the boundary: $\vec{w} \cdot \vec{x} - b \geq 1$ is classified as a circle, the below one: $\vec{w} \cdot \vec{x} - b \leq -1$ is classified as a square.

When using a non-linear SVM, the algorithm uses a kernel function K to the data: $\left(\vec{x}, \vec{x}'\right) \rightarrow K\left(\vec{x}, \vec{x}'\right)$ to map the data. The bottom of Fig. 2.3 shows a 2D illustration of an example of non-linear SVM with a circular transformation. In this example, we assume each point is (a_k, b_k), and we transform the data into a new feature space which is defined as $c = \sqrt{a_k^2 + b_k^2}$ (circular transformation); therefore, the decision boundary is shown as the circular shape in the input space (i.e. $a-b$ space), but shown lines in feature space (c space). In this study, we use a non-linear SVM, in particular, the Radial Basis Function (RBF) kernel function (Orr and Science, 1996): $\left(\vec{x}, \vec{x}'\right) \rightarrow K\left(\vec{x}, \vec{x}'\right) = exp\left(-\gamma \left\|\vec{x} - \vec{x}'\right\|^2\right)$.

SVM was expected to be an alternative option for the neural network due to the capability of dealing with high-dimensional data (Zanaty, 2012). The application of this in astronomy is very popular, e.g., Gao et al. (2008); Huertas-Company et al. (2008, 2009); Kovács and Szapudi (2015). There are two standard regularisation hyper-parameters for SVM: C-SVM and Nu-SVM (Scholkopf and Smola, 2001) methods. Both C and Nu are the hyper-parameter of regularisation which are related to the number of support vectors and the number of misclassification. The range of C can be any positive value, but the range of Nu is limited to 0 and 1 which is easier to control. Therefore, in this study, to get a better control of the hyper-parameter, we use Nu-SVM and apply the package from SCIKIT-LEARN - `NuSVC`. The value of nu is determined by the package `GridSearchCV` (Hsu et al., 2003).

2.3.5 Random Forest (RF)

Random forest (RF) is an ensemble learning method developed by Breiman (2001) which aggregates the results from a number of individual decision trees to decide the final classification (Fawagreh et al., 2014). Each tree is trained by a randomly picked subset from the training set. The RF is a well known machine learning technique applied in Astronomy using 'parameter input' (e.g., Dubath et al., 2011; Beck et al., 2018) but the application that directly using pixel such as our study is untested.

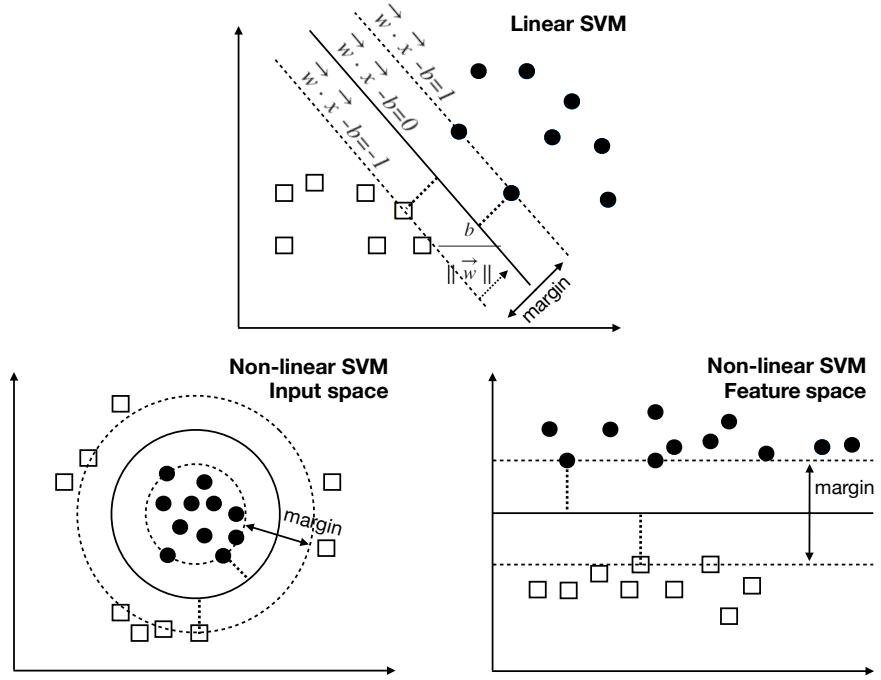

Figure 2.3: Illustration of the linear and non-linear SVM method. Different markers represent two different classifications. *Top*: linear SVM. *Bottom Left*: non-linear SVM in input space. *Bottom Right*: non-linear SVM in feature space (kernel space).

We use `RandomForestClassifier` from the SCIKIT-LEARN module. The number of trees (*n_estimators*) used in this study is determined by plotting the accuracy (Equation 2.7) versus different values of *n_estimators*, and we ultimately use 200 trees. The maximal number of features to consider for each split (*max_features*) is equal to $\sqrt{N_f}$, where N_f is the total number of features. Each tree grows until all leaves are pure or all leaves contain the number of leaves less than 2.

2.3.6 Multi-Layer Perceptron Classifier (MLPC)

Multi-layer perceptron classifier (MLPC) is a supervised artificial neural network with multiple hidden layers (Rosenblatt, 1958; Fukushima, 1975; Fukushima et al., 1983). Hidden layers which have several hidden units are invisible layers between input and output layer in neural networks, and are used to connect input features with each other. Each hidden unit is an activation function calculated by the product of weights and input. Using a neural network with one hidden layer as an example (Fig. 2.4), $X1$ and $X2$ are input features, $f1$ and $f2$ are the activation functions of hidden units calculated by (using $f1$ as an example) $f1 = f\left(w0 \cdot 1 + w1X1 + w2X2\right)$, where w are weights and f represents an activation function as well. Through the calculation, it connects each input feature with hidden units by weights. Therefore, more hidden layers and more hidden units in each hidden layer can form more complicated connections of input features; however, the architecture with more hidden layers and hidden units is more time-

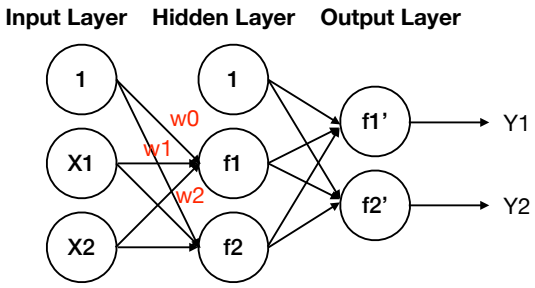

Figure 2.4: Illustration of a neural network. This structure is for illustration only and this includes one hidden layer, and two hidden units. Two input features, $X1$ and $X2$, work with the activation functions, $f1$ and $f2$, then obtain the outputs, $Y1$ and $Y2$.

consuming and can lead to overfitting problems. Similarly, the output layer also can be calculated from this concept.

MLPC uses a back-propagation algorithm (Werbos and John, 1974; Rumelhart et al., 1986), which returns the error of predicted classification compared with the true label to the algorithm when the neural network is activated and the preliminary output is obtained. Algorithm adjusts the weights through the error until the error is lower than the tolerance which we set 10^{-5}. There are two hidden layers and 1,024 hidden units for each hidden layer in MLPC method we used. The learning rate is fixed to 0.001.

2.3.7 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) started from the design of LeNet-5 (Lecun et al., 1998). However, CNN were not applied to the morphological classification of galaxies utill Dieleman et al. (2015) in the Galaxy Zoo Challenge[7]. There are two main differences between artificial neural networks (e.g., MLPC) and CNN. One is that CNN has convolutional layers which are able to extract notable features from the input images by applying several filter matrices, and the other difference is the dimension of the input.

Most machine learning algorithms are designed for dealing with 1D array input (e.g., parameter input), but some of them (e.g., SVM and neural networks) are able to deal with higher dimension data. However, the input still needs to be reshaped to 1D arrays for SVM and MLPC. On the contrast, CNN is designed for image input with three dimension arrays which means that in addition to the image itself, CNN has an extra dimension to store more information of image such as colours (RGB).

Fig. 2.5 shows the architecture of CNN that we use in this study. The input size of image is 50 by 50 pixels (Section 2.2.1.2). We have 3 convolutional layers with filter sizes of 3, 3, 2, respectively, and each of them is followed with a pooling

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

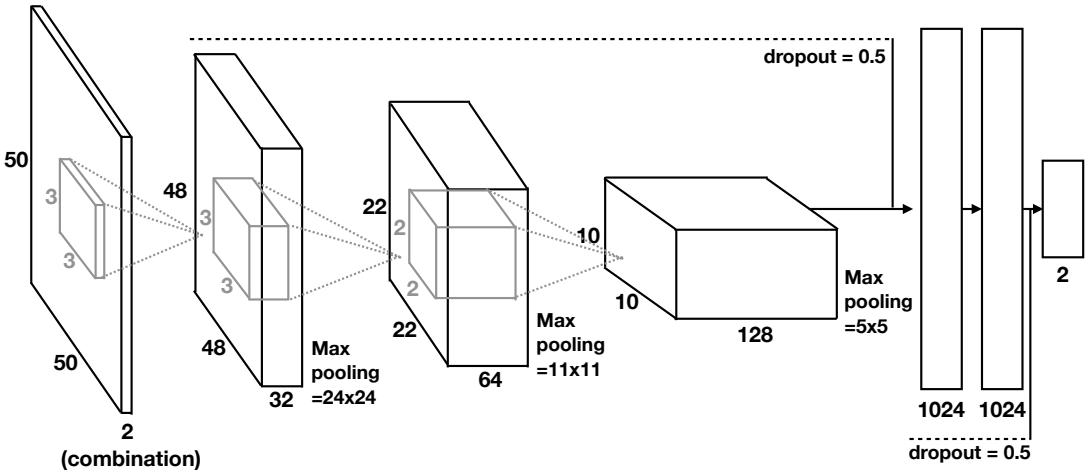

Figure 2.5: The schematic overview of the architecture of CNN. The architecture starts from an input image with size 50 by 50 pixels, then three convolutional layers (filter: 32, 64, and 128). Each convolutional layer is followed a pooling layer. Two hidden layers with 1,024 hidden units for each are following the third convolutional layer. One dropout (p=0.5) follows after the third convolutional layer and the other follows after the second hidden layer. At last, there are two outputs in our CNN, 'Ellipticals' and 'Spirals'.

layer with size 2. These are then connected with two hidden layers with 1,024 hidden units for each layer. Additionally, two dropout layers are used to prevent overfitting, one follows the third convolutional layer (pooling layer), and the other comes after two hidden layers. The rectification of non-linearity is applied for each convolutional layer and hidden layer, and the `softmax` function is applied to the output layer to get the probability distribution of each type (all from the package `lasagne.nonlinearities`). We use `Adam` Optimiser (Kingma and Ba, 2014), *Nesterov momentum*, and set *momentum*=0.9 according to Dieleman et al. (2015), and the learning rate 0.001 and maximum 500 iterations for the CNN training.

2.4 Results

2.4.1 The evaluation factors for models

We use the Receiver Operating Characteristic curve (ROC curve) (Fawcett, 2006; Powers, 2011) to examine the performance of each method and dataset. On a ROC curve the y-axis is the true positive rate and the x-axis is the false positive rate; therefore, the closer the ROC curve gets to the corner (0,1), the better the performance is. The definition of true positive and the false positive are shown in Fig. 2.6 in terms of the confusion matrix. Therefore, the true positive rate (TPR) and false positive rate (FPR) are defined as below,

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}. \tag{2.4}$$

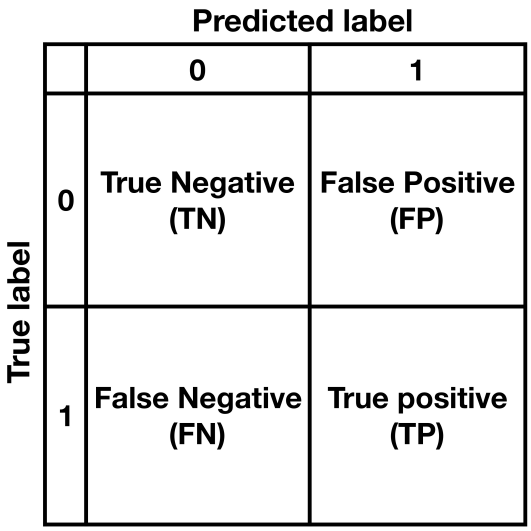

Figure 2.6: The confusion matrix. The x-axis label is the predicted label and the y-axis label is the true label. The '0' means negative as well as Ellipticals type while '1' represents positive signal and Spirals type in this study.

The definition of TPR is identical to 'recall (R)' in statistics which represents the completeness that shows how many true types have been picked, while 'precision ($Prec$)' indicates the contamination which means how many picked types (predicted types) are true types. We are doing binary classification - positive: Spirals and negative: Ellipticals. Therefore, the recalls for Spirals and Ellipticals are shown below,

$$Prec = \frac{TP}{TP + FP}; \tag{2.5}$$

$$R\,(1) = \frac{TP}{TP + FN}; \quad R\,(0) = \frac{TN}{TN + FP}. \tag{2.6}$$

Additionally, we also use the factor - the area under the ROC curve (AUC) as a performance evaluation for machine learning (Bradley, 1997; Fawcett, 2006). The meaning of AUC is the probability that a classifier ranks a randomly chosen positive example greater than a randomly chosen negative example. This factor also indicates the separability - how well the classifications can be correctly separated from each other.

2.4.2 The impact of rotated images

The ROC curves of each method and datasets are shown in Fig. 2.7. We show the results of raw images input (i) in this figure. Different colours represent different datasets such that the yellow, orange, cyan, blue lines represents datasets 1, 2, 3, 4, respectively (Table 2.2). The datasets 1 and 2 contain both the original images and the rotated images, and the datasets 3 and 4 only contain the rotated images. Meanwhile, the datasets 1 and 3 have an unbalance number of each type, conversely, the datasets 2 and 4 have an identical number for each classification. The lighter colour shadings are the scatters defined by the minimum and maximum over three reruns. The black diagonal dashed line indicates a random

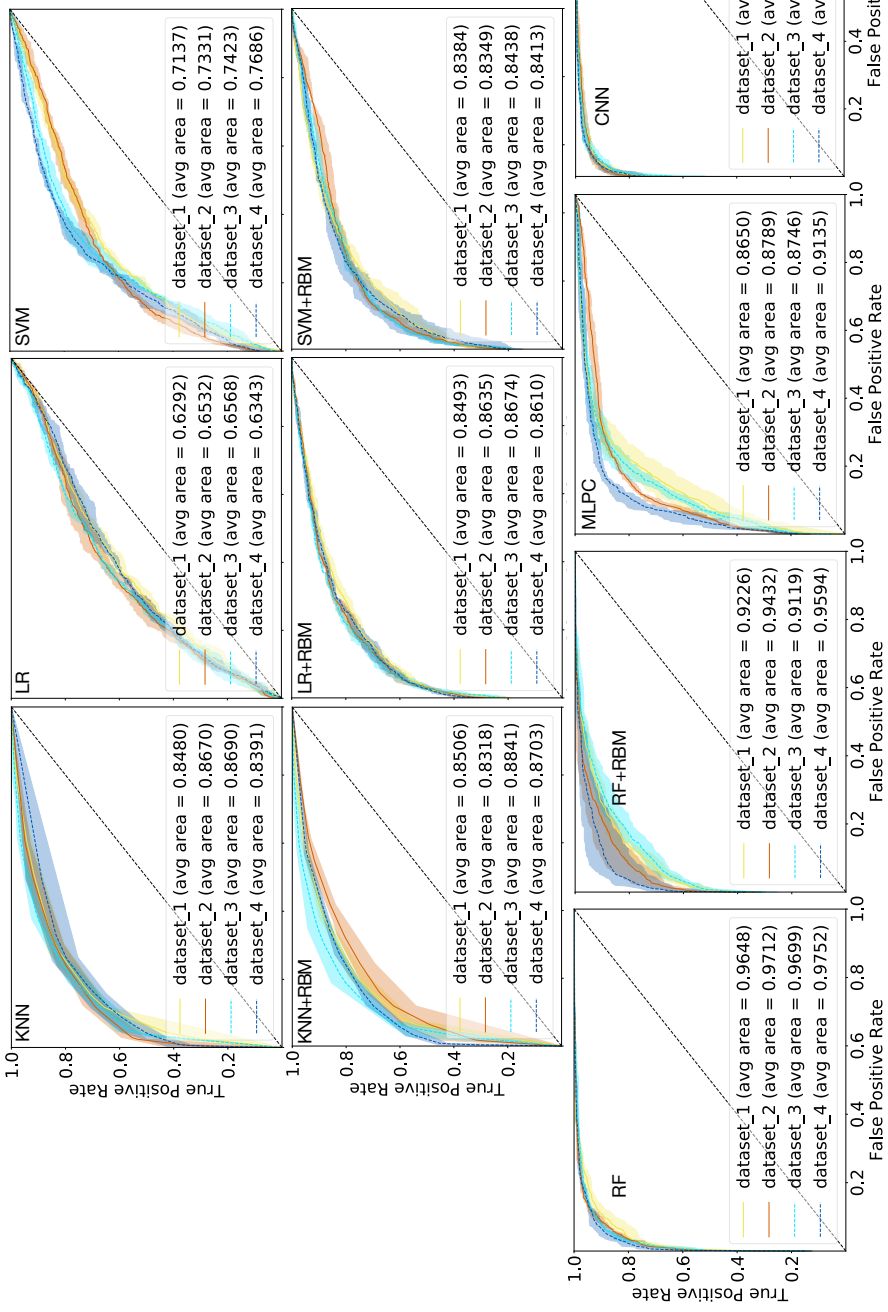

Figure 2.7: The ROC curve of each method and each dataset using the raw images input (i). The abbreviation of the methods are the same as Table 2.1. Different colours are for the different datasets (Table 2.2). Yellow, orange, cyan, blue are for dataset 1, 2, 3, 4, respectively. The lighter colour shading shows the scatters defined by the minimum and maximum of three reruns, and the lines inside are the averages of the three reruns. The black diagonal dashed line represents a random classification.

classification.

First, the results of the LR and SVM methods, with and without combining with neural network, RBM show an improvement for LR and SVM when combining with RBM in Fig. 2.7. On the contrary, the performance of RF+RBM method shows slightly worse performance than the one of the RF method. Secondly, the scatters of the three reruns show small variance for each dataset, confirming the consistency of the reruns with each other. Additionally, as can be seen there are not large differences in the results between the different datasets. However, the slight shifts of the ROC curve occur within a few methods between the different datasets (e.g., MLPC). These are due to the slight differences in the total number of training samples for different datasets (Table 2.2). For example in MLPC, the dataset 4 has the maximum number of training data within the 4 datasets used ($\sim$12400 galaxies), so the performance of this dataset is the best in MLPC; the datasets 2 and 3 have very similar number of training data (the differences in number is only 67), thus they have a similar performance to each other. The dataset 1 has the least number of training data ($\sim$10400 galaxies), therefore, the performance is relatively worse. The shifts seen are also influenced by the condition of the balance between the ratio of each type (e.g., SVM and RF), for example, the datasets 1 and 3 are the unbalanced training data, so the shape of their ROC curve are similar to each other. This is also the case for the datasets 2 and 4. To summarise, from Fig. 2.7, data augmentation through rotated images works fair to improve the performance of classification with machine learning.

2.4.3 Balance or Unbalance?

Here we investigate the influence of the balance between the number of each type in training data. Fig. 2.8 shows the recalls of Ellipticals and Spirals for the different datasets using the different methods. The colour representation is the same as the ROC curve of Fig. 2.7, and the different methods are marked by the different shape markers. We obtain the value of the recall from Equation 2.6 for Fig. 2.8 by averaging the values from the three reruns. Different pattern types represent different types of input. The colour-filled points are the raw images input (i) while the points with diagonal-filled marker are the HOG images (ii), and with dotted-filled marker are the combination input (iii). The black diagonal dashed line shows the condition that $R(0) = R(1)$ (Equation 2.6), and the black dotted lines show that the recall differences between these two types are within ± 0.1.

We observe that the unbalanced training dataset 1 (yellow) and dataset 3 (cyan) are all above the upper dotted line which means that these two datasets generally have relatively higher recalls for Spirals compared to Ellipticals, and the differences of the recalls between Spirals and Ellipticals are larger than 0.1. For example, the result of the LR with the raw images input (i) (using the dataset 3 as an example shown as the leftmost cyan square in Fig. 2.8) has the recall of (0.34, 0.81) for Ellipticals and Spirals, respectively. We also observe that the LR, LR+RBM, SVM, and SVM+RBM methods have more seriously unbalanced results than other methods when using the unbalanced datasets (close to top-left

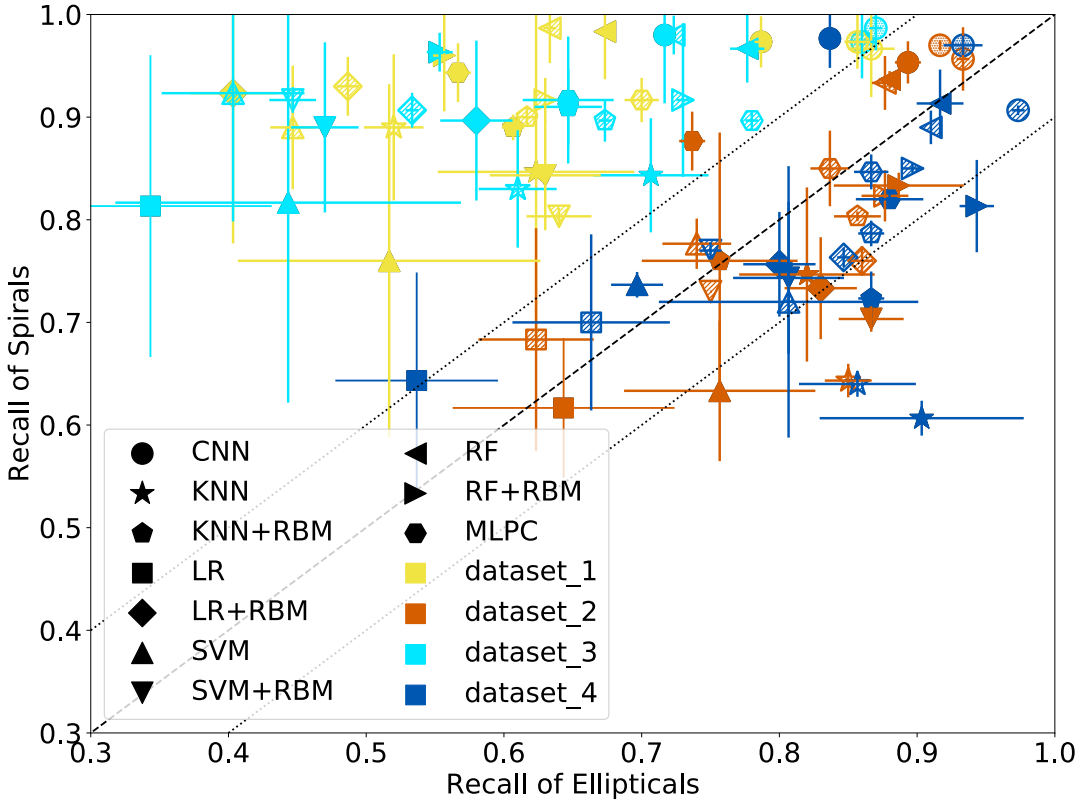

Figure 2.8: The recalls of the Ellipticals and Spirals for all methods and the different types of the input data used. The colours represent the different datasets, while the different shape markers are the different methods. The different types of filled-points represent the different types of input. The fully-colour-filled markers are the raw images only (i), the diagonal-line-filled markers are the HOG images (ii), and those with dots are the combination input of the raw and HOG images (iii) which is only for CNN. The black dashed line represents the condition that $R(0) = R(1)$ (Equation 2.6). The black dotted lines indicate that the differences in the recalls between these two types are within ± 0.1. The error bars are from the standard deviation of the three reruns.

in Fig. 2.8). This situation is due to the characteristics of these methods. For example, LR simply uses logistic functions to determine the decision boundary which can be easily shifted by unbalanced number of each type. On the other hand, Wu and Chang (2003) discusses the skewed decision boundary of SVM caused by an unbalanced data such that the decision boundary is likely to be dominated by the support vector for the majority class.

On the other hand, most of the balanced dataset 2 (orange) and dataset 4 (blue) are located within two dotted lines which implies that these two datasets have similar recalls between Ellipticals and Spirals (the differences are smaller than 0.1). However, a few results of the balanced datasets in KNN have a higher recall of Ellipticals, but a relatively lower recall of Spirals (the orange and blue stars which are below the lower dotted line). The KNN algorithm obtains the similarity between two images by calculating the 'distance' between each pixel of two images (Section 2.3.2). Spirals have various shapes (e.g., different numbers of the spiral arms) while Ellipticals have a relatively simple appearance similar to one another. Therefore, it is easier for KNN to recognise Ellipticals than Spirals when we have the same numbers of both types within the training data.

We apply ten different common machine learning algorithms in this study and they show the consistent result in their balance except for KNN which we have discussed above; therefore, according to this discussion, the balance between the number of each type in training process is of great importance while using pixel input in most machine learning algorithms. In this figure, we also observe that the CNN method with a balanced datasets obtains the best recalls of both Ellipticals and Spirals.

2.4.4 The effect of different types of input data

Here we show the comparison results between the different types of input for each method (Fig. 2.9). We have 2 (3 for CNN) different types of input - the raw images (i) , the HOG images (ii), and the combinations input (iii) (for CNN only). Different colours in Fig. 2.9 indicate different types of input such that cyan, orange, and blue are for the raw images (i), the HOG images (ii), and the combination input (iii), respectively. According to the discussions in section 2.4.2 and section 2.4.3, the results of the balanced datasets 2 and 4 are basically equivalent, and are better representations in our four datasets (Table 2.2). Therefore, we show the averages of the balanced datasets 2 and 4 after three reruns in Fig. 2.9, and the lighter colour shadings show the scatters defined by the standard deviation of the three reruns.

Fig. 2.9 shows that the HOG images input successfully improves the performance in most of methods, except for KNN. Although the HOG image is able to extract the characteristics of the morphologies according to the value of the gradients, it also loses some of the detailed information (i.e. the smaller fluctuations or gradients) and the smooth structure as well. Therefore, for KNN, the loss of the smooth structure in HOG images causes difficulties in determining the correct

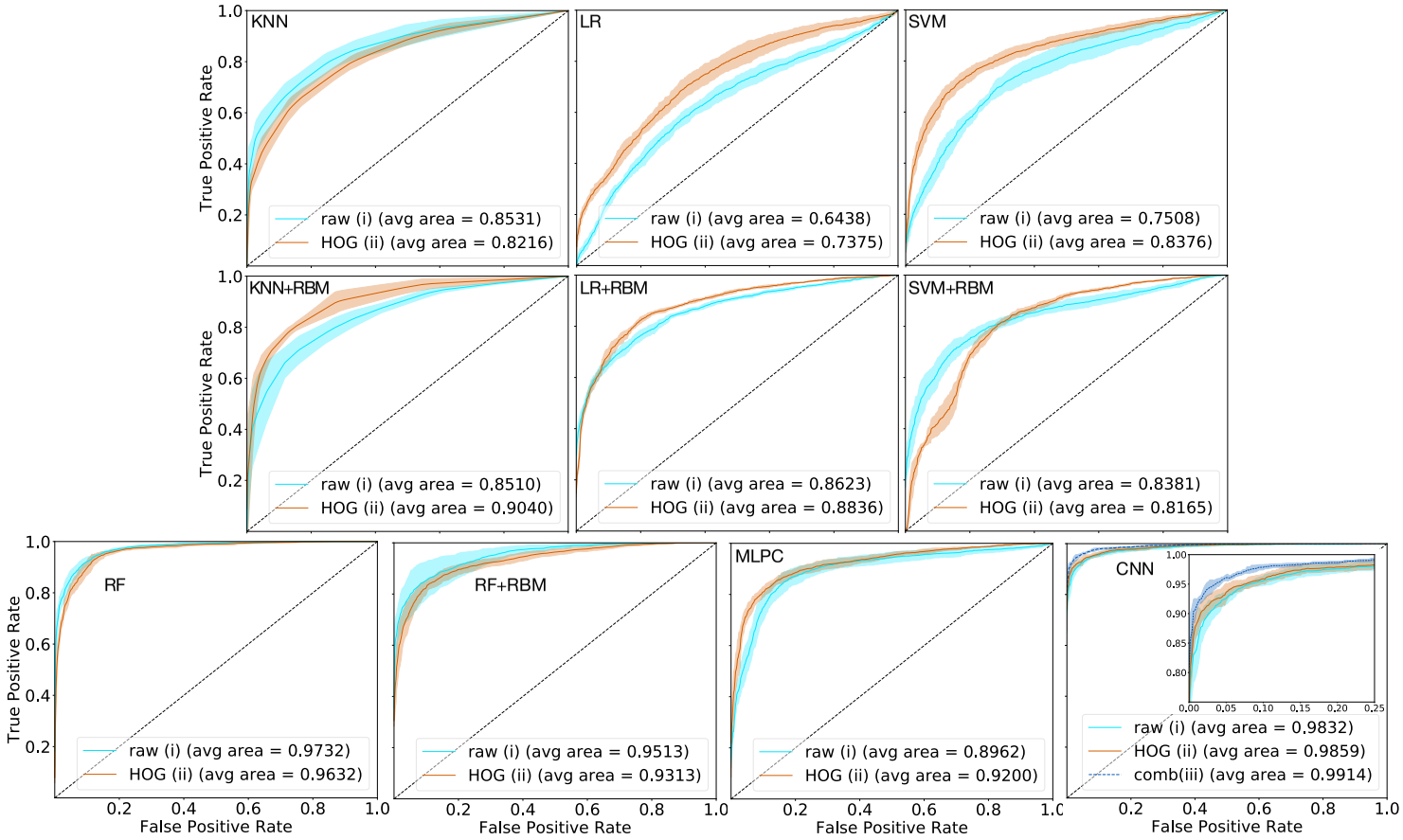

Figure 2.9: The ROC curve for different types of input within each method. Different colours are for different input types of data. Cyan, orange, and blue are for raw images (i), HOG images (ii), and combination input (iii), respectively. The lighter colour shadings show the scatters defined by the standard deviation calculated through three runs of the balanced datasets 2 and 4. The lines inside the shading are the averages of the three reruns of the datasets 2 and 4. The black diagonal dashed line represents a random classification. The subplot in the CNN method is the zoom-in area from 0.75 to 1.0 in y-axis and from 0.0 to 0.25 in x-axis.

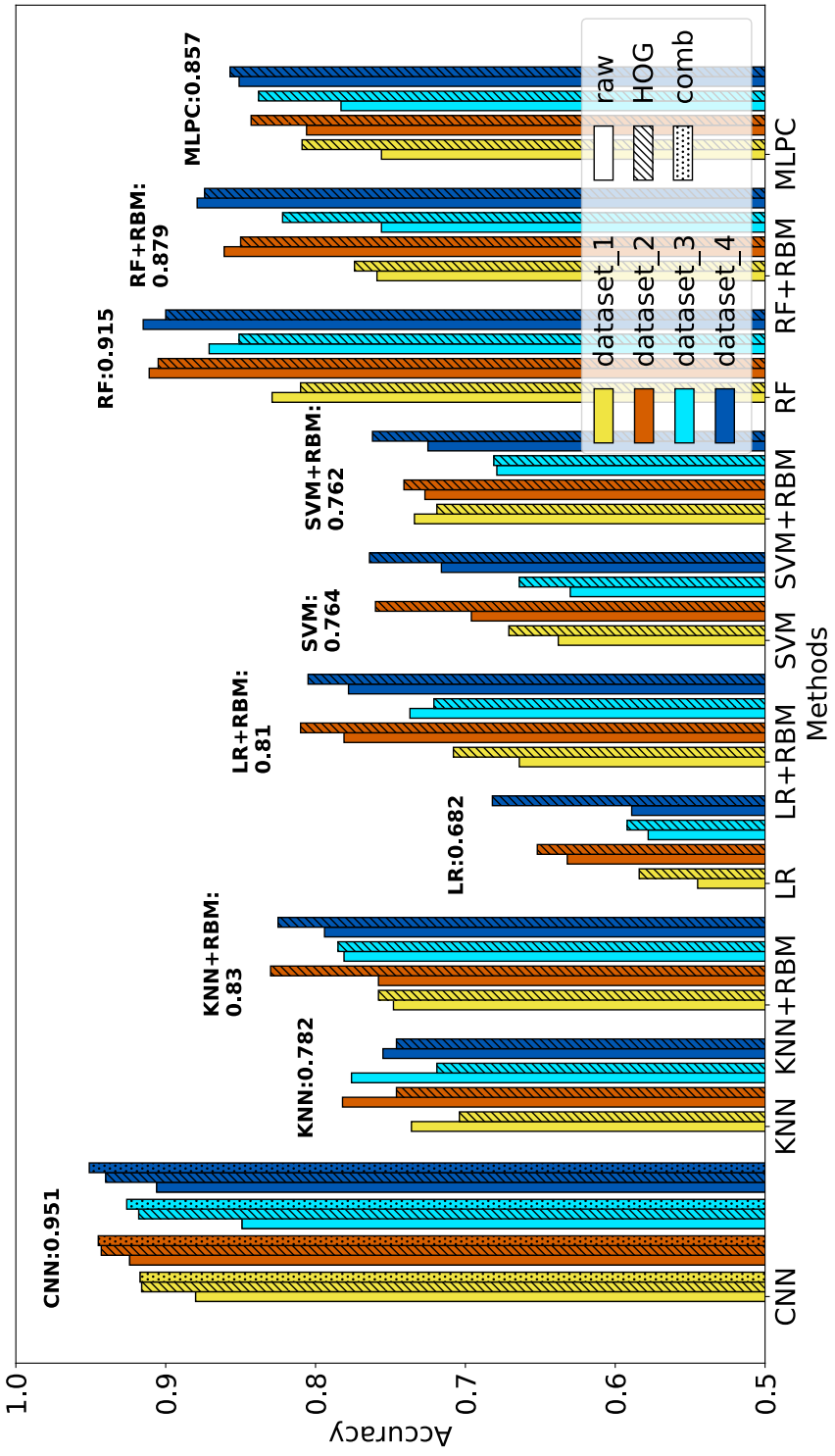

Figure 2.10: The average accuracy (Equation 2.7) of the three reruns versus each method with the different datasets and the different types of input shown. The y-axis is from 0.5 to 1.0. Colours represent different datasets such that yellow, orange, cyan, blue represents dataset 1, 2, 3, 4 (Table 2.2), respectively. The different styles of shading are the different types of input data such that the fully-filled, the diagonal-line-filled, the dotted-filled represents the raw images (i), the HOG images (ii), and the combination input (iii), respectively. The labels above bars are the highest value of the accuracy for each method.

decision boundary. This result can be significantly improved by combining KNN with RBM when using the HOG images.

On the other hand, we observe that the application of HOG images shows an unapparent effect when combining RBM in LR+RBM, SVM+RBM and RF+RBM. We infer that this phenomenon is caused by the fact that the RBM interlinks with the HOG features which have less information in the images than the raw images input. Therefore, it 'annihilates' the effect of RBM and HOG which leaves an unapparent change in these three methods. This effect is shown in both MLPC and CNN as well such that the HOG images input shows only a slight improvement in these two methods as well. However, increasing the number of hidden layers or more neurons in the neural networks helps to connect the HOG features with each other. Therefore, the improvements with HOG images in MLPC and CNN are qualitatively better than LR+RBM, SVM+RBM, and RF+RBM. A more qualitatively significant improvement is shown in CNN when we combine both the raw images input and the HOG images input (blue colour in CNN plot of Fig. 2.9).

2.4.5 Comparison between methods

The definition of the accuracy used in Fig. 2.10 is shown below,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.7}$$

such the meaning of this is defined as how many successfully classified samples there are out of all the samples tested. The comparison of the accuracy for the different datasets and the different methods is shown in Fig. 2.10. Through this figure we can observe the same situations as we have discussed in section 2.4.4 such that most methods have a better performance when using the HOG images as input, except for the KNN where the HOG image input slightly reduces the performance, and the LR+RBM, SVM+RBM, and RF+RBM methods which the HOG images input gives no apparent improvement in performance. We also make another comparison of efficiency between all methods (Table 2.3). Most methods were run on the 2.3GHz Intel Core i5 Processor with 16GB 2133 MHz LPDDR3 memory except for the 'CNN (GPU)' which was run on the NVIDIA GeForce GTX 1080 Ti GPU.

Interestingly, the performance of RF wins the performance of MLPC with a faster computation time (Table 2.3) using raw images which was totally unexpected. The further investigation for the capability of the RF on imaging data will be very helpful considering both the computing speed and a high accuracy the RF can reach. On the other hand, we can see that KNN and MLPC need less computation time but can reach a relatively good accuracy compared to other methods. Therefore, KNN and MLPC can be a good option when using pixel input. Additionally, although the KNN method has lower accuracy than MLPC, it applies raw images input which saves the preprocessing time that generates the HOG images (or other types of scaling).

Methods	Training time	Testing time	accuracy
KNN	$\sim$ 0.2 sec	$\sim$45 sec	0.782$\pm$0.027 (raw)
KNN+RBM	$\sim$3000 sec	$\sim$45 sec	0.830$\pm$0.007 (HOG)
LR	$\sim$7-8 sec	$\leq$ 1 sec	0.682$\pm$0.040 (HOG)
LR+RBM	$\sim$3000 sec	$\leq$ 1 sec	0.810$\pm$0.012 (HOG)
SVM	$\sim$800 sec	$\leq$ 8 sec	0.764$\pm$0.029 (HOG)
SVM+RBM	$\sim$3000 sec	$\leq$ 8 sec	0.762$\pm$0.001 (HOG)
RF	$\leq$1 sec	$\leq$ 5 sec	0.913$\pm$0.009 (raw)
RF+RBM	$\sim$3000 sec	$\leq$ 5 sec	0.870$\pm$0.031 (raw)
MLPC	$\sim$18 sec	$\leq$ 3 sec	0.857$\pm$0.010 (HOG)
CNN	$\sim$3000 sec	$\leq$ 5 sec	0.951$\pm$0.005 (comb)
CNN (GPU)	$\sim$360 sec	$\leq$ 5 sec	0.951$\pm$0.005 (comb)

Table 2.3: The comparison of the computing time (per $\sim$1000 galaxies) for each method. The 'accuracy' is the best accuracy shown in Fig. 2.10. The first ten methods were run on the 2.3GHz Intel Core i5 Processor with 16GB 2133 MHz LPDDR3 memory, while the sixth method 'CNN (GPU) was run on the NVIDIA GeForce GTX 1080 Ti GPU.

The most successful methods when using pixel input in our study according to both the ROC curve (Fig. 2.9) and the comparison of accuracy (Fig. 2.10) between each method is certainly CNN. Both of these two figures indicate that the HOG image input helps the performance of CNN (Table 2.4).

Additionally, we create a new way to utilise the third dimension in CNN when we combine the raw image (i) with the HOG images (ii) which together we call a 'combination input (iii)'. This shows a slight but qualitatively great improvement when using the combination input (iii) to do training in CNN (see CNN plot in Fig. 2.9). With the combination input (iii) and the balanced datasets, we can reach $\sim$0.95 accuracy with CNN using pixel input in this study (Table 2.4).

On the other hand, Sreejith et al. (2018) proposes an 'unanimous disagreement' indicating an object that all the classifiers agree with each other but disagree with the visual classification. In our study, we found only 3 galaxies out of 1,000 galaxies show an unanimous disagreement when considering all classifiers. These galaxies are all labelled as Spirals by the Galaxy Zoo 1 classification (GZ1) but classified as Ellipticals by our classifiers. We also visually confirmed that these galaxies are indeed Ellipticals. This unanimous disagreement is more likely caused by the debias process applied in GZ1 to statistically adjust the population of galaxies at a higher redshift rather than a simple visual misclassification.

2.5 Conclusion

In this chapter, we have examined ten supervised machine learning methods to determine the most successful method for classifying galaxies into ellipticals and spirals with the imaging data from the Dark Energy Survey (DES) using only pixel input on a single band (i-band). As part of the investigation, we have also tested

Input Types	accuracy	R_{01}
raw (i)	dataset 2: 0.924±0.013	0.933
	dataset 4: 0.906±0.018	0.907
HOG (ii)	dataset 2: 0.943±0.016	0.940
	dataset 4: 0.940±0.003	0.940
comb (iii)	dataset 2: 0.945±0.004	0.947
	dataset 4: 0.951±0.005	0.953

Table 2.4: The comparison between the different types of input in CNN when using the datasets 2 and 4 (Table 2.2). The total number of testing images is 1,000 galaxies. The definition of the accuracy is according to Equation 2.7. The value of R_{01} is the recall value of Ellipticals and Spiral (Eqaution 2.6) after taking a weighted average, and the value of this is shown in the table as the three reruns average of R_{01}.

how using rotated images to augment our data influences on our classification. In addition, we also confirmed that the balance between the number ratio of each type is rather important when using pixel input in machine learning.

We show that the machine learning algorithms, logistic regression (LR) and support vector machine (SVM) improve the performance of machine learning when combining with neural networks features, such as Restricted Boltzmann Machine (RBM). On the other hand, we find that using the image input along with the the Histogram of Oriented Gradient (HOG image) helps the performance in most methods, except for k-nearest neighbour (KNN). Additionally, we also observe that the application of HOG images gives less help when combining with a neural network (e.g., LR+RBM, SVM+RBM, RF+RBM) because the RBM interlinks the HOG image features which have less information than the raw images. However, increasing the number of hidden layers and neurons qualitatively helps the connection between the HOG image features according to the performance of multi-layer perceptron classifier (MLPC) and convolutional neural networks (CNN).

According to the Receiver Operating Characteristic (ROC) curve, the computing accuracy and the efficiency of each method, the performance of RF is comparable with a neural network (i.e. MLPC) with a faster computation time. In addition to RF, both the KNN and MLPC are alternative options can be considered when using pixel input because both of them have a relatively good accuracy with much less computing time than other conventional machine learning algorithms (e.g., LR, SVM) shown in this study (Table 2.3).

The most successful method within the ten methods we test is the convolutional neural networks (CNN) with the combination input of raw images and HOG images and when using a balanced training data. Through this we are able to reach an initial accuracy of ∼0.95 using ∼12,000 galaxies (including rotated images) as the training set. A further investigation of the application of CNN on the morphological classification of galaxies using the DES imaging data is carried out in Chapter 3.

Chapter 3

Morphological Classification of Dark Energy Survey Galaxies using Convolutional Neural Networks

This chapter is based on published material by ***Ting-Yun Cheng****, Christopher J. Conselice, Alfonso Aragón-Salamanca, Nan Li, Asa F. L. Bluck, Will G. Hartley, et al. Monthly Notices of the Royal Astronomical Society, Volume 493, Issue 3, April 2020, Pages 4209–4228.*

Abstract

In this chapter we present a further study of the application of the convolutional neural networks (CNN) method to the morphological classification of Dark Energy Survey (DES) galaxy images. The CNN method was identified as the most optimal supervised machine learning method among the ten methods tested in Chapter 2. We use a sample of $\sim 2,800$ galaxies with visual classifications from GZ1 and show that, using the maximal available number of the training data - which includes rotated images - and a probability threshold $p = 0.8$, we are able to improve the accuracy of our results from ~ 0.95 to ~ 0.99 when classifying galaxies into Ellipticals and Spirals.

As a part of the work, we investigate the galaxies that have a mismatched label between machine learning and visual classification, but with a high predicted probability from our CNN method. Some of them show a spiral structure in the DES imaging data that did not appear in the SDSS images. We also find that the galaxies having a low probability of being either Spirals or Ellipticals are visually Lenticulars (S0). This result demonstrates that supervised learning is able to rediscover that this galaxy class is distinct from both Ellipticals and Spirals. In our datasets, we confirm that $\sim$2.5% galaxies are mislabelled by GZ1 when using the DES imaging data. After correcting these galaxies' labels, we improve our CNN performance to an average accuracy of **over** 0.99.

3.1 Introduction

Since the work of Dieleman et al. (2015) from the Galaxy Zoo Challenge, convolutional nerual networks (CNN) have been widely applied in a variety of astronomical studies such as strong lensing (e.g., Lanusse et al., 2018; Pearson, J. et al., 2019), star-galaxy classification (e.g., Kim and Brunner, 2017), galaxy mergers (e.g., Pearson et al., 2019; Bottrell et al., 2019; Ferreira et al., 2020), and galaxy morphology (e.g., Huertas-Company et al., 2015, 2018, 2019; Domínguez Sánchez et al., 2018; Walmsley et al., 2020). These studies have shown the capability of CNN in capturing meaningful features from images and their usefulness for processing a variety of astronomical imaging data.

In Chapter 2, we also concluded that the best method in terms of classification accuracy using imaging data from all of the supervised machine learning methods tested (Table 2.1) is CNN. Our CNN (see details in Section 2.3.7) provides a preliminary accuracy of $\sim$0.95 for classifying galaxy morphology with the imaging data from the Dark Energy Survey (DES) and the visual classification from the Galaxy Zoo 1 project (GZ1; Lintott et al., 2008, 2011) when using a balanced training set (see Table 2.2 in Section 2.2) with a mixture of raw images and the images processed through the feature extractor, 'Histogram of Oriented Gradient' (HOG images; Section 2.2.1.3).

The DES imaging data has a better resolution and is deeper than images from the Sloan Digital Sky Survey (SDSS; see Section 2.2). With our CNN, these properties of DES data help us to build a larger, deeper, and better catalogue of galaxy morphologies containing the largest sample to date (Chapter 4). We will also investigate misclassification issues when comparing machine-learning results with Galaxy Zoo 1 (GZ1) labels, particularly for galaxies with high predicted probabilities from the CNN method.

In this chapter, we further analyse the CNN results obtained from Chapter 2, and improve the performance of our CNN by increasing the number of training data and applying a probability threshold. The analysis is shown in Section 3.2. As part of the work, we also investigate galaxies which 'fail' by our CNN algorithms in Section 3.3. The conclusion is drawn in Section 3.4.

3.2 Analysis of Convolutional Neural Networks (CNN)

Here we discuss in more details for the results of our CNN classification from Chapter 2. We used a default criterion for the classification in CNN such that the probability (p) > 0.5 is the criterion for classification; namely, Ellipticals or Spirals with $p > 0.5$ will be classified as that type in Chapter 2. We then change the criterion to $p \geq 0.8$, namely, any types with $p \geq 0.8$ are classified as the predicted type, and if both types have $p < 0.8$ then that galaxy will be classified as 'Uncertain type'. With this criterion, we separate our testing data into three different classes: Ellipticals, Spirals, and Uncertain. Furthermore,

	accuracy	R_{01}	$N_{\text{classifiable}}$	N_{uncetain}
dataset 2 ($p \geq 0.8$)	0.974±0.004	0.973	912	88
dataset 4 ($p \geq 0.8$)	0.974±0.003	0.973	927	73
Max ($p \geq 0.8$)	0.987±0.001	0.99	958	42

Table 3.1: The average result of the classification success with the classification criterion $p > 0.8$ through using CNN for dataset 2, dataset 4 (Table 2.2), and the result of the maximum available number of training data in our study with the combination input (iii) which includes both raw and HOG images. The total number of testing galaxies is 1,000. The definition of accuracy (Equation 2.7) and the meaning of R_{01} are same as in Table 2.4. $N_{\text{classifiable}}$ and $N_{\text{uncertain}}$ are the number of testing data which are classifiable (namely $p \geq 0.8$) and uncertain (probabilities of both types (p) < 0.8), respectively.

with the combination input (iii), the accuracy of classification increases to ∼0.97 (Table 3.1).

Secondly, increasing the number of training samples should intuitively improve the performance; however, we investigate whether this assumption is correct. We increase the number of our training samples by the rotated images, and keep the balance between the number of both types of galaxies. The maximum balanced number of the training data used in this study is 53,663 (S: 26,839; E: 26,824).

In Fig. 3.1, we observe that the increased rate of accuracy remains basically positive, but this decreases as the number of training data increases. This shows that there is likely a maximum accuracy limitation within the CNN method for galaxy classification. This figure also shows that our combination input (iii) (Table 2.2) has a better performance than the other two types of input data as we increase the number of training data, and the combination input (iii) is the only one which is able to reach over the accuracy of ∼0.97 without any condition.

Finally, we apply the maximum number of our training data (53,663) with the combination input (iii) to do the training, and combine it with the classification criterion $p = 0.8$. We then obtain a high accuracy of ∼0.987 in the morphological classification of galaxies. The result is shown in the third row of Table 3.1.

3.3 Origin of Classification Failures

As shown in the above section, we are able to reach a high classification accuracy of ∼0.987 by using CNN with the maximum number of the training data with a combination of input (iii) (Table 2.2), and the criterion of the probability $p \geq 0.8$. However, the < 100 percent accuracy indicates that there are a few galaxies misclassified but with high predicted probabilities ($p \geq 0.8$). On the other hand, there are also a few galaxies (∼42 out of 1,000 testing galaxies) which are non-classifiable (lower predicted probability $p < 0.8$ in both Ellipticals and Spirals). Table 3.2 shows the fraction of the samples within a range of probability (out of 1,000 testing galaxies), and the number of misclassification

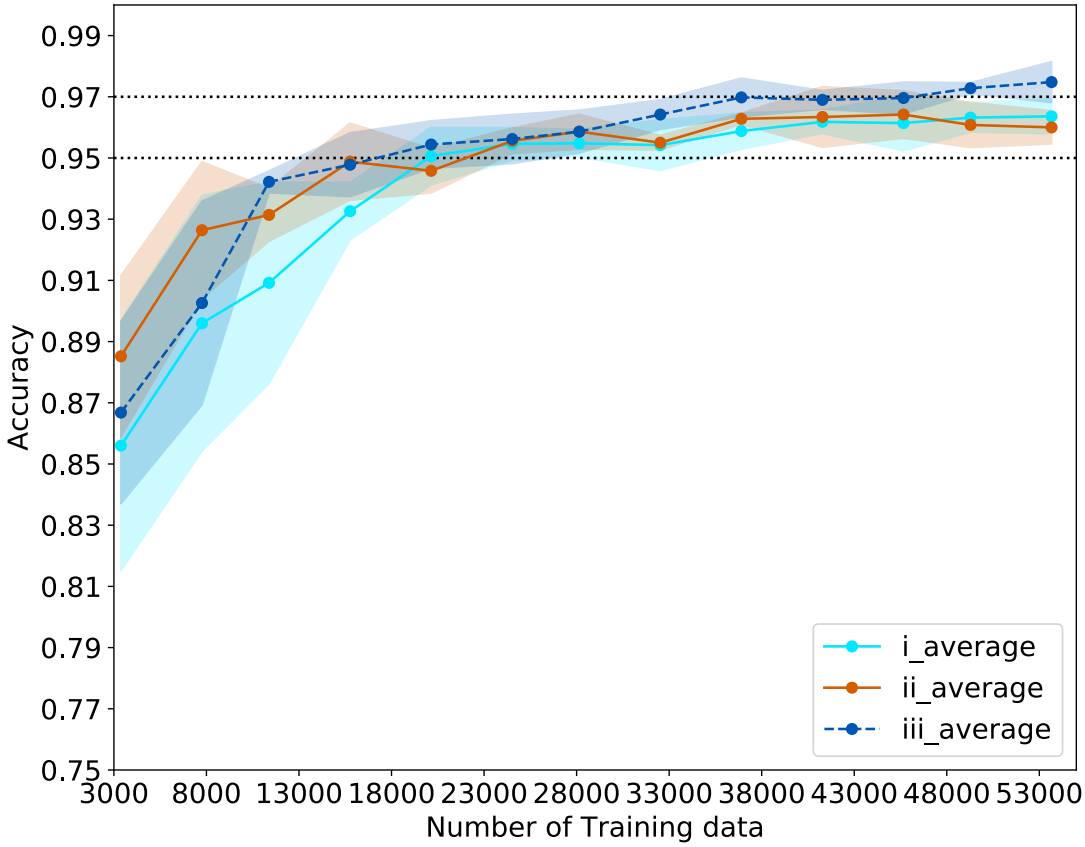

Figure 3.1: The accuracy versus the number of training data with different types of input. Different colours show different types of input such that cyan, orange, blue are for the raw images (i), the HOG images (ii), and the combination input (iii), respectively. The lighter colour areas show the scatters of the standard deviation calculated by the five reruns, and the lines inside shadings show the average of the five reruns. The two dotted horizontal lines indicate the accuracy of 0.95 and 0.97.

probability	sample fraction	misclassification
$p \geq 0.8$	0.958	0.0142
$0.7 \leq p < 0.8$	0.0184	0.239
$0.6 \leq p < 0.7$	0.0302	0.132
$0.5 \leq p < 0.6$	0.0114	0.368

Table 3.2: The fraction of the samples out of 1000 testing galaxies, and the fraction of misclassification within a certain probability range calculated by being divided by the sample number. The results are the average of five reruns.

out of the galaxies within a probability range. It indicates that the classifications with higher probabilities ($p \geq 0.8$) are much less often misclassified. However, it also shows that galaxies with the predicted probabilities between 0.7-0.8 have a higher misclassified rate than the predicted probabilities between 0.6-0.7. This means that there are some galaxies with relatively higher predicted probabilities but which have different morphology labels compared with the GZ1.

In this section, we define two types of failures by our CNN. One is the misclassification that are galaxies which were classified with high probabilities with CNN ($p \geq 0.8$) but which later turned out to have a different classification in Galaxy Zoo. The other type of 'failed' classification are those galaxies with low predicted probabilities ($p < 0.8$ in both types) of being either elliptical or spiral. We then investigate the origin of these 'failures' in this section.

3.3.1 The failure with high probability: the misclassification of the classifiable galaxies

We rerun five times the best combination of our method (i.e. the CNN trained by the maximum balanced number of training data and the combination input (iii) (Table 2.2), and classified by the criterion $p = 0.8$), and we then collect all the misclassification of the classifiable galaxies from these five reruns together, obtaining 22 galaxies in total (Fig. 3.2). Misclassification in this sense is that what we get from our CNN analysis differs from the Galaxy Zoo classification. Most of these 22 galaxies are repeatedly misclassified between these five reruns, in Fig. 3.2, objects 1-7 only show up once, objects 8-17 are repeated more than twice, and objects 18-22 are repeatedly showing up in five reruns.

There are two main probable reasons for these misclassifications with a high probability through our CNN method. One is that we use the galaxy images with linear scale (including HOG images) on our CNN training, so in some cases, even if it shows the feature of Spirals in logarithmic scale, it is just a point source, a round object, or a large bright area in linear scale. Therefore, they prefer to be classified as Ellipticals rather than Spirals in our CNN. This will be further discussed in the section 3.3.3. The other reason for the differences is due to the incorrect labels from the GZ1 which revealed because of a better imaging data used in this study. We apply visual classifications which have over 80% agreement between volunteer classifiers in the GZ1 catalogue in which we use to label our data from the Dark Energy Survey (DES). When we compare the Sloan

Digital Sky Survey (SDSS) imaging to the DES imaging, we can see some GZ1 classifications based on the SDSS data were simply wrong. Some examples are shown in Fig. 3.3. Most of them are revealed to be misclassifications due to the better resolution and deeper depth of the DES data than the SDSS data. With higher resolution of the DES data, we reveal more detailed structure than the SDSS data (e.g the number 4 and 8 in Fig. 3.3 which show clear spiral structures in the DES data but nothing in the SDSS data). We will further discuss this in Section 3.3.4.

On the other hand, we also discover that some galaxies with large, bright, and oval structure are easy to misclassify using our method. These galaxies are lenticular galaxies when examined on the DES imaging. The main reason for their misclassifications is because there is not a class for lenticular galaxies in the Galaxy Zoo 1 project. Lenticular galaxies are difficult to see by visual classification and typically require high resolution and deep imaging, even for nearby galaxies. Some of them are therefore classified as Spirals, and some of them are recognised as Ellipticals in the GZ1 catalogue. The details will be discussed in the next section (Section 3.3.2) as most of these galaxies generally have lower predicted probabilities of being either elliptical or spiral.

3.3.2 The failures at low probability: Uncertain type

In this section, we investigate the galaxies with lower predicted probabilities ($p < 0.8$) for classification as either elliptical or spiral in the five reruns of our best method. The majority of the samples with lower probabilities are repeated between five reruns, and some of them also show up in the previous section (Section 3.3.1) which are misclassified but with high probabilities. The probabilities of these galaxies vary significantly between each rerun.

The appearance of these galaxies can be separated into two types. One type are the galaxies which look large, oval, and bright (*Top 1-12* in Fig. 3.4), and the other type are those which do not appear this way, e.g., galaxies which are relatively fainter or with large bulge and spiral structure at the same time, or the target galaxy is shifted significantly away from the centre of the image (*Bottom 1-12* in Fig. 3.4).

The galaxies with large and oval structure are lenticular galaxies which we discussed in the previous section (Section 3.3.1). As discussed there is not a lenticular galaxy class in the GZ project, nor can these types be easily seen in SDSS data, therefore, the classification of these galaxies in the GZ1 catalogue are such that half of them are classified as Spirals, and half of them are classified as Ellipticals. Because lentinculars are neither spirals or ellipticals, their structure confuses our CNN such that it gives lower probabilities for these galaxies to be of either type. This is a 'rediscovery' of lenticulars, and shows the power of machine learning for discovering new types of galaxies, as we did not expect this to occur.

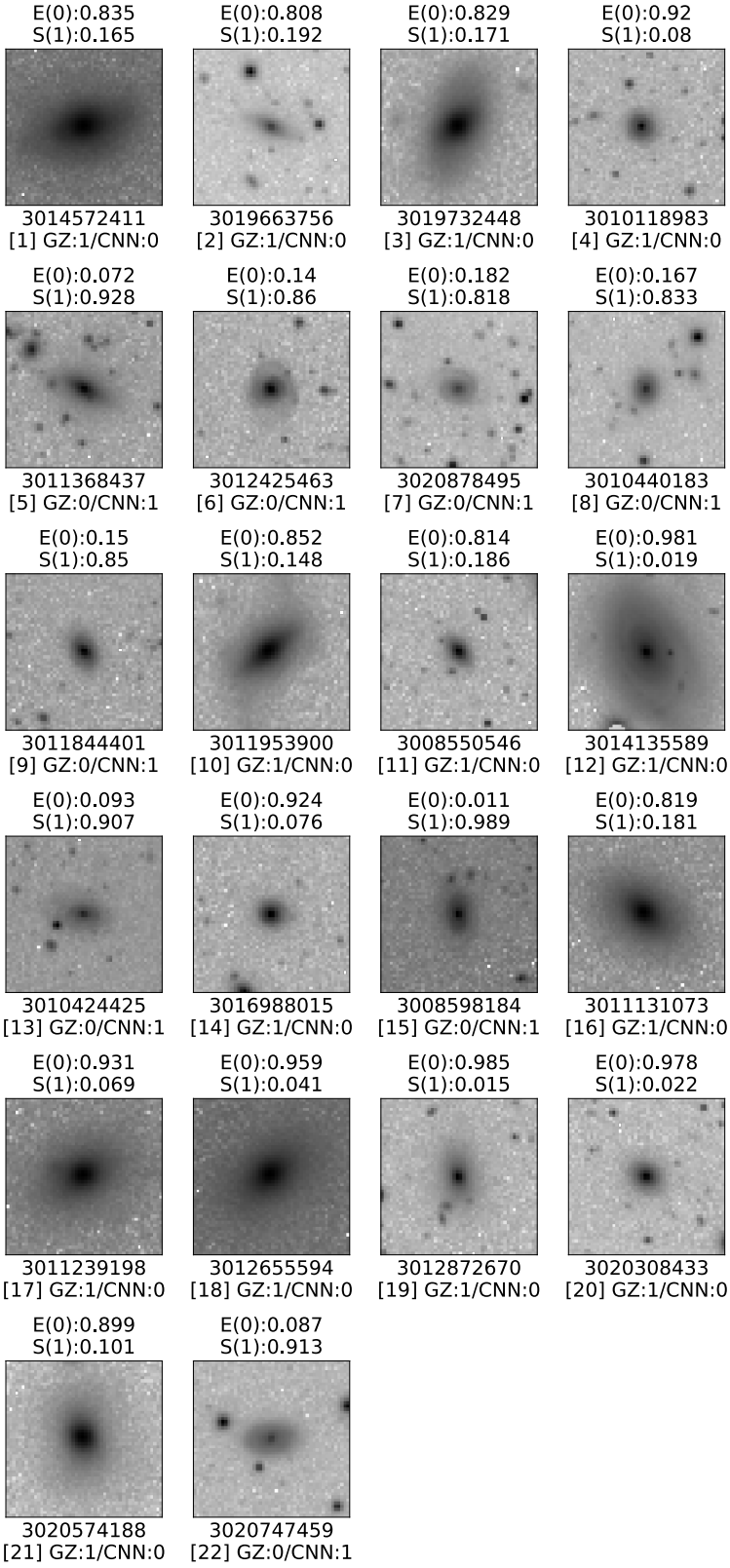

Figure 3.2: The misclassified galaxies with high probabilities ($p \geq 0.8$) comparing the classification of Galaxy Zoo 1 and our CNN. On the top of the images shows the probabilities of being Ellipticals, E(0) and Spirals, S(1) by our CNN. The line below the image shows the ID number of the galaxies in Dark Energy Survey (DES), and the second row shows the classifications by Galaxy Zoo and our CNN.

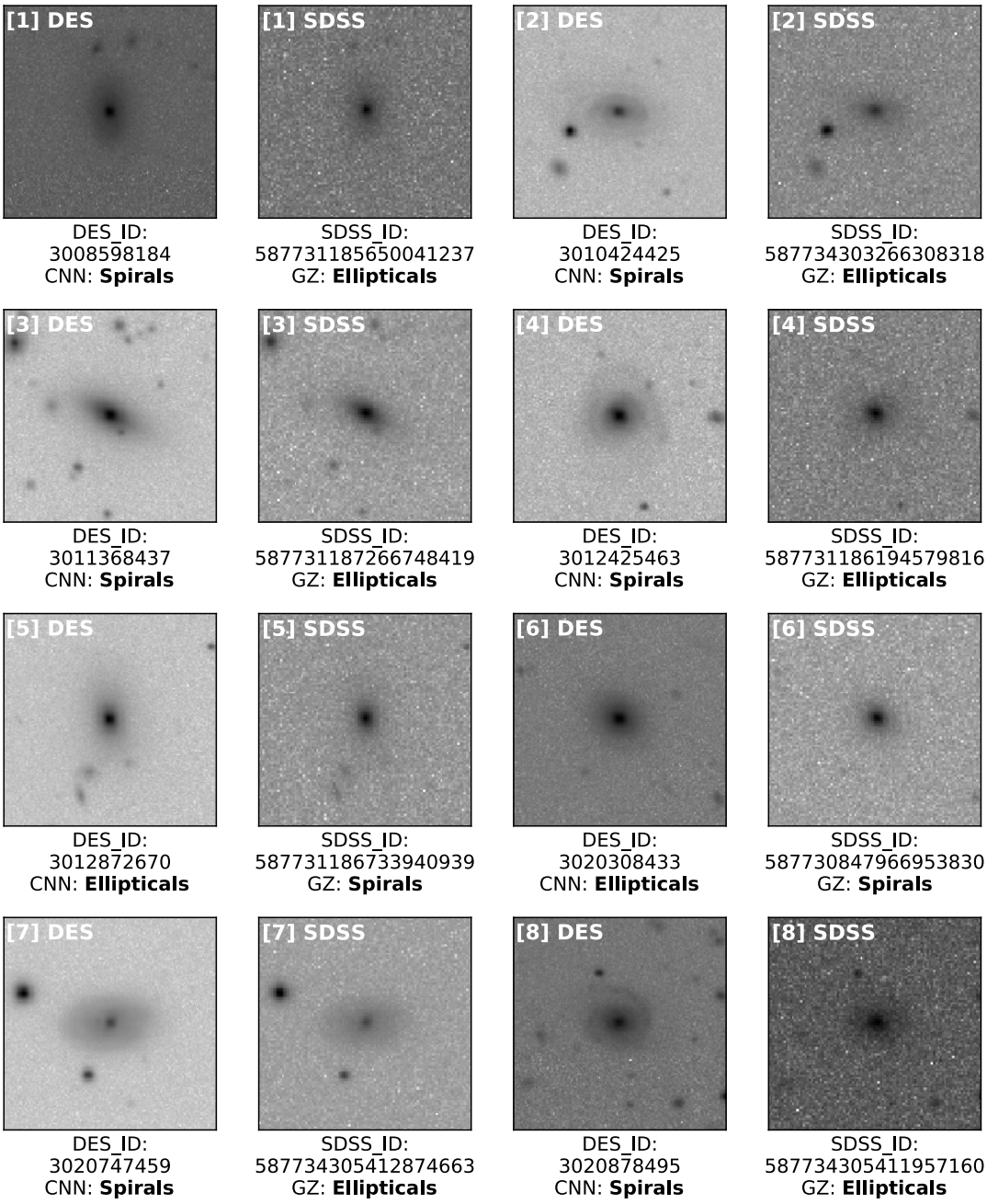

Figure 3.3: Examples of the incorrect label from GZ1 with SDSS imaging. The figures under each number show the galaxy images of DES and SDSS, and their ID numbers. The label of 'CNN' shows the predicted label from our method, and which of 'GZ' shows the label from the Galaxy Zoo 1 catalogue.

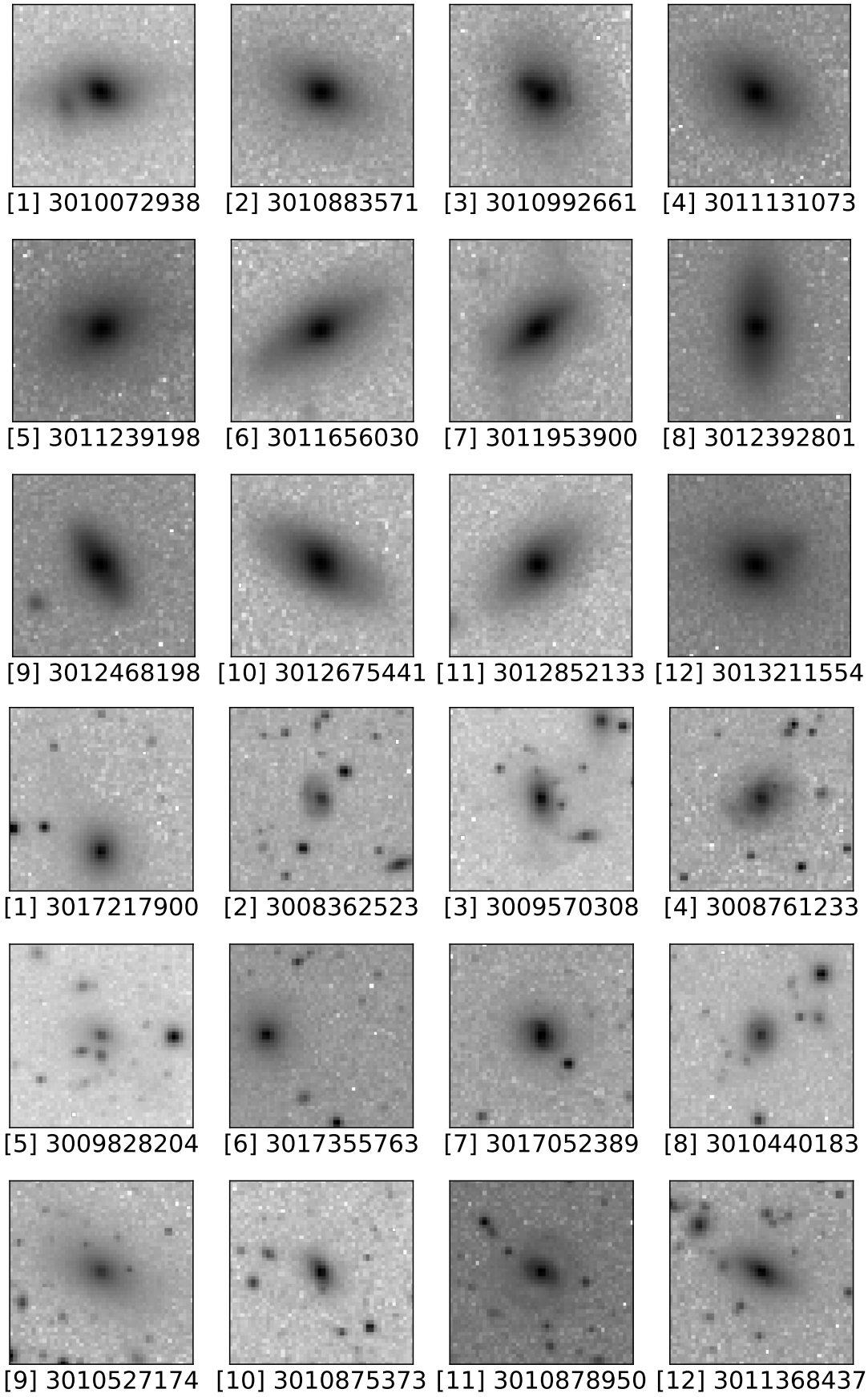

Figure 3.4: Examples of the galaxies with low probabilities of classification as either spiral or elliptical. *Top 1-12:* these objects are turned out to be lenticular galaxies (S0) in cluster inspection. *Bottom 1-12:* the other types of galaxies.

	log image		combination input(iii) +log image	
	accuracy	R_{01}	accuracy	R_{01}
dataset 2	0.950±0.006	0.947	0.952±0.006	0.950
dataset 4	0.954±0.004	0.953	0.964±0.007	0.967
Maximum	0.973±0.002	0.970	0.971±0.005	0.973
Max ($p \geq 0.8$)	0.987±0.004	0.987	0.987±0.003	0.987

Table 3.3: The comparison of the accuracy (Equation 2.7) and the recalls (Same as Table 2.4) between the inputs of the log images and the combination of log images and combination input (iii) by using the dataset 2, dataset 4 (Table 2.2), and the maximum number of training data.

3.3.3 Combined with logarithmic scale images

According to the discussion in section 3.3.1, we investigate the impact on our classification with CNN when using images with a logarithmic scale (hereafter, log images) to train our CNN algorithm by using datasets 2 and 4 (Table 2.2). In addition to the log images, we also combine the log images with our combination input (iii) as the input to train our CNN. The comparison of the results are shown in Table 3.3.

Comparing Table 3.3 with Table 2.4 shows a significant improvement when using the log images, and the combination of the log images and our combination input (iii) shows a better accuracy than just using the log images as input. However, comparing the row of '*Max ($p \geq 0.8$)*' Table 3.3 with Table 3.1 shows that there are not significant differences in the performance when we train our CNN through the maximum available number of the training data and apply a classification threshold of $p = 0.8$. This means that there is an intrinsic limitation of our method. This limitation can also be seen in Fig. 3.1 in Section 3.2.

We conclude that although adding the log images as input helps the performance, it still has no apparent difference from our result when we apply the maximum number of training data to our CNN.

3.3.4 The advantage of Dark Energy images and the misclassifications by Galaxy Zoo project

We have discussed the incorrect labels by Galaxy Zoo in previous sections. As discussed, the main reason to reveal the misclassification by SDSS imaging Galaxy Zoo is because of the better resolution ($0.''263$ per pixel) and deeper depth of DES data ($i = 22.51$) (Abbott et al., 2018).

These wrong labels not only influence the results of our CNN, but also contaminate the training set. Therefore, we remove the potential misclassified galaxies from the training set. We purify our training set by excluding the suspected misclassified galaxies then use the criteria shown in Table 3.4 to confirm or dismiss our suspected misclassifications. We then rerun our CNN classification five

	Criteria:
Confirmed	(1) Appearing ≥ 4 times in total failures. (2) Appearing at least once in the high-p failures.
Suspected	(1) Appearing ≥ 2 but ≤ 4 times in total failures. (2) Does not satisfy the criteria for 'confirmed'.
Not misclassification	(1) Appearing ≤ 1 time in the test of new models

Table 3.4: The criteria for selecting the suspected misclassified galaxies by the Galaxy Zoo project and purifying the training set.

times on each new training set and obtain five new CNN models on the new classifications. After carrying out this purification twice, and then retraining and updating our list of suspects, we obtain two lists of these galaxies: one is the confirmed misclassified galaxies by the Galaxy Zoo, and the other are the suspected misclassified galaxies.

The images of these systems are shown in Fig. 3.5 and Fig. 3.6. There are $\sim 2.5\%$ misclassified galaxies in the Galaxy Zoo 1 catalogue out of 2,800 in our study as revealed by using DES images and our CNN, and $\sim 0.56\%$ are suspected candidates in our study. We then correct our training set according to these two lists. We change the label of the confirmed misclassified galaxies, and exclude the suspected misclassified galaxies from the training set, then do the training with the maximum available number which is 53,141 galaxies in total (E: 26,344; S: 26,797). We then change the label of the confirmed misclassified galaxies in the testing set as well.

The results are shown in Table 3.5. The first row of Table 3.5 is the testing result excluding 8 suspected misclassified galaxies out of 1,000 testing galaxies. Comparing this result with the results in Table 3.1, our new models predict the highest accuracy, and end up having a resulting fewer number of uncertain type (about half the original number) than the previous results. Therefore, Fig. 3.7 shows the best testing result in our study. In this result, we change the label of the confirmed misclassified galaxies and exclude the suspected misclassified galaxies in testing set. We obtain the accuracy of 0.994 for the best model within five reruns, and the average accuracy of five reruns is 0.991.

The second and third rows of Table 3.5 show the results including suspected galaxies which retain the initial label from the Galaxy Zoo in test and change the label of them to the opposite label, respectively. We have lower accuracy in these two conditions than the result of the first row. This indicates that part of our suspected galaxies have incorrect labels in Galaxy Zoo catalogue, and part of them are not, based on our CNN. Some examples of the successful classifications by the purified CNN training are shown in Fig. 3.8 and Fig. 3.9.

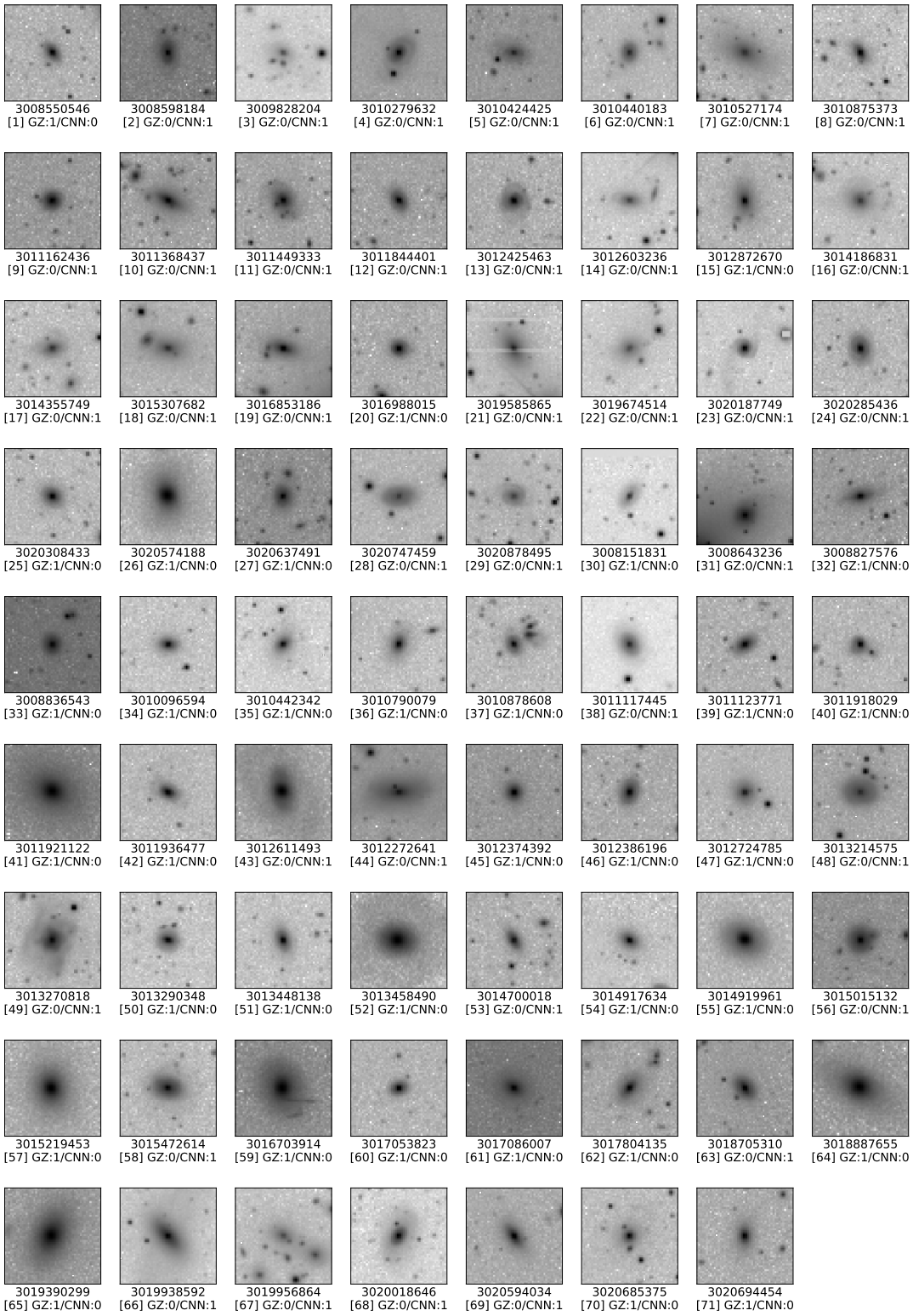

Figure 3.5: The confirmed list of the misclassified galaxies in the Galaxy Zoo 1 catalogue. The first row underneath the images is the ID numbers of galaxies, and the second row shows the classification by Galaxy Zoo (GZ) and our CNN (CNN).

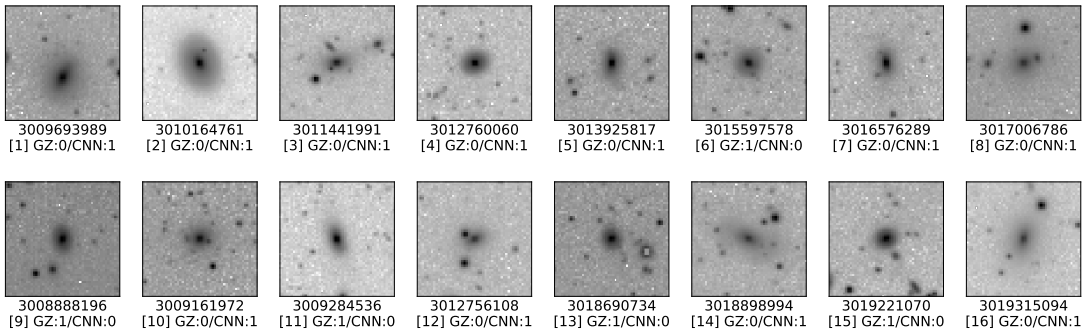

Figure 3.6: The suspected list of the misclassified galaxies in the Galaxy Zoo 1 catalogue. The first row underneath the images is the ID numbers of galaxies, and the second row shows the classification by Galaxy Zoo (GZ) and our CNN (CNN).

	accuracy	R_{01}	$N_{\text{classifiable}}$	N_{uncetain}
No suspected galaxies	0.991±0.003	0.990	976	16
with suspected galaxies	0.989±0.001	0.990	981	19
label changed	0.987±0.003	0.986	981	19

Table 3.5: The testing result after using the purified training set. The meaning of each column are same as Table 3.1. There are 8 suspected misclassified galaxies out of 1,000 testing galaxies. The first row is the testing result excluding suspected galaxies. The second row shows the result with the suspected galaxies which retain their initial labels from the Galaxy Zoo catalogue. The third row is the result with the suspected galaxies but their initial labels changed – for instance, the label changes to Elliptical if the initial label was Spiral.

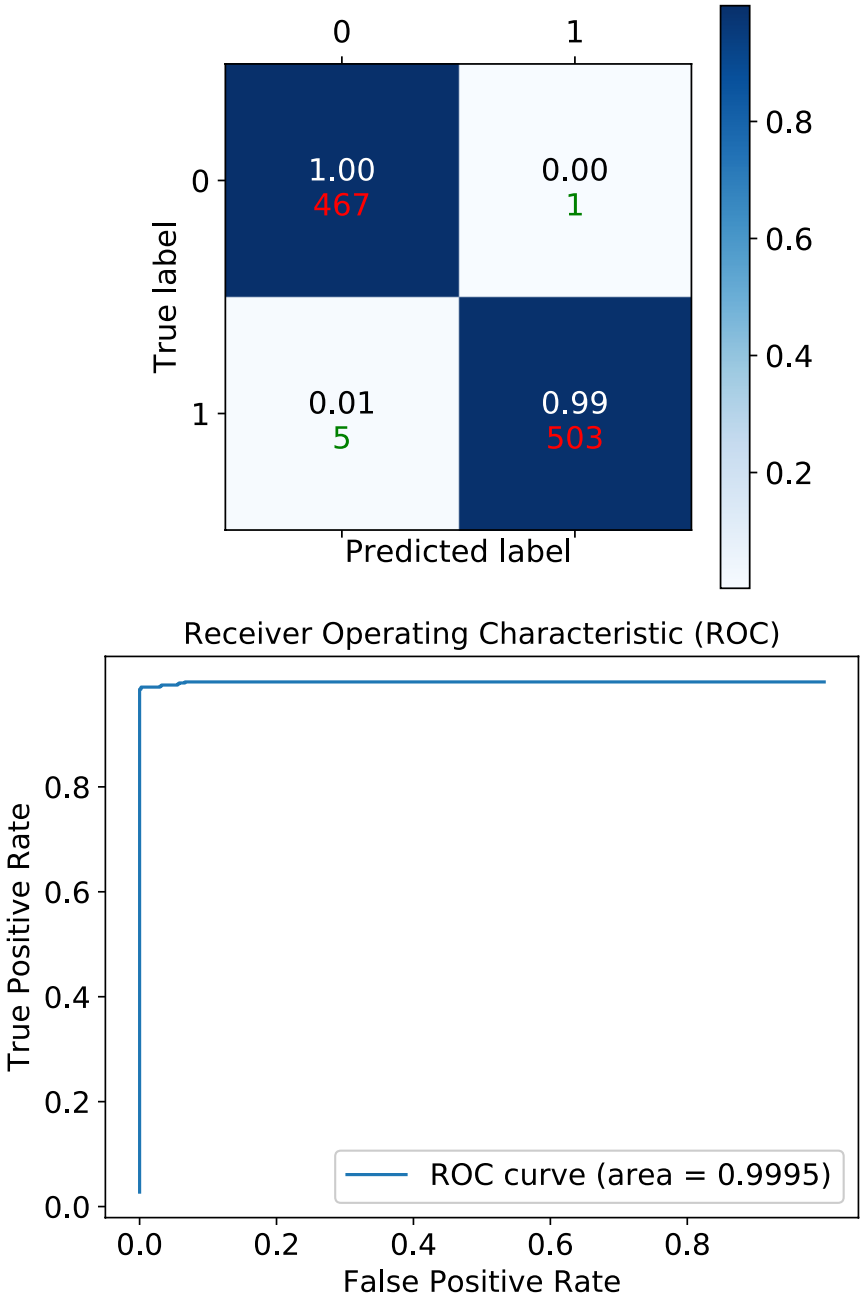

Figure 3.7: The best testing result which we changed the label of the confirmed misclassified galaxies and excluded the suspected misclassified galaxies in both training and testing set. *Top*: Confusion matrix. The '0' means Ellipticals and '1' represents Spirals. The colour bar shows the fraction of each true label (Galaxy Zoo), and the number shows the corresponding number of the fraction. *Bottom*: The ROC curve of this testing result.

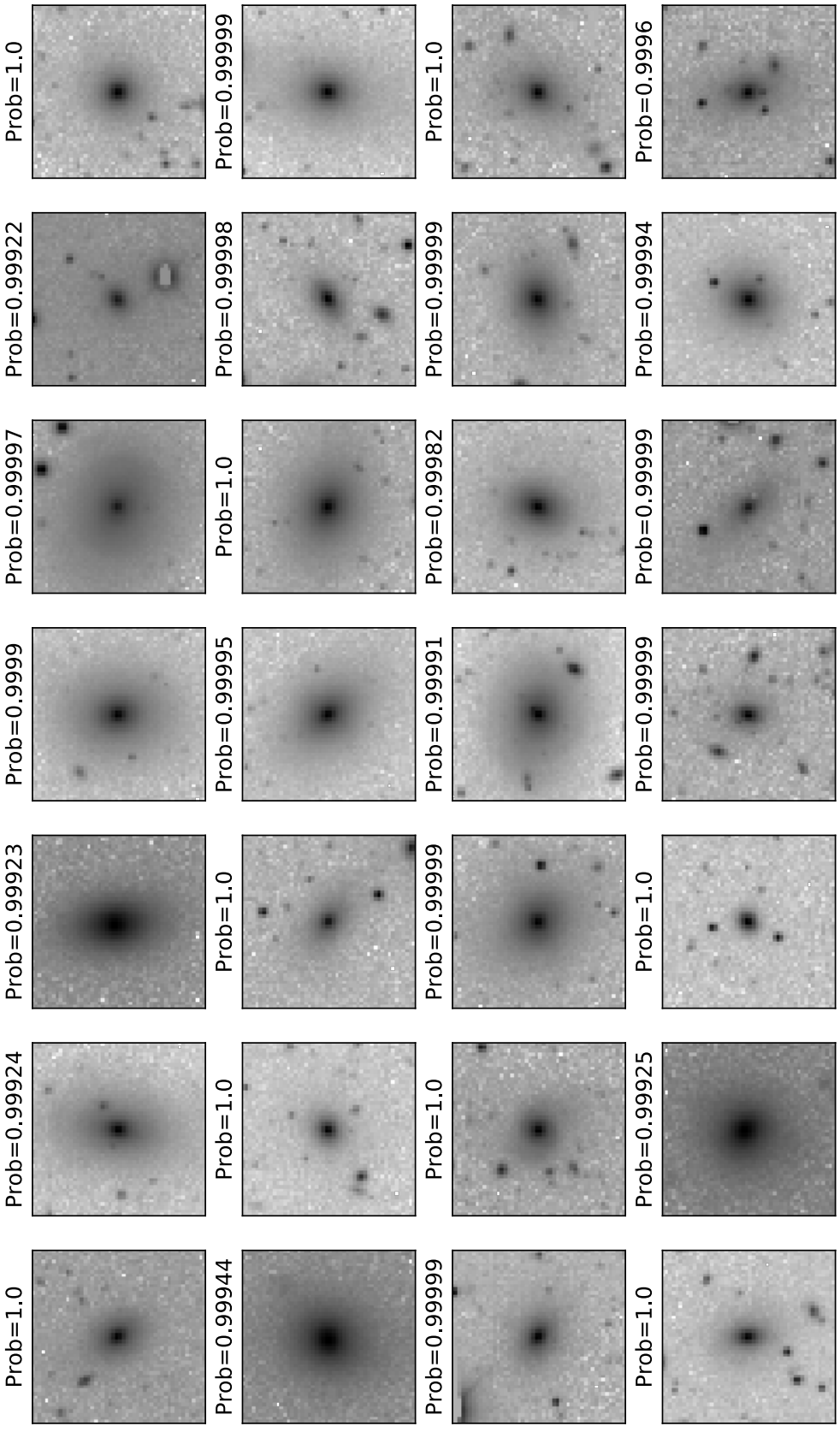

Figure 3.8: Successful examples of classified Ellipticals. The 'prob' on the top of the images show the predicted probability of being Ellipticals.

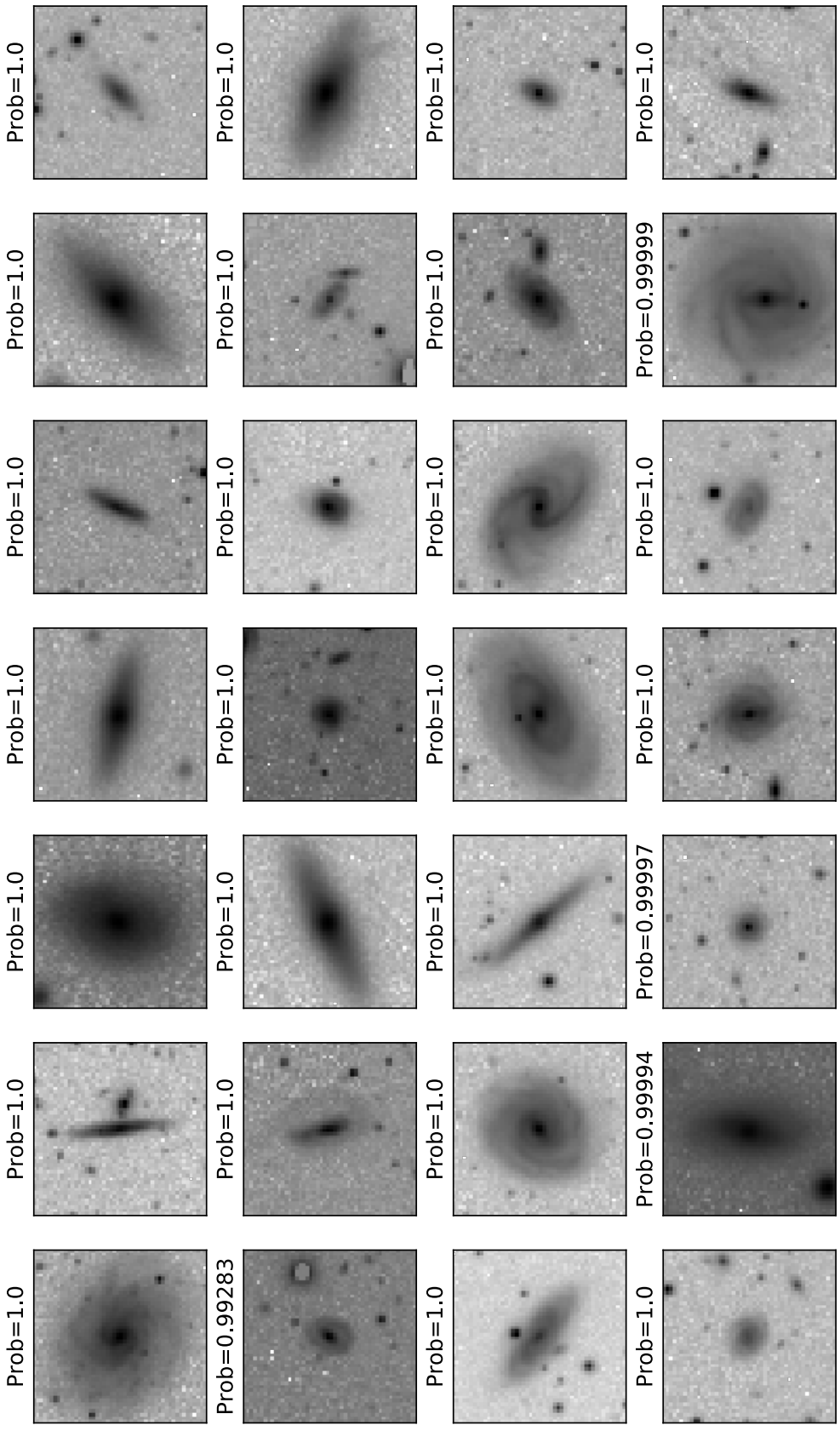

Figure 3.9: Successful examples of the classified Spirals. The 'prob' on the top of the images show the predicted probability of being Spirals.

3.4 Conclusion

In this chapter, we further analyse the convolutional neural networks (CNN) algorithm used in Chapter 2 and successfully improve the performance of our CNN from an accuracy of ∼0.95 to ∼0.99 by investigating the impact of the number of training data and the 'classification failures' by our CNN.

When using a classification criterion for the probability of the predicted type, $p > 0.8$, we firstly increase the accuracy to ∼0.97 and we are able to separate the classification into three types - Ellipticals, Spirals, and Uncertain. In the final test of this part, when we apply the available maximum number of training data to train our CNN, and classified our testing galaxies by the criterion $p > 0.8$, we reach a very high accuracy of ∼0.987 in the automated morphological classification of Ellipticals and Spirals.

Furthermore, we investigate the probable reasons for the failures in a small number of our classifications. We separate the failure into two situations - galaxies with high probabilities but still misclassified according to Galaxy Zoo, and galaxies with lower probabilities of being either elliptical or spiral. Most of galaxies in these two situations are repeated between the five reruns we do; therefore, these galaxies have some features in common which cause the difficulties within our CNN algorithm.

We conclude that these 'failures' are not true failures of the CNN. First of all, there is not a class for lenticular galaxy classification in the Galaxy Zoo catalogue, therefore, the confusion of lenticular galaxies with various labels cause difficulties to our CNN, resulting in low probability classifications for both ellipticals and spirals. Secondly, the better resolution ($0.''263$ per pixel) and deeper depth (i =22.51) of the data from the Dark Energy Survey (DES) compared to the data from the Sloan Digital Sky Survey (SDSS) reveals a more detailed structure of our sample of galaxies. Ultimately, this reveals incorrect labels from the Galaxy Zoo 1 (GZ1) catalogue, due to the lower resolution and shallower depth of that data. As a result we find a few misclassifications by the Galaxy Zoo 1 project, identified through our machine learning. We find that about 2.5% of the Ellipticals and Spirals are mislabelled out of $\sim 2,800$ galaxies from Galaxy Zoo. After correcting the labels of these confirmed misclassified galaxies by Galaxy Zoo, we reach an average accuracy of over 0.99 (0.994 in the best result within five reruns, Fig. 3.7) on the classification of Ellipticals and Spirals by our CNN.

In summary, the purpose of the studies in Chapter 2 and Chapter 3 is to pick the most successful machine learning method through pixel input for future usage in DES. With this method, we can quickly classify millions of galaxies in the DES data using a pre-trained model. The most optimal method found amongst the 10 methods used in Chapter 2 is convolutional neural network (CNN; Section 2.3.7). With this result, in Chapter 4, we apply our CNN models trained by corrected GZ1 labels of galaxies on DES data to build the largest morphological catalogue ever with machine learning classifications. There is not a catalogue of morphological classification of galaxies for DES yet. Therefore, this catalogue

as a reference will be useful for a comparison or further investigation with other studies.

On the other hand, we look forward to developing unsupervised machine learning techniques (UML) for galaxy classification using images (Chapter 5 and Chapter 6). Supervised machine learning needs a certain amount of training data and may be biased by the way labels are assigned and the composition of the training sets (Rosenfeld et al., 2018). On the contrary, UML has no need for (much) pre-labelled data. This advantage saves time and effort for data preparation as well as reduces the potential biases from humans and simulations. Therefore, it will be interesting to explore the evolution of galaxies and galaxy morphologies with UML (Chapter 6).

Chapter 4

The Largest Catalogue of Galaxy Morphological Classification for the Dark Energy Survey Year Three Data

*This chapter is based on unpublished material by **Ting-Yun Cheng**, under the supervision of Christopher J. Conselice, and Alfonso Aragón-Salamanca.*

Abstract

We present in this chapter one of the largest galaxy morphological classification catalogue to date, including over 20 million galaxies, using the Dark Energy Survey Year 3 data based on Convolutional Neural Networks (CNN). A binary classification, ellipticals and Spirals, is provided with an analysis of the confidence level of our predictions. Monochromatic i-band images with linear, logarithmic, and gradient scales, matched with debiased visual classifications from the Galaxy Zoo 1 (GZ1) catalogue, are used to train our CNN models. As stated in Cheng et al. (2020a, Chapter 3), a correction is applied to the GZ1 labels due to the better imaging quality of the DES data, which reveals more detailed galaxy structures. Training with the corrected debiased GZ1 labels makes our CNN classifier self-debiased and provides a more reliable morphological labels to the galaxies that humans have difficulties classifying correctly. For example, the CNN classifier correctly categorises disky galaxies with rounder and blurred features while humans often incorrectly classify them as Ellipticals. The CNN classifications show an accuracy of over 99% when comparing with the GZ1 classifications. As a part of the validation, we carry out one of the largest examination of non-parametric methods including $\sim$100,000 classifications with morphological measurements from Tarsitano et al. (2018) using the most confident CNN predictions in our study. We then reassure the robustness of the Gini coefficient in discriminating Ellipticals and Spirals in this study. Given the largest number of galaxy morphology classifications to date, this catalogue will provide an invaluable resource to DES and the galaxy evolution community.

4.1 Introduction

Galaxy morphology is linked to the stellar populations of galaxies, providing essential clues to their formation history and evolution. Hubble's system (Hubble, 1926) initially had two broad galaxy morphological type: early-type galaxies (ETGs) and late type galaxies (LTGs), based on their appearance in optical light (Section 1.3). These two broad categories connect galaxy morphology with a variety of stellar and structural properties. For instance, ETGs are dominated by older stellar populations and have no spiral structure, while LTGs usually contain a younger stellar population and often have spiral arms. These differences in stellar properties indicate that galaxies with different morphologies follow different formation and evolution paths. Therefore, the availability of galaxy morphologies for very large samples is of great importance when studying the formation and evolution of galaxies.

Visual assessment is the main method of galaxy morphological classification (e.g. de Vaucouleurs, 1959, 1964; Sandage, 1961; Fukugita et al., 2007; Nair and Abraham, 2010; Baillard et al., 2011). However, individual visual classification can be extremely time-consuming. Since around 2000 there has been a significant growth in the size of imaging data sets and increasingly complex ones from e.g., the Hubble Space Telescope (also see Section 1.1). Due to this and the development of computational capacity, non-parametric methods were developed such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), the Gini coefficient, and the M20 parameter (Abraham et al., 2003; Conselice, 2003; Lotz et al., 2004; Law et al., 2007). There are good indications that these parameters, which make no assumptions about the galaxy, are largely free from subjective biases. However, even these computational methods become challenging to apply when the astronomical data become too large and we have to use Big Data techniques and machine learning (Section 1.1).

Before this work, one of the largest galaxy morphological classification catalogue was built using the power of the citizen science - the Galaxy Zoo projects (Lintott et al., 2008, 2011; Willett et al., 2013, see Section 1.1) and contains up to ∼900,000 galaxies. In this study, we apply the convolutional neural networks (CNN) investigated in Chapter 3 to predict binary galaxy morphological classification for the DES Year three GOLD data (hereafter, DES Y3 data). This project allows us to build one of the largest catalogue of galaxy morphological classification to date which includes ∼20 million resolved galaxies.

The arrangement for this chapter is as follows. The data sets are described in Section 4.2, and we introduce the CNN used in this work in Section 4.3. Other catalogues used for validating our CNN predictions are introduced in Section 4.4. The content of our classification catalogue is presented in Section 4.5, while the validation of the predictions are shown in Section 4.6. Finally, we summarise this study in Section 4.7.

4.2 Data Sets

The Dark Energy Survey (DES; DES Collaboration, 2005; DES Collaboration et al., 2016), as mentioned in Section 2.2, is a wide-field optical imaging survey covering 5000 square degrees (∼1/8 sky) which partially overlaps with the survey area of the Sloan Digital Sky Survey (SDSS), but has a better imaging quality than the SDSS images. The spatial sampling of the DES images is 0.″263 per pixel. These images are taken in natural seeing conditions and reach a depth of 22.51 AB magnitudes in the i-band (Abbott et al., 2018).

To create the galaxy stamps for this chapter, we follow the guideline shown in Section 2.2.1 (see Section 4.2.1) to preprocess both the training set (Section 4.2.2) and the DES Y3 data (Section 4.2.3).

4.2.1 Pre-processing

The data preparation we use closely follows the procedure described in Chapter 2. There are two main parts of the data preparation: (1) stamp creation and (2) image processing. Fig. 4.1 shows the pre-processing procedure used in this study. Using the DES GOLD catalogues, we cut the original coadd images, which have a size of 10000 by 10000 pixels, into many different 'postage stamp' images – creating millions of galaxy stamps with sizes of 50 by 50 pixels (approximately $13'' \times 13''$). When a galaxy size, as given in the DES catalogue, is larger than the size threshold (30 by 30 pixels), a larger 200 by 200 pixel stamp is cut from the images, and then re-sampled to produce a 50 by 50 pixel image by calculating the mean value in 4 by 4 pixel blocks. This is done for a very small fraction of the galaxy sample since over 99% of all DES galaxies are smaller than 25 by 25 pixels. Additionally, when creating stamps for the training set, each image is rotated by different angles to increase the number of training images (see Section 2.2.1.1 and Section 4.2.2).

In the second step, we create two extra images which are both included in training our CNN models. One is an image with gradient features that we obtain by a feature extraction technique called the Histogram of Oriented Gradient (HOG; Dalal and Triggs, 2005, see details in Section 2.2.1.3). The HOG, as a feature extractor, is a well-known technique within pattern recognition studies, e.g. human detection, face recognition, and handwriting recognition (e.g., Dalal and Triggs, 2005; Shu et al., 2011; Kamble and Hegadi, 2015, etc). In astronomy, it has already been used in a few of studies such as spectral lines observation (Soler et al., 2019), gravitational lensing detection (Avestruz et al., 2019a), and galaxy morphological classification such as our previous work (Cheng et al., 2020a). The key feature of HOG is to characterise the local appearance and the shape of objects based on local intensity gradients. We rescale the HOG output images so that their pixel values are between 0 and 1 (hereafter, HOG images), and use them as one of the inputs to train our CNN models.

In addition to the HOG images, the other input we use is the image itself with a logarithmic scale (hereafter, log images). In Section 3.3.3, we tested the impact

of using log images to train our CNN algorithms. It initially showed a clear improvement compared with the results using linear images or HOG images only, but decreased in its impact when the number of the training data increased to the maximal available number in the work. To have as complete a set as possible using different significant features in our images we decided to include the log images with rescaled pixel values betwen 0 and 1 when training the final CNN models for the task of catalogue construction.

4.2.2 Training Data

The training data used throughout are described in Chapter 3, which is the subset of the first year DES GOLD data (DES Y1 data), the DES observation of SDSS stripe 82, selected at magnitude $i < 22.5$ and redshift $z < 0.7$ (Drlica-Wagner et al., 2018) and matched with the visual binary classifications from the Galaxy Zoo 1 project (hereafter, GZ1[1]; Lintott et al., 2008, 2011). Monochromatic i-band images are used to select the optimal method for the task as described in Chapter 2. In this work, the morphological classification catalogue is built based upon monochromatic i-band images only, considering the cost in computational time and memory on the enormous size of the DES Y3 data.

We directly used the visual classification provided in Lintott et al. (2011), giving us 2,862 galaxies in total to train our machine. The magnitude range of the overlap data goes from ∼12.5 to 18 in the i-band, and their redshifts are at $z \leq 0.25$ (peak at $z \sim 0.1$). However, in Chapter 3, we found that the better resolution and deeper depth of DES data reveal more detailed structures such as spiral arms that did not show in the data used in GZ1 from the SDSS. This condition resulted in a few mismatches between our CNN predictions and the GZ1 labels.

Additionally, we note that the morphological flags provided in Lintott et al. (2011), as clarified on the website of the GZ1 data release[1], are constructed using the bias correction flags based upon the Ellipticals/Spirals ratio (hereafter, E/S ratio) after applying a vote fraction threshold of 0.8 when counting galaxies in each type. Using this correction shows a worse bias than the one based on the E/S ratio using the likelihoods directly (the 'debiased votes' provided in Lintott et al. (2011)). In Chapter 3, some galaxies with less accurate morphological flags after the bias correction from the GZ1 also showed a questionable label when comparing with our CNN predictions. Therefore, through repeated tests of our CNN and visual assessment, we corrected the labels for $\sim 2.5\%$ of our sample galaxies, and excluded $\sim 0.56\%$ galaxies that have suspected labels according to our test in Chapter 3. We then train our final CNN models with the corrected GZ1 labels which better correspond to the ones based upon the 'debiased votes' in Lintott et al. (2011). Therefore, when comparing our CNN predictions with the GZ1 classifications we use the 'debiased votes' in the GZ1 catalogue to determine the final morphology types for comparison purposes (See Section 4.4.1 and Section 4.6.1).

[1]https://data.galaxyzoo.org/

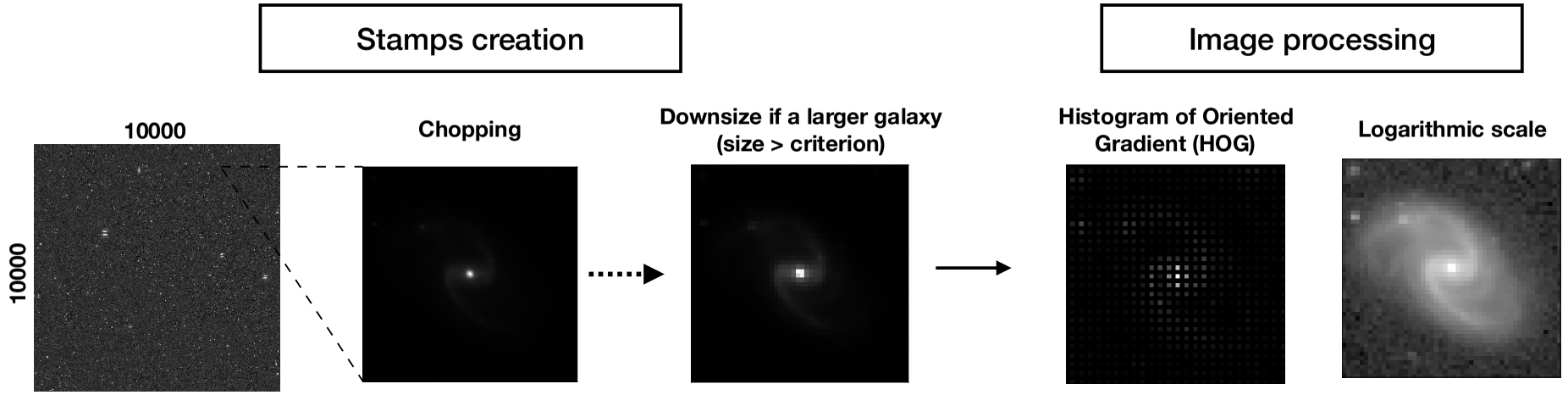

Figure 4.1: Pre-processing procedure pipeline. The pipeline starts from the initial coadd images, then we chop them into different sizes based on the size of galaxies. A downsizing procedure is applied if the size of a galaxy is larger than the threshold of 30 by 30 pixels; otherwise, we simply chop the stamps into the size of 50 by 50 pixels. After which, two individual image processing procedures are carried out to generate the HOG images and the log images (see details in Section 4.2.1)

Selection Flags	
EXTENDED_CLASS_COADD	= 3
EXTENDED_CLASS_WAVG	= 3
FLAG_FOOTPRINT	= 1
FLAG_FOREGROUND	= 0
bitand(FLAGS_GOLD,120)	= 0
bitand(FLAGS_BADREGIONS,1)	= 0

Table 4.1: The flags used to select data in the DES Y3 GOLD catalogue. The first two flags guarantee that the astronomical objects which are the most likely to be a galaxy are selected, and the last four flags indicate the data with a reliable analysis from the SExtractor (Bertin and Arnouts, 1996).

The training set is prepared following the pipeline shown in Fig. 4.1. To prevent from overfitting during the training process, important as considering we have a limited amount of labelled data, an extra process of rotating images is performed to increase the number of the training data. An extra amount of Gaussian noise is also added, which is negligible towards causing any impact to the visual appearance and the structure of galaxies, but able to enhance a detectable change of pixel values (Dieleman et al., 2015; Huertas-Company et al., 2015). Finally, we retain the balance between the number of elliptical (E) and spiral galaxies (S) in the training set; therefore, the rotational operation increases the number of training data to 54,133 galaxy stamps with the ratio of number of types held to $E/S \sim 1$.

4.2.3 DES Year 3 Data

We build the catalogue of galaxy morphological classifications based on the DES Year 3 (Y3) GOLD data that are selected with the flags shown in Table 4.1 and within a magnitude range $16 \leq i \leq 22$. The top two flags guarantee that astronomical objects selected using these flags are most likely to be galaxies, based on the analysis from SExtractor (Bertin and Arnouts, 1996). The bottom four flags are used to select the data with a reliable SExtractor analysis. This selection provides over 50 million galaxies for the task, with the redshift distribution of the selected data peak at $z \sim 0.4$ with over 99.9% of the galaxies at $z \leq 1.2$. The galaxies in this catalogue have a wider range of magnitudes and redshifts than those in the training set - the training set galaxies are, typically, brighter and have lower redshifts. A subsample of about 670,000 galaxies have magnitudes and redshifts in the ranges covered by the training set.

The selected data is separated into six magnitude bins from $i = 16$ to $i = 22$ for further analysis (Section 4.6). The number of galaxies in each magnitude bin increases exponentially when going fainter. A pre-processing procedure described in Section 4.2.1 is also applied to the selected data.

4.3 Convolutional Nerual Networks (CNN)

Convolutional Neural Networks (CNN, Lecun et al., 1998) are a type of neural network which includes convolutional layers used to extract strongly weighted features from input images for a given classification problem (Section 2.3.7). CNNs were first applied on galaxy morphological classification in Dieleman et al. (2015). Since then, this technique is widely used in a variety of astronomical studies (see in Section 1.2).

The architecture of the CNN used throughout this chapter is shown in Fig. 2.5. This design is inspired by the best performing architecture used in Dieleman et al. (2015), but with fewer convolutional layers and parameters. The dimension of the inputs is $50 \times 50 \times 3$, with the depth including the linear images, HOG images, and log images. Three convolutional layers with filter sizes of 3, 3, 2, respectively, are used in this study, and each of them is followed by a max-pooling layer with a size of 2. The max-pooling layer is also referred to as a 'downsampling' layer, which is used to reduce the spatial size and the numbers of parameters involved in the architecture. After the third convolutional layer, two dense layers with 1,024 hidden units for each layer follow. In addition, dropouts (= 0.5) are applied to reject irrelevant parameters and prevent overfitting in training the CNN. A dropout follows the third convolutional layer (max-pooling layer), and the other one comes after the two dense layers.

The activation function used in the convolutional layers and the dense layers is the Rectified Linear Unit (`ReLu`; Nair and Hinton, 2010) such that $f(z) = 0$ if $z < 0$ while $f(z) = z$ if $z \geq 0$. Finally, the `softmax` funciton (Bishop, 2006), $f(z) = \exp(z)/\sum \exp(z^j)$ is applied to the output layer and provides the probability distribution of each type. For the CNN training, we apply *Adam* Optimiser, *Nesterov momentum*, and set *momentum* = 0.9 according to Dieleman et al. (2015). The learning rate is set to 0.001, and the maximum number of iterations is 500, with an early-stopping mechanism that triggers when the validation set hits the local minimal loss.

A CNN has the technical advantage of not requiring the pre-processing procedure commonly used in artificial neural networks. However, in Chapter 2 and Chapter 3, we have proven that combining pre-processed images such as HOG images and log images qualitatively improves the performance of our CNN and reaches a final accuracy of over 0.99. In this study, we independently train the CNN five times with the same training set. After this, the final prediction is obtained by averaging the predicted probabilities of these five independent CNN models for each type, 'Ellipticals' and 'Spirals'.

4.4 Catalogues for Cross-validation

Once we have the morphological predictions from the convolutional neural network (CNN) for millions of galaxies, it is of great importance to validate the reliability of these classifications. In this study, we compare our CNN predic-

tions with three different resources: (1) the Galaxy Zoo 1 (GZ1) catalogue using the galaxies that were not present in the training set (Section 4.4.1); (2) visual classifications carried out by TC, CC, and AAS[2] (Section 4.4.2); and (3) non-parametric methods using the structural measurements from Tarsitano et al. (2018) (Section 4.4.3).

4.4.1 The Galaxy Zoo 1 catalogue (GZ1)

The Galaxy Zoo projects are amongst the most successful attempts using citizen science to obtain large numbers of galaxy morphological classifications (Section 1.1). A set of questions are asked to the volunteers for each galaxy image. Based on the answers from the volunteers, the GZ1 statistically provides the morphological classification of $\sim 900,000$ galaxies. Of these, $\sim 670,000$ galaxies with spectroscopic redshifts have been bias corrected (Bamford et al., 2009).

In this study, we have three main pieces of classification information from GZ1: *raw votes*, *debiased votes*, and *morphological flags*. The *raw votes* are the likelihood calculated directly from the volunteers' votes for each image. The *debiased votes* and *morphological flags* are derived after applying bias corrections based upon different assumed E/S ratios (see Lintott et al., 2011).

In GZ1, a correction factor is necessary to account for a classification bias that depends on the apparent brightness and size of each galaxy. For example, when viewing a spiral galaxy at higher redshift, its decreasing apparent brightness and size makes it more difficult to appreciate morphological details such as spiral arms, resulting in an increased likelihood of it being classified as an elliptical galaxy. The corrections needed to account for this bias are calculated by assuming that the morphological mix does not evolve significantly in the narrow redshift range covered by GZ1 (Bamford et al., 2009). This assumption has been shown to be a reasonable one (Conselice et al., 2005).

In order to perform this correction, GZ1 use two different values for the E/S ratio, one to obtain the *morphological flags* and a different one to estimate the *debiased votes*. The *morphological flags* provided by Lintott et al. (2011) are determined using the E/S values that only take into account the classifications with at least a 0.8 morphological vote fraction. On the other hand, the *debiased votes* provided by GZ1 are based on E/S ratios that use the raw likelihood.

Our CNN model is trained with the ***corrected*** *morphological flags* based on DES imaging data (details of this are given in Section 4.2.2). This correction was performed in Chapter 3 and, when applied, it decreases the number of inaccurate classifications. These classifications are claimed to be incorrect when compared with the GZ1 labels due to both the better quality of the DES imaging data than the SDSS images and the use of the bias correction in the 'morphological flags' from the GZ1 catalogue.

[2]TC: Ting-Yun Cheng; CC: Christopher Conselice; AAS; Alfonso Aragón-Salamanca

labels	primary votes	combined votes
0	Ellipticals	Ellipticals
1	Early Spirals	Sprials
2	Late Spirals	
3	Edge-on Spirals	
4	Irregulars	Irregulars
5	Unknown	Unknown

Table 4.2: The classification system we applied in the visual classification carried out by TC, CC and AAS[2]. Galaxies are classified into 6 categories (*primary votes*) which are then merged into 4 categories, i.e. Ellipticals, Spirals, Irregulars, and Unknown (*combined votes*; see text).

For the latter situation, the ***corrected*** *morphological flags* carried out in Chapter 3 better corresponds to the results using the *debiased votes*. Therefore, we use the GZ1 classifications based upon the *debiased votes* as the comparison match with our CNN predictions.

In summary, in the training phase we use the *morphological flags* from GZ1 as corrected in Chapter 3, and test the predictions of our CNN system using the *debiased votes* from GZ1 (Section 4.6.1).

4.4.2 Visual classification of randomly selected subsamples

Galaxy morphologies are needed for validation of the CNN predictions at all magnitudes, but they are not available for faint galaxies ($i > 18$). Therefore, visual classification (VIS) was carried out by the author (TC) and her supervisors (AAS and CC) for a reasonably large number of galaxies. We randomly selected 500 galaxies per magnitude bin from the DES Y3 dataset for galaxies with $16 < i < 22$. For the brightest bins ($16 < i < 18$), only galaxies in GZ1 were included. In doing so, we covered the whole magnitude range of the DES sample with a significant overlap with GZ1 for cross-validation.

The classification system we use is displayed in Table 4.2. We classify galaxies into six categories: Ellipticals (0), Early Spirals (1), Late Spirals (2), Edge-on Spirals (3), Irregulars (4), and Unknown (5). To compare with our CNN predictions, which is a binary classification system, we merge three subcategories of spiral galaxies into one - Spirals (1), and others retain the original label. The label with the most *combined votes* (Table 4.2) from our visual classifiers is set as the final visual type of a galaxy. The combined vote is the morphological type which is picked by at least 2 out of 3 of the classifiers. Those galaxies without a dominant label are categorised into the class of 'Unknown'; the relative fraction of these 'unknown' types increases with magnitude. The distribution of each visual type in each magnitude bin is shown in Fig. 4.2.

In order to validate the visual classifications (hereafter called VIS), we compared the classifications of brighter galaxies ($i < 18$) with the GZ1 classifications

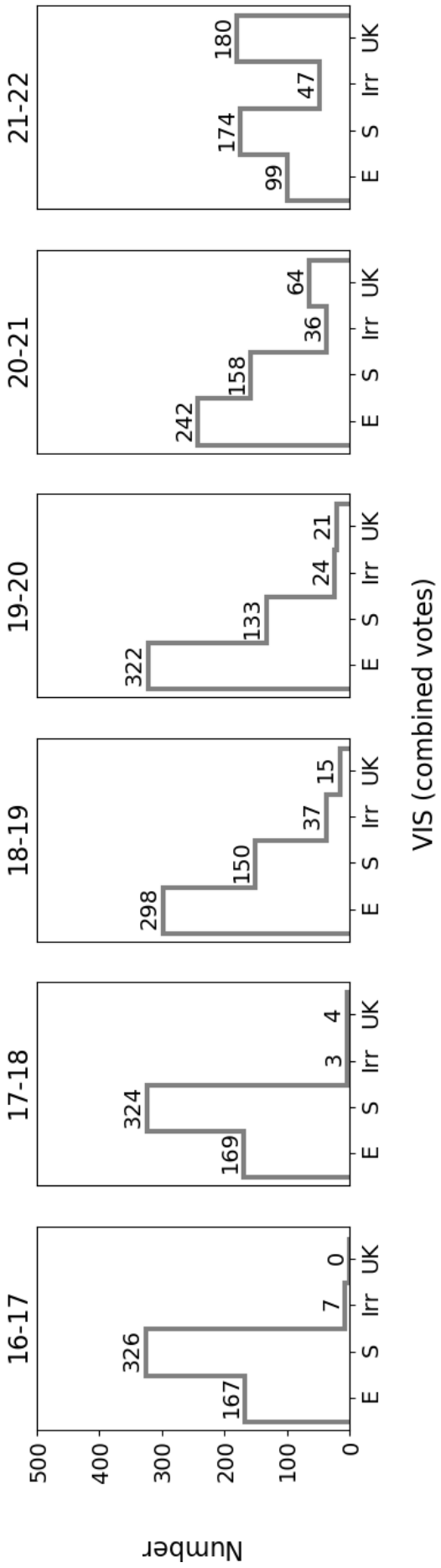

Figure 4.2: The frequency distribution of each visual classification in each magnitude bin. On the x-axis, 'E', 'S', 'Irr', and 'UK' is short for Ellipticals, Spirals, Irregulars, and Unknown, respectively. The number above each bar represents the number of galaxies visually classified in that category. The range of the magnitude in i-band is shown at the top of each column.

based upon the *debiased votes* and *raw votes* (Fig. 4.3). The *raw votes* directly reflect the votes from the volunteers of the GZ1. The *debiased votes*, as described in Section 4.4.1, are bias corrected using the E/S ratio measured directly from the raw likelihood. We apply a threshold of 0.8 to both votes to decide the morphology type with a higher confidence.

In Fig. 4.3, our VIS classifications show apparently better agreement with the raw votes from the GZ1 volunteers when comparing with the GZ1 debiased votes. The majority of the mismatched cases when comparing with the labels based on the debiased votes occur when a galaxy is classified as Elliptical by our visual classifications. This indicates that our judgement for galaxy morphology is also biased by the size, magnitude, and redshift of the galaxies. This gets worse when a galaxy is fainter which is shown in Fig. 4.2. It is clear that significantly more galaxies are visually classified as ellipticals.

Although our visual classification suffers from the same type of biases as GZ1, unfortunately we cannot perform a bias correction similar to the one they carried out. There are several reasons for this. First, the broader redshift range of our sample makes the assumption of unevolving morphological mix unreliable. Second, number of galaxies we have been able to classify is too small to provide reliable correction statistics. And third, the lack of spectroscopic redshifts would render any redshift-dependent correction highly uncertain. Therefore, additional factors such as Sérsic index (Sérsic, 1963, 1968) and colour will be considered to validate the CNN predictions (Section 4.6.2).

4.4.3 DES Y1 catalogue of morphological measurements

To obtain a reliable analysis of the quality of our CNN labels, in addition to using visual classifications, parametric factors such as the Sérsic index and non-parametric coefficients such as *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), Gini coefficient, and M20 are used in this study. Tarsitano et al. (2018) included 45 million objects selected from the first year DES data, and provided the largest structural catalogue to date for galaxies. The selected samples from this catalog cover the magnitude range of $i \leq 23$. According to the suggestions from the paper, we apply an initial cut as follow,

- MAG_AUTO_I ≤ 21.5
- SN_I > 30
- SG > 0.005,

where MAG_AUTO_I represents the cut in i-band apparent magnitude and SN_I is the signal-to-noise ratio in the i-band. The SG is used for optimising the separation between stars and galaxies while maintaining the completeness. The cut ($SG > 0.005$) recommended in Tarsitano et al. (2018) is the optimal compromise between the completeness and purity of the galaxy sample. These selections provides 12 million galaxies with 90% completeness in Sérsic measurements and 99% completeness in non-parametric measurements in the i-band.

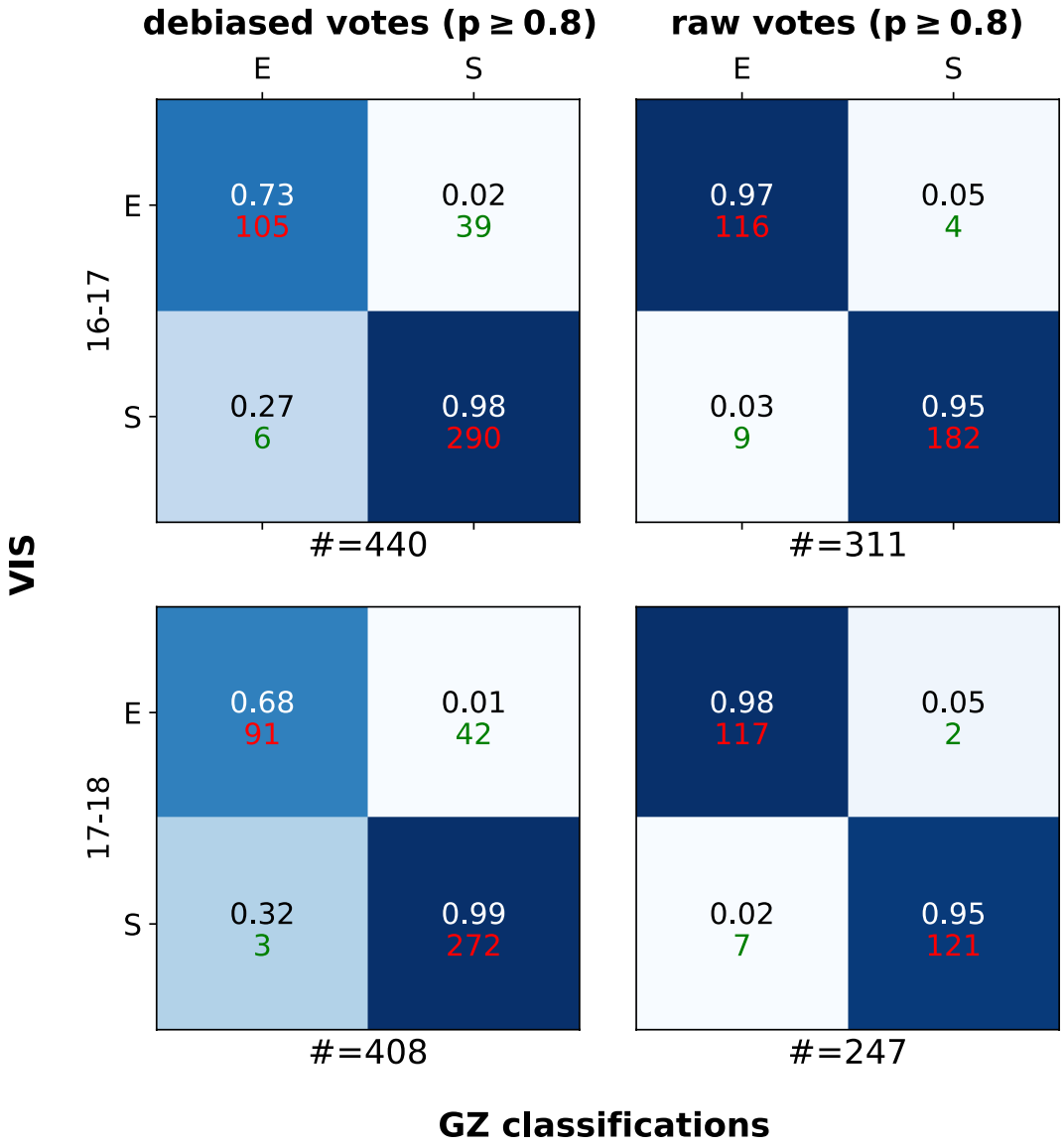

Figure 4.3: The confusion matrices between our visual classifications (VIS) and the GZ1 classifications based on the *debiased votes* (first column) and *raw votes* (second column) (Lintott et al., 2011). A threshold of 0.8 is applied to both votes here to select high confidence classifications. Rows are separated by different magnitude bins: $16 \leq i < 17$ (first row) and $17 \leq i < 18$ (second row).

The parameters provided from the single Sérsic fits (e.g. Sérsic index, ellipticity, etc) are measured with GALFIT (Peng et al., 2010). We then apply a further cut suggested in Tarsitano et al. (2018) to select the galaxies that are successfully validated and calibrated. The calibration is made based upon four parameters: size, magnitude, Sérsic index, and ellipticity using simulated galaxies generated with these parameters (Tarsitano et al., 2018):

- FIT_STATUS_I = 1

On the other hand, the non-parametric parameters (CAS parameters, Gini, and M20) are measured using the Zurich Estimator of Sturctural Types (ZEST+) (Scarlata et al., 2007a,b). The calibration is applied with the same procedure as the parameter fit but uses concentration instead of Sérsic index for non-parameteric parameters, and the validation is discussed on the Gini-M20 plane as a function of other morphological measurement such as concentration (C), asymmetry (A), and clumpiness (S) (Tarsitano et al., 2018). One criterion is applied in non-parametric coefficients to select the objects with successfully validated and calibrated measurements.

- FIT_STATUS_NP_I = 1

4.5 Galaxy Morphological Classification Catalogue

In this chapter, with the convolutional neural network (CNN) trained with the subset of the DES Y1 data with the GZ1 labels corrected in Chapter 3 (Section 4.2.2), we provide one of the largest catalogue to date with galaxy morphological classifications for over 20 million galaxies from the DES Y3 data (Section 4.2.3). The items provided in our catalogue of morphological types are listed in Table 4.3. The average predicted probabilities from the five individual CNN models (Section 4.3) are used as the final probabilities of being Ellipticals (pE) and Spirals (pS). With the predicted probabilities of both types, we provide the classification label based on a threshold of 0.8 (*MORPH_FLAG*) for the user's convenience. The analysis shown in Section 4.6 uses this most probable morphological label.

Machine learning is sensitive to image qualities such as the signal-to-noise ratios and resolution, but have a certain level of tolerance for variations within these effects (Dodge and Karam, 2016). The apparent magnitude of a galaxy, which is influenced by the redshift, affects the signal-to-noise ratio of the galaxy observed in the image, which can affect how easily structure can be seen. Additionally, due to the effects of distance, a galaxy at a higher redshift shows less detailed structure, namely the resolution of the galaxy images decreases. Therefore, we statistically investigate the confidence of our CNN predictions at fainter magnitudes and higher redshifts by comparing the quality of our morphologies with our visual assessments, structural measurements such as the Sérsic profile, and galaxy properties such as colour. The detailed discussion of this is in Section 4.6.2 and Section 4.6.3. Based on the analysis in Section 4.6.3, we provide a *confidence_flag*

Col.	Keyword	Description
1	DES_Y3A1_ID	DES Y3 ID
2	RA	right ascension
3	DEC	declination
4	pE	probability of being Ellipticals
5	pS	probability of being Spirals
6	MORPH_FLAG	CNN predictions with a threshold of 0.8
7	MAG_I	i-band magnitude. *(MAG_AUTO_I)*
8	MAGERR_I	i-band magnitude error. *(MAGERR_AUTO_I)*
9	ZMEAN	photometric redshift. *(DNF_ZMEAN_MOF)*
10	ZSIGMA	redshift uncertainty. *(DNF_ZSIGMA_MOF)*
11	confidence_flag	confidence level of predictions

Table 4.3: Content of the catalogue published with this work. Columns 7 to 10 are quantities which are taken directly from the DES Y3 GOLD catalogue, and the corresponding column names are highlighted and placed within brackets in the description.

(column 11 in Table 4.3) which can be used to select CNN classifications with different confidence levels depending on the science goal. The confidence labels used for this flag are shown in Table 4.4.

We give a short description of these confidence flags here. We categorise our CNN predictions into six confidence levels (Table 4.4). The 'superior confidence' level are the classifications using the data with the same magnitude and redshift ranges as the training set, and the detailed analysis is shown in Section 4.6.1. Other confidence levels are defined based upon the distribution of colour, $g-i$, and Sérsic index for galaxies in a specific region of magnitude and redshift. A detailed discussion is given in Section 4.6.3. We recognise the distribution of galaxies with the 'superior confidence' as the reference. Samples with the 'high confidence' show a clear separation in colour and Sérsic index as well as a reasonable peak in both distributions. The 'confidence' label indicates a slightly worse distribution, e.g., a broader distribution in Sérsic index, etc, compared with the reference. The 'less confidence' label is assigned when at least two significant differences in distributions compared with the reference are recognised. The '*' label shown in Table 4.4 is there to represent that only the classifications of Spirals are good within the assigned confidence level, and the rest are labelled as 'no confidence'. Finally, the 'no confidence' is for galaxies with a messy distribution, e.g., bimodal distribution, within a specific range of colour and redshift.

Overall, over 20 million CNN classifications with an assigned confidence level are included in our final catalogue; of which, ∼670,000 galaxies have a 'superior confidence', ∼5 millions of galaxies are assigned as a 'high confidence' classifica-

labels	representation
4	superior confidence
3	high confidence
2	confidence
1	less confidence
1*	less confidence (for Spirals only)
0	no confidence

Table 4.4: Content of the *confidence_flag* (column 11) shown in Table 4.3. The 'superior confidence' flag is for classifications within the same magnitude and redshift ranges as the training set. The details of other levels are described in Section 4.6.3.

tion, and ∼7 million galaxies have a 'confidence' label. Finally, in columns 7-10 in Table 4.3 we provide magnitude and redshift information directly from the DES Y3 GOLD catalogue to allow customised magnitude/redshift cut when applying our predictions.

4.6 Validation & Discussion

In this section, we carry out the cross-validation of our CNN predictions using multiple sources. Included among this, we also discuss the confidence levels assigned to the predictions and the uses of this catalogue with these confidence levels. That is, we explain how to use our catalog for determining galaxy morphologies.

4.6.1 Galaxy Zoo 1 catalogue (GZ1)

To validate our CNN predictions, first we compared the CNN classifications with the GZ1 labels based upon the *debiased votes* (Section 4.4.1). The distribution of the DES Y3 data for this test is in the same magnitude and redshift range as the training set (Section 4.2.2) as shown in Fig. 4.4. Note that there are significantly fewer faint galaxies at $i < 17.3$ in our sample with overlapping GZ1 classifications. Therefore, a cut of $i = 17.3$ is applied when carrying out this analysis in this subsection. We then discuss the performance of the CNN predictions below and above this magnitude limit in later sections.

First, in Fig. 4.5, we show the change in accuracy when applying different likelihood thresholds to the GZ1 debiased votes. The first two columns are separated by the magnitude cut $i = 17.3$, and the third column contains all overlapping data between GZ1 and DES Y3 data used in this work. The accuracy of the training set is represented by the black lines. One applies a probability threshold of $p = 0.5$ to our CNN predictions (dashed line), the other applies a threshold of $p = 0.8$ (solid line). The comparison of the results using different likelihood threshold at various GZ1 debiased votes are shown by the blue lines. The line styles reflect the same meaning as the black lines. Meanwhile, the second y-axis is used for the shading bars which show the number of galaxies under the likelihood threshold.

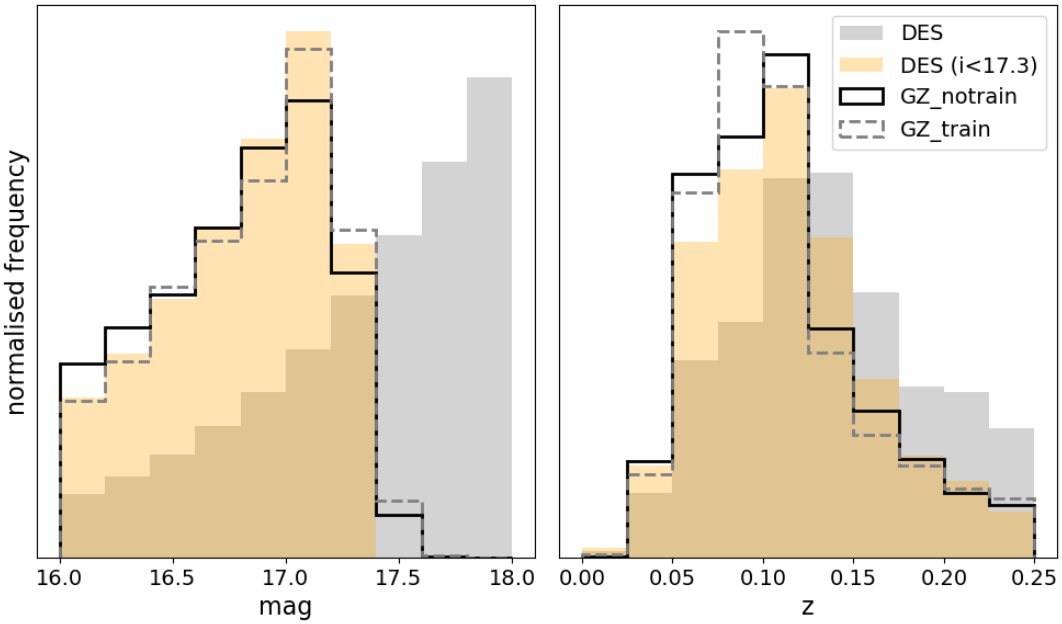

Figure 4.4: The magnitude and redshift distribution of the DES Y3 data with the same coverage as the training set (Section 4.2.2). The gray and yellow shading represents the DES Y3 data without and with a cut at $i = 17.3$, respectively. The solid lines show the overlap region with the GZ1 catalogue, excluding the training set, while the dashed lines show only the training set.

The GZ1 label, after this correction made in Chapter 3, is used for the training set. This mostly corresponds to using the *debiased votes* with a threshold of 0.8. Using this, we note that the accuracy of our CNN predictions compared with the GZ1 classifications based upon a debiased likelihood threshold of 0.8 shows a good consistency with the accuracy of the training set (the first column in Fig. 4.5). In the second column ($17.3 \leq i < 18$), the CNN show a slightly better performance than the brighter range ($16 \leq i < 17.3$) and training set. However, the scatter for the CNN predictions are larger because there are significantly fewer samples in the second plot. When taking the scatter into account, the performance of our CNN predictions in this magnitude range also shows a good consistency with the training set. Therefore, based upon Fig. 4.5, we interpret that there is a 'superior confidence' level to the CNN predictions within the brighter magnitude range $16 \leq i < 18$ and redshift range $z < 0.25$. Additionally, in later analysis, we apply a likelihood threshold of 0.8 to the GZ1 debiased votes to determine the GZ1 classifications for comparison, as well as a probability threshold of 0.8 to our CNN predictions to reject samples with low predicted probabilities from the CNN model.

In the first three column of Fig. 4.6, we show confusion matrices within a certain magnitude range as listed above the graph. The x-axis indicates the CNN predictions, while the y-axis shows the GZ1 classifications. In this study, we work on binary classification, namely, Ellipticals (E) and Spirals (S). The numbers at the bottom of the confusion matrices show the number of galaxies within the

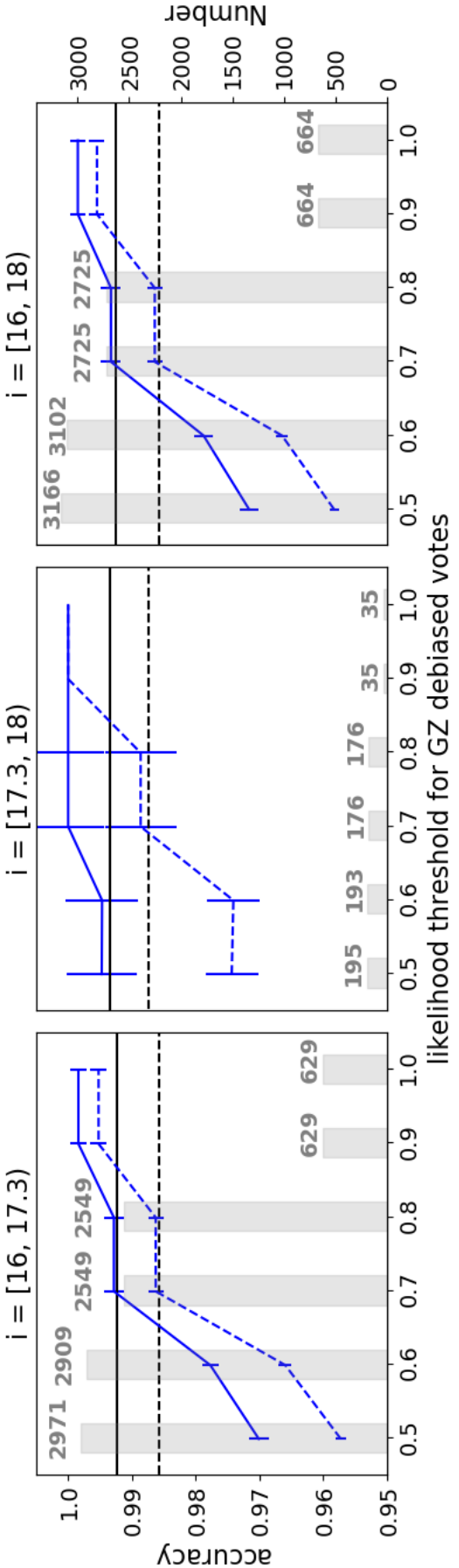

Figure 4.5: The comparison between the accuracy of the CNN predictions when applying different likelihood thresholds to the GZ1 debiased votes. A magnitude cut $i = 17.3$ is applied in the first two columns, and the third column includes total data in the first two columns. The black lines are the accuracy of the training sets when applying a probability threshold, $p = 0.5$ (dashed line) and $p = 0.8$ (solid line) to our CNN predictions. The blue lines represent the change of accuracy when comparing our CNN predictions, based on a probability cut $p = 0.5$ (dashed line) or $p = 0.8$ (solid line), with the GZ1 classifications based on different likelihood thresholds. The second y-axis on the right reflects the height of the shading bars which gives the number of data points left after applying a likelihood threshold.

ranges in each column. For the first three plots, we exclude the training set, and compare the performance with the training set in the last column. In this figure, we notice that the two labels (GZ1 and CNN) match well, and the majority of mismatches occur in the case where the CNN classification is Spiral, but the debiased GZ1 classification disagrees. Fig. 4.7 showcases the galaxies which are classified as Spirals by the CNN but Ellipticals by the GZ1. Some galaxies in this category show disky structures (e.g. [1], [3], [5], [15]) or asymmetric features (e.g. [8] and [9]) in the DES imaging data. In Chapter 3, we proved that the higher quality DES imaging data reveals detailed structures that were not detected in the data from SDSS. This condition is responsible for a significant fraction of the mismatched classifications in Fig. 4.7.

Ideally, we would have liked to examine the Sérsic index distribution of these mismatched galaxies to determine whether these misclassified galaxies have particular structural properties. However, in this case, there are fewer than three overlapping galaxies with mismatched labels in Tarsitano et al. (2018). Therefore the mismatched test sample is far too small for any statistically meaningful analysis. Therefore, we leave this additional cross-validation to future work, when more structural measurements are measured for the DES Y3 data. We note that, although we cannot carry out this additional test, we are confident of the excellent performance of our CNN predictions within the magnitude and redshift range covered by the GZ1 training set. Based on the discussions above and, in particular, the confusion matrices shown in Fig. 4.6, we conclude that in this magnitude and redshift range, which includes $\sim 670,000$ galaxies, our CNN classifier has an accuracy of over 99%.

The last column in Fig. 4.6 shows a Receiver Operating Characteristic curve (ROC curve, Fawcett, 2006; Powers, 2011) which is used to examine the performance of our machine learning technique by comparing the probabilities predicted by the machine with the true labels (see details in Section 2.4.1). Another important indicator on the ROC curve is the 'area under the curve', AUC, which shows a better performance of a machine learning model when having a larger value. From the ROC curve, the CNN predictions within the coverage of the training sets in magnitude ($16 \leq i < 18$) and redshift ($z < 0.25$) show a perfect consistence with the results of the training set. This result doubly confirms our confidence on these predictions. Therefore, the CNN predictions within this range are labelled as 'superior confidence (4)' in the *confidence_flag* (Table 4.4) in the catalogue, Table 4.3.

4.6.2 Visual classification

To allow us to test the quality of the CNN classifications of fainter galaxies, we carried out a visual classification of 500 randomly picked galaxies in each magnitude bin with an interval of 1 magnitude using the DES imaging data. The first five columns in Fig. 4.8 show the confusion matrices in each magnitude range, and the ROC curve is shown on the last column. The performance quality of our CNN method drops with magnitude when comparing with the visual classifications. Through the confusion matrices, we notice that the majority of

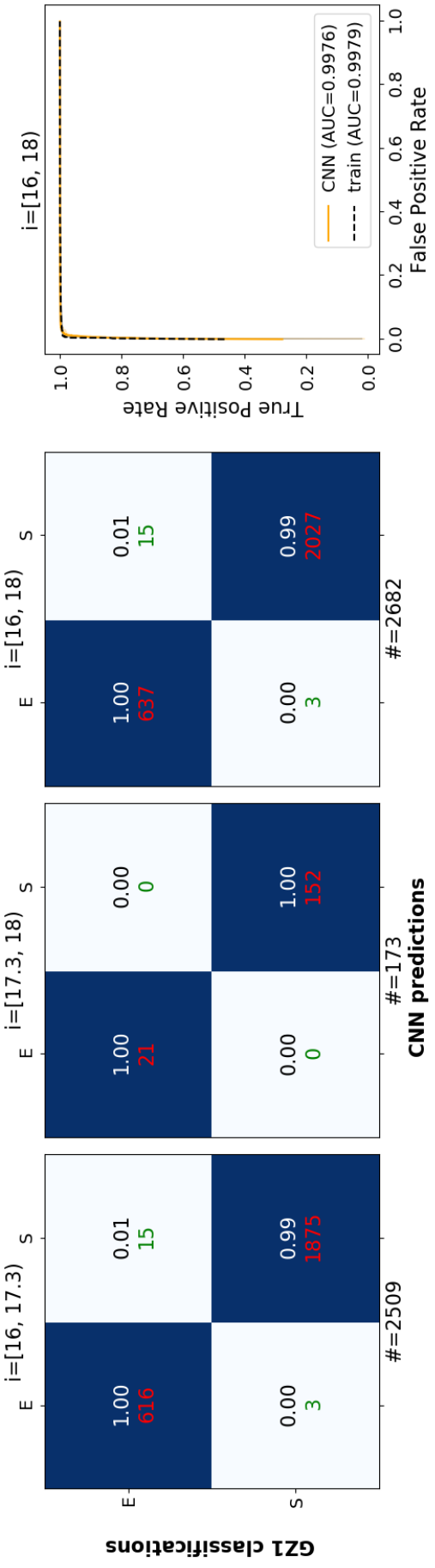

Figure 4.6: The combined graph of the confusion matrices and the ROC curve comparing our CNN predictions with the GZ1 labels defined by the debiased votes with a threshold of 0.8. The first three columns show the confusion matrices within certain magnitude ranges: $16 \leq i < 17.3$, $17.3 \leq i < 18$, and $16 \leq i < 18$, and the number of galaxies within the ranges are shown below each graph. The last column shows the ROC curve where the y-axis represents the true positive rate, and the x-axis is the false positive rate. The orange line shows the curve from our CNN predictions while the black dashed line shows the one of the training set.

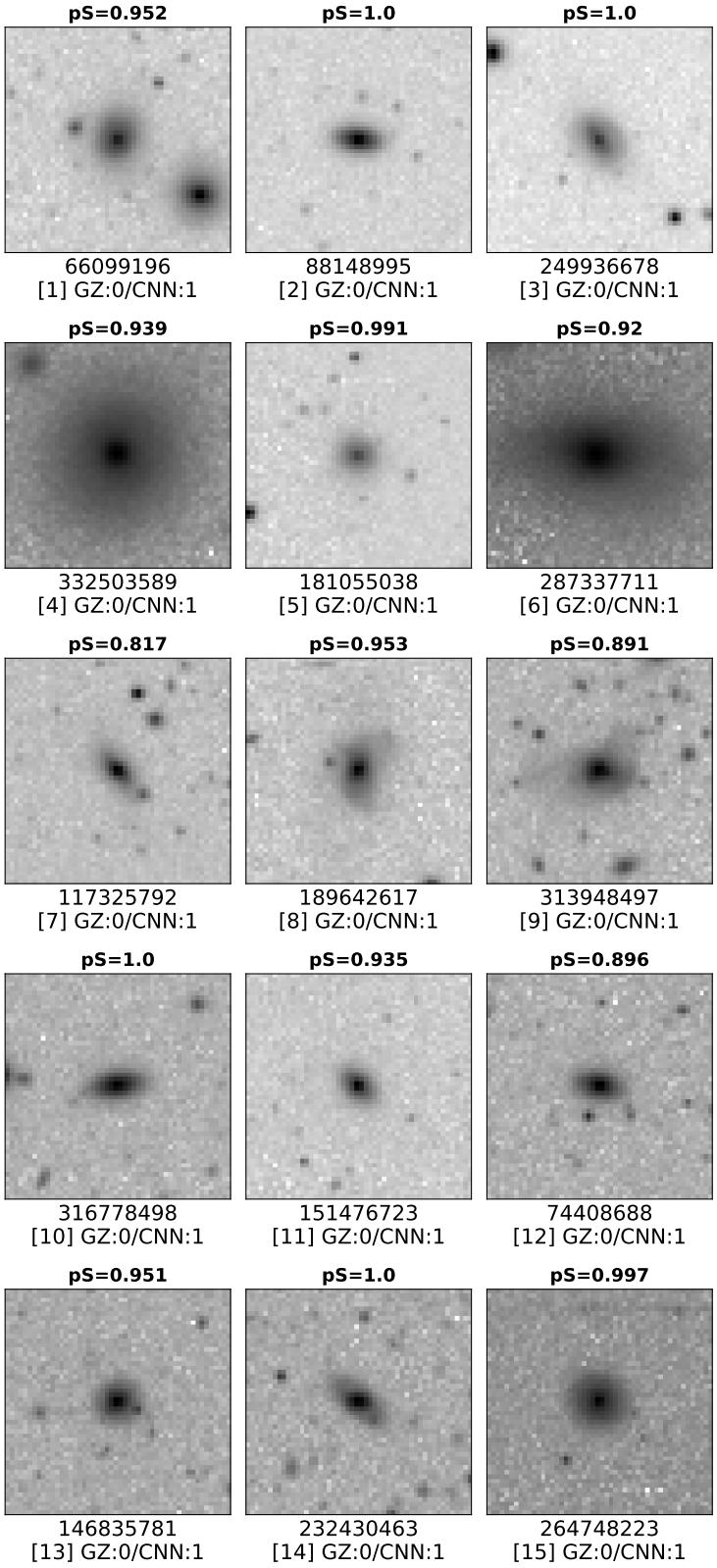

Figure 4.7: Examples of galaxies that our CNN classified as Spirals while the GZ1 labelled as Ellipticals. The predicted probability of being Spirals from the CNN is shown above each stamp (**pS**).

mismatches happened in the cases where our CNN method classified a galaxy as a spiral galaxy but we visually classified it as an elliptical galaxy. This situation is caused by the fact that our CNN is trained with the corrected debiased GZ1 classifications (Section 4.4.1); however, the visual classification used here is a raw classification. In Section 4.4.2, we pointed out that our visual classifications suffer from a similar classification bias compared with the raw GZ1 classifications which are influenced by the magnitude, size, and redshift of the targets. Therefore, in Fig. 4.9, we combine the Sérsic index and colour of each galaxy to cross-validate our results. The colour information is from the DES Y3 GOLD catalogue, and the Sérsic index is from the DES Y1 morphological measurements (Tarsitano et al., 2018) selected based upon the suggested flags (described in Section 4.4.3). Due to the applied cut in magnitude up to $i = 21.5$ used in Tarsitano et al. (2018), the last column in Fig. 4.9 only shows galaxies within the magnitude range of $21 \leq i \leq 21.5$.

In Fig. 4.9, the central contour shows the density distribution of the Sérsic index and the (g-i) colour at each magnitude. The histograms at the top and the right show their respective normalised frequency distribution. The bottom and left histograms show the misclassified samples colour-labelled by the visual classifications. From this figure, it is clear that the majority of misclassified galaxies labelled as Ellipticals by our visual assessment are in fact diskier and bluer. Since our CNN is self-debiased by training with the corrected debiased GZ1 labels, it shows a more sensible classification of the images than humans have difficulty to classify correctly. That is, our CNN classifications are more likley to be correct than the visually based ones.

We remind the reader that our CNN classifier is trained with monochromatic i-band images, without any colour information. Therefore, the strong colour segregation between CNN-classified Ellipticals and Spirals is reassuring: the connection between CNN morphology and colour is independent, and not based on the training process – colour and galaxy morphology are linked through galaxy formation and evolution processes, and are not strongly the result of classification biases.

For the faintest magnitude range in our study ($i \geq 21$), the 'self bias correction' of our CNN classifier is over applied due to the very low signal-to-noise ratio compared with the training set. This overdone bias correction gives us an artificially low number of ellipticals classified by the CNN. The ratio of the CNN-classified ellipticals to Spirals in this magnitude range is $\sim$0.00006. This is shown in both the confusion matrix and the colour-Sérsic diagram: no visually classifiable ellipticals is picked out by our CNN classifier (Fig. 4.8), and there is not a clear separation between ellipticals and Spirals in the Sérsic index distribution (Fig. 4.9). Interestingly, even though the CNN-classified ellipticals are rare and do not have the expected Sérsic index distribution, we still find a fairly good separation in their colour distribution. This indicates that the CNN-classified ellipticals with $21 \leq i \leq 21.5$ share some similarities among themselves. Therefore, this particular class of galaxies might have a different formation history from

other ellipticals, resulting in a relatively disky structure but redder colours. It would be interesting to test this hypothesis with multicolour data in the future.

Nonetheless, we exclude the CNN classifications in this magnitude range ($i \geq 21$) from our final catalogue due to the strong imbalance of the CNN classifications between the two types and the poor division in the colour-Sérsic diagram.

4.6.3 Confidence level scheme

In order to make our catalogue more accessible and easier to use reliably, we statistically assess the confidence given to our CNN classifications for each magnitude and redshift range. The confidence scheme is shown in Table 4.4. From the discussion in Section 4.6.1, our higher confidence class, 'superior confidence', is assigned to the CNN classifications for galaxies with $16 \leq i < 18$ and $z < 0.25$.

For galaxies with $18 \leq i < 21$, we carry out further statistical analyses by subdividing the galaxies in each magnitude bin into 0.25-wide redshift bins. Galaxies are excluded from the catalogue if the number of galaxies with a given morphology type falls below 30 in a given bin since we do not have the necessary statistics to assess their reliability. The excluded galaxies are generally at the highest redshifts in their magnitude bins.

Each row in Fig. 4.10 shows the diagrams within a given magnitude range, while each column presents them in a different redshift bin. We use the 'superior confidence' classifications, top-left diagram, as reference to assess the confidence level of other ranges. To be qualified as 'high confidence', the CNN predicted classifications need to follow a similar distribution to the reference sample in both Sérsic index and colour. Specifically, (1) a clear distinction needs to appear in both quantities between the two galaxy types; (2) the peaks of the distributions needs to be located at reasonable locations for both morphologies (e.g., the Sersic index distributions should peak close to ~ 1 for spirals and ~ 4 for ellipticals); and (3) no unusual features should be apparent in any of the distributions (e.g., no bimodal or messy distributions). With these criteria, a confidence level will be allocated to our CNN classifications for galaxies in each specific bin, as discussed below within the various magnitude bins.

4.6.3.1 Magnitude bins: $16 \leq i < 18$

In Section 4.6.1, we established the excellent performance of our CNN predictions for galaxies in the same magnitude and redshift ranges as the training set ($16 \leq i < 18$ and $z < 0.25$). On the first column of the first row in Fig. 4.10, we show this robust conclusion again using a parametric morphology indicator, the Sérsic index, and a generic galaxy property – its colour. The distributions of both quantities in this range are used as reference to determine the confidence level of other ranges.

First, we extend this examination to higher redshift but remain within the same magnitude range (second column at the first row in Fig. 4.10). A clear distinction

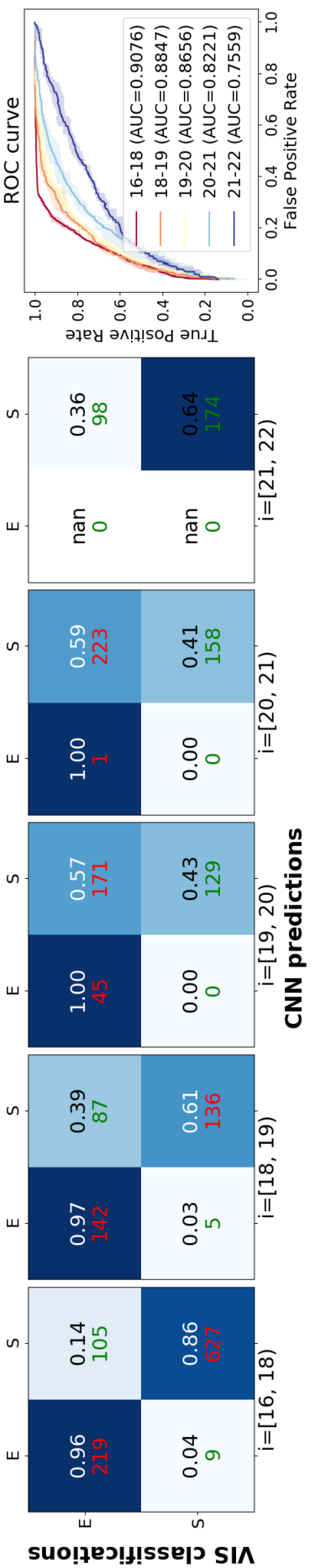

Figure 4.8: The confusion matrices and the ROC curve of different magnitude ranges. The x-axis of the confusion matrices is the CNN predictions and the y-axis is our visual classifications. On the ROC curve, the x-axis is the false positive rate while the y-axis represents the true positive rate. Different colours indicate different magnitude ranges.

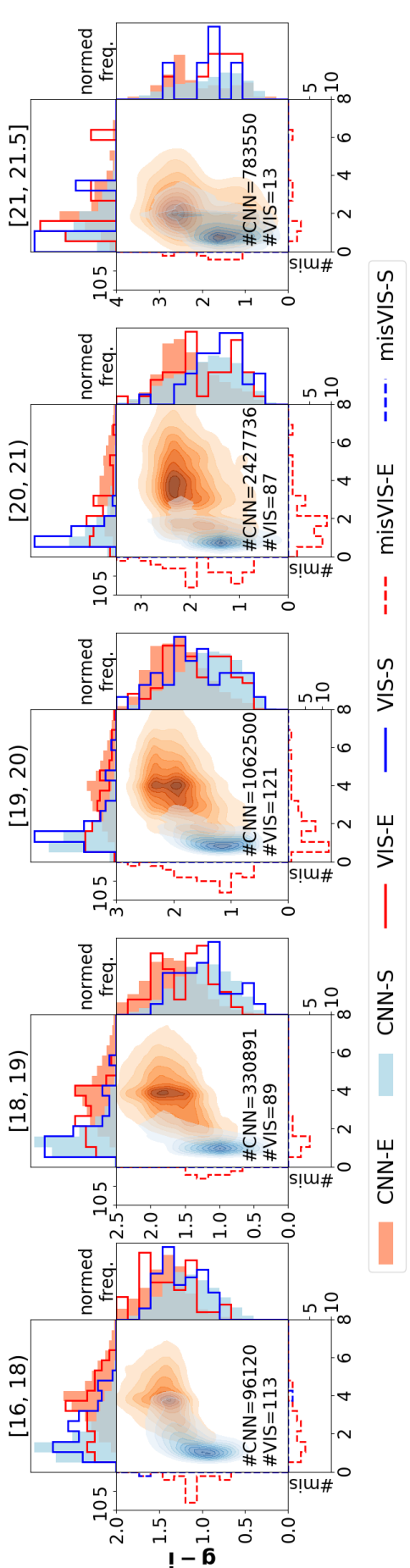

Figure 4.9: The diagram shows the colour and the Sérsic index distribution of our samples. The y-axis presents the colour g-i and the x-axis shows the Sérsic index. The central contour plot shows the two dimensional density distribution while the histograms at the top and the right show the normalised frequency distribution of each quantity, Sérsic index and g-i colour, respectively. The histograms at the bottom and the left show the samples with mismatched labels between the CNN and visual classifications for the Sérsic index and the colour, respectively. The red/orange colour represents the ellipticals, while the bluish colour is for the spiral galaxies. The shadings represents the CNN predictions, and the solid lines show the distribution labelled by the visual classifications. Finally, the dashed lines show the misclassified samples with the labels from the visual classifications.

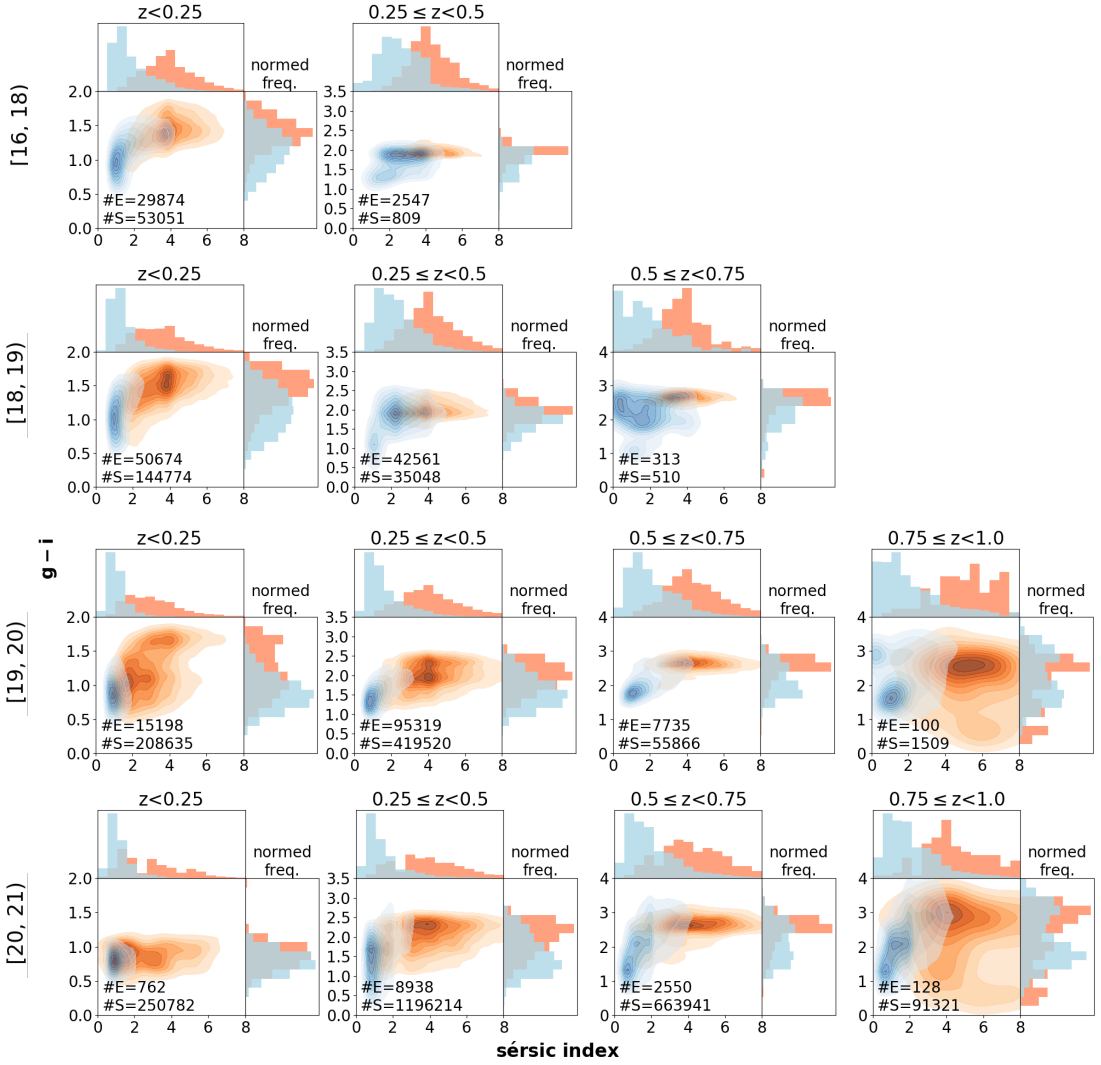

Figure 4.10: The colour-Sérsic diagrams of different redshift bins for each magnitude ranges. The histograms at the top and the right of each diagram show the normalised frequency distribution of Sérsic index and colour $g - i$, respectively. The red shading represents the ellipticals classified by our CNN, while the blue shading shows the CNN-classified Spirals. The magnitude range is shown at the left of each row while the redshift range is presented above each graph. The textual information in the diagrams shows the number of ellipticals (E) and spirals (S) classified by our CNN and with the DES Y1 morphological measurements from Tarsitano et al. (2018).

between two CNN predicted types in the Sérsic index distribution can be seen within this redshift range, $0.25 \leq i < 0.5$, and the peaks of both types are located in a sensible region. However, the CNN-classified spirals have a broader distribution compared with the reference sample; additionally, their colour distribution shows overlap with the CNN-classified ellipticals. This suggests two possibilities: (1) our CNN classifier is being less accurate within this range, and/or (2) there are a fair number of galaxies with the features of spirals but red in colour, particularly within $g - i$. Overall, we label the CNN predictions within this range as 'less confidence'.

4.6.3.2 Magnitude bins: $18 \leq i < 19$

In the magnitude range $18 \leq i < 19$, we have three redshift bins which include more than 30 galaxies with morphological measurements within each type: $z < 0.25$, $0.25 \leq z < 0.5$, and $0.5 \leq z < 0.75$. In the first plot, we notice a good differentiation between the features of ellipticals and spirals, which is reasonably consistent with the reference. Therefore, we label the predictions of this range as 'high confidence'.

The second diagram ($0.25 \leq z < 0.5$) shows similar features to the plot in the same column on the first row; however, it has a cleaner distribution in Sérsic index for CNN-classified spirals as well as a distinguishable separation in the colour distribution. Hence, we recognise the CNN classified labels in this range as 'confidence', indicating a slightly better performance than the ones within $16 \leq i < 18$ and $0.25 \leq z < 0.5$.

Finally, although the CNN-classified ellipticals show a representative Sérsic index distribution, the colour distribution has a clearly bimodal structure. Additionally, there is no sharp separation between ellipticals and spirals within the colour distribution. We thus label the classifications in this range as 'no confidence'.

4.6.3.3 Magnitude bins: $19 \leq i < 20$

In this fainter magnitude range, we observe an interesting result: a good confidence for our CNN predictions is found in the two higher redshifts bins, $0.25 \leq z < 0.5$ and $0.5 \leq z < 0.75$, than in the lower one. In these ranges, the distribution of galaxy properties are clearly separated for the different morphologies, and the peaks of the distributions are also reasonable. We therefore give the morphological classifications for galaxies in these redshift ranges a 'high confidence' label.

The low redshift interval ($z < 0.25$; first column) shows a worse performance. We find a flat Sérsic index distribution for the CNN-classified ellipticals which peaks at roughly $n \sim 2$. Additionally, although there is a separation in the colour distribution between the two types, the CNN-classified ellipticals show a clearly bimodal colour distribution which partially overlaps with the one of the CNN-classified spirals. Although the performance for ellipticals in this redshift

range is clearly worse, the behaviour for spirals is significantly better: there is a fairly good discrimination in both the Sérsic index and the colour distributions. This means that in this redshift range, our CNN-classified spiral sample has a high purity but not a high completeness. We therefore label the classifications made in this range as 'less confidence' but with a '*' mark (Table 4.4). The '*' indicates that this confidence level is only defined for CNN-classified spirals, and the classified ellipticals are labelled as 'no confidence'. Clearly this sample cannot be used to find all spirals, but we do have some confidence in the morphologies for the ones it does classify.

It seems counter-intuitive that a better performance is found for higher redshift galaxies than for lower redshift ones at these faint magnitudes. However, the reason is that the fainter galaxies in the training set tend to be at higher redshifts. Therefore, there is a somewhat better overlap in the properties of faint higher redshift galaxies than there is for faint lower-redshift ones between the general DES Y3 sample and the training set.

Finally, for this magnitude range we give a 'no confidence' label to the highest redshift range ($0.75 \leq z < 1.0$). This is due to the messy galaxy property distributions reflected in the bimodal colour distributions for both morphological types, a significantly higher Sérsic index than expected for the CNN-classified ellipticals, and a relatively low Sérsic index for the CNN-classified spirals. Interestingly, despite the relatively anomalous Sérsic index, a fairly sharp differentiation between both types is shown in the Sérsic index distributions. For the CNN-classified ellipticals, it suggests a class of red galaxies which has a higher concentration and a more peaked surface brightness distribution than expected. This is an interesting conclusion from our CNN classification analysis that deserves to be explored further in future work.

4.6.3.4 Magnitude bins: $20 \leq i < 21$

As we get to fainter magnitudes using our CNN methodology to classify galaxies becomes more of a challenge. From Fig. 4.8, we notice that there are also significantly fewer galaxies classified as ellipticals by our CNN, such that the CNN-classified E/S ratio is $\sim$0.003 in this range, while the ones in other brighter ranges have a ratio over 0.1. This indicates that the bias self-correction by our CNN classifier is overdone in this range compared to the brighter ranges. However, unlike the result shown in the range $21 \leq i \leq 21.5$ in Fig. 4.9, a better and clearer separation between both types in Sérsic index and colour is presented. Hence, we carry out a further investigation within different redshift bins for this range.

In the first plot on the bottom row in Fig. 4.10, the distributions of CNN-classified spirals are fairly reasonable. However, a bimodal distribution and incorrect peak assignment of the Sérsic index occurs within the CNN-classified ellipticals. Therefore, we decided to assign a class of 'less confidence' with * for this range, where * means this confidence label is for the Spirals (no confidence to the classification of Ellipticals).

The second plot for this magnitude range in Fig. 4.10 shows a good separation between the two types of galaxies in Sérsic index and colour space. Although a strong imbalance in the number of ellipticals and spirals still exists here, the differentiation in two types proves a certain degree of confidence to our CNN predictions. Hence, we label the predictions in this range as 'confidence'. The reason for this good separation in this significantly fainter magnitude range is also due to the effect discussed in Section 4.6.3.3 that the galaxies in this magnitude ($20 \leq i < 21$) and redshift ($0.25 \leq z < 0.5$) range have similar galaxy features and galaxy properties to the reference samples. The shift in magnitude for these galaxies is due to the change in redshifts.

This situation is also demonstrated within the third plot of Fig. 4.10 ($20 \leq i < 21$ and $0.5 \leq z < 0.75$) whereby both types are distinguished in Sérsic index and colour distributions. However, CNN-classified spirals have a relatively flat colour distribution, preventing a clear separation. Hence, a class of 'less confidence' is assigned to this range. Finally, the last diagram shows a messy distribution. Therefore, we simply label this range as a 'no confidence' class.

4.6.4 Non-parametric methods and galaxy properties

Another examination is carried out using non-parametric methods such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), Gini coefficient, and M20. In this study, the non-parametric measurements are from Tarsitano et al. (2018) using the i-band images, and we use the measurements after applying the selection criteria described in Section 4.4.3.

Furthermore, this validation can work in both directions. We can use the non-parametric measurements to check the robustness of our CNN-based morphological classifications, while also use out most reliable morphological classifications (those with 'superior confidence') to assess the ability of non-parametric methods to separate the ellipticals and the spirals (Fig. 4.11). Such an analysis of non-parametric measurements as proxies for morphology has never before been carried out with samples as large as ours. In this work we include over 100,000 galaxies in the 'superior confidence' category from our DES Y1 morphological classifications.

In Fig. 4.11, we show the pair plots of six different parameters: concentration (C), asymmetry (A), clumpiness (S), Gini, M20, and Sérsic index. For the A, S parameters, we only showcase the data with values smaller than 0.2 to focus on 'typical galaxies'. The Sérsic index is used as a comparison to the non-parametric methods, and it is one of the main features used to define the confidence level (Section 4.6.3). It shows a clear separation between the two morphological types here. In addition to this, we note that only the Gini coefficient shows a consistently distinguished difference between the two types in the histogram.

The Gini coefficient (G) reflects the inequality of the flux distributed among the pixels of a given galaxy; if $G = 1$, the light is concentrated in one pixel, conversely, $G = 0$ means that the light is uniformly distributed to every pixel.

Therefore, the Gini coefficient is somewhat analogous to the concept of concentration, and ellipticals generally have a higher value than spirals. Nevertheless, the concentration does not show a separation as good as the one for the Gini coefficient. A slight shift between the peaks of the two morphological types is shown in the histogram of the concentration; however, a large overlapping area is also shown. Additionally, the difference of the mean concentration values between both types is relatively small compared with previous studies (Conselice, 2003; Hernández-Toledo et al., 2008; Hambleton et al., 2011). On the other hand, both asymmetry and clumpiness also fail to show a consistent distinction between the two morphological types in our analysis.

Finally, the M20 histogram does not show a clear separation between the two morphological types neither. However, a clean separation shows in the contour of the Gini coefficient and M20. The black dashed line indicates a cut used to separate Ellipticals and Spirals and described in Lotz et al. (2008) such that

$$G = 0.14M_{20} + 0.8. \tag{4.1}$$

Thus we find that the Gini coefficient is a possible better tracer of the overall structure of a galaxy than any other non-parametric morphological quantities such as C, A, S, and M20 (Zamojski et al., 2007) when separating ellipticals from spirals.

4.7 Conclusion

We present in this chapter one of the largest galaxy morphological classification catalogue produced to date, using the Dark Energy Survey (DES) Y3 data with over 20 million galaxies. We carry out these classifications using convolutional neural networks (CNN) trained with the subset of a DES Y1 data. The **corrected** debiased labels, which are initially from the Galaxy Zoo 1 (GZ1) catalogue and corrected in Chapter 3, are used to label our training set (Section 4.2.2). With a combination of three different types of input, including: linear images, log images, and HOG images (Section 4.2.1), our CNN classifier reaches an accuracy of over 99% when compared with the GZ1 classifications. The majority of mismatches occurs in the case when a galaxy is classified as a Spiral by our CNN but as a Elliptical by GZ1. The reason behind this situation is likely to be the better resolution and deeper depth of the DES imaging data which reveals unnoticeable structure in the data used in GZ1 from the Sloan Digital Sky Survey (SDSS) (Cheng et al., 2020a, Chapter 3).

Additionally, training with the **corrected** debiased labels, our CNN classifier is shown to be self-debiased and more accurate in classifying disk galaxies which human visual classifications have difficulty detecting at faint magnitudes down to $i \sim 21$ (see Section 4.6.2).

Using a cross-validation with the Sérsic index and galaxy colour, we provide a confidence evaluation scheme to our CNN classifications (Table 4.4) through a statistical analysis of data in different magnitude and redshift bins (Section 4.6.3).

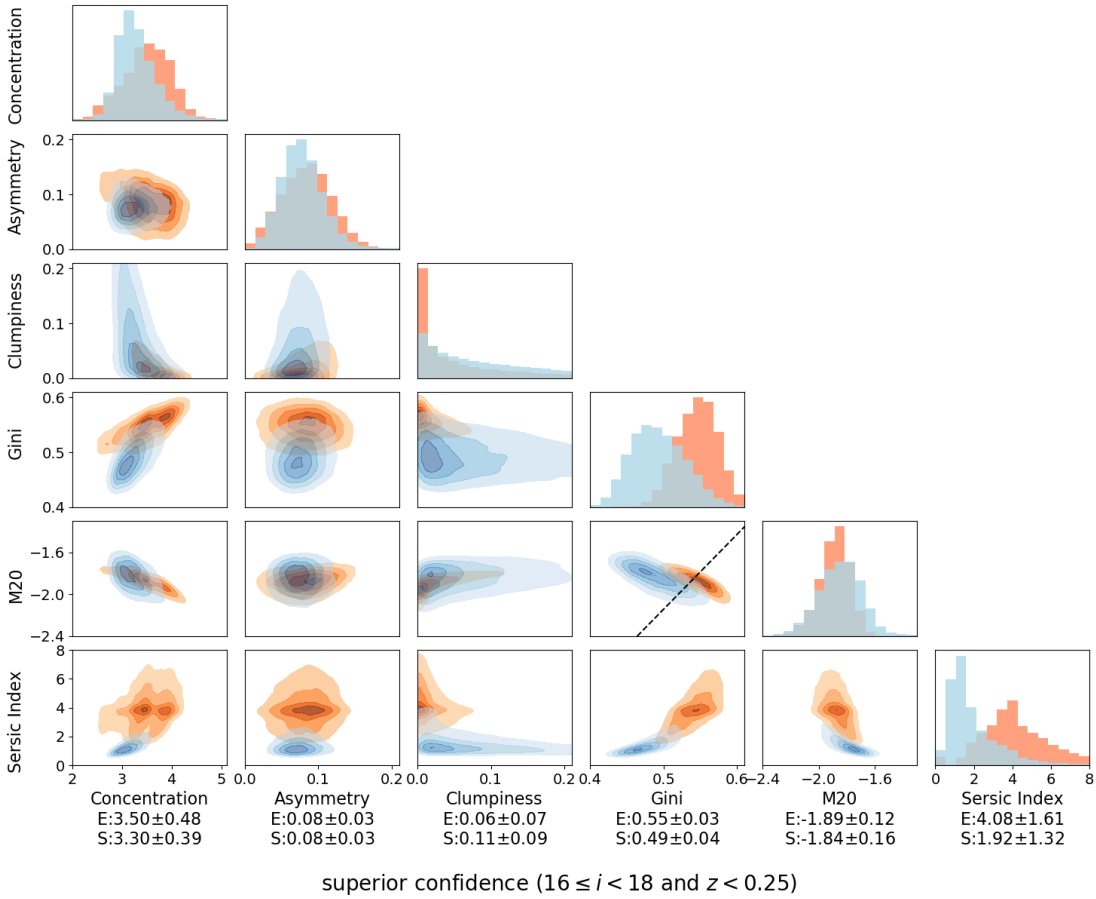

Figure 4.11: The pair plots of six morphological parameters: concentration, asymmetry, clumpiness, Gini, M20, and Sérsic index labelled by the CNN classifications with 'superior confidence'. The colour shadings represent the CNN classifications. The red/orange and blue colour are for Ellipticals (E) and Spirals (S), respectively. The mean value of each parameter for both types with the standard deviation is shown below each column. The black dashed line shows a cut from Lotz et al. (2008) to separate ellipticals and spirals based on the M20 and the Gini coefficients.

As a part of the validation, we carry out a large examination of non-parametric methods such as the *CAS system* (Concentration, Asymmetry, and Smoothness/Clumpiness), the Gini coefficient, and M20 using over 100,000 classifications with structural measurements from Tarsitano et al. (2018).

From this we conclude that the Gini coefficient shows the most significant distinction, as a single parameter, between ellipticals and Spirals within all parameters tested. This is such that a straight line can be drawn to separate these two types on a Gini coefficient and M20 diagram (Fig. 4.11). Our new morphological catalogue allows a variety of new approaches towards understanding galaxy properties and evolution that involve morphology that could not be carried out before. For example, non-parametric analysis methods of galaxy structure can be assessed using an unprecedented sample not only in size but also in quality. Our catalogue can also be used to cross-validate other classification methods, and to explore galaxy properties as a function of morphology with superb statistics. Such investigations could include, among others, the dependence of the E/S ratio with redshift and magnitude and studies of galaxy morphology divided into many divisions of galaxy property and environment.

Scientifically, there are of course a myriad of uses for our catalog, as morphology is one of the fundamental properties of galaxies. Future work within and outside the DES collaboration will investigate these issues. For the time being this will remain one of the largest set of mythologies available for analysis for any survey done to date. Euclid and LSST will however supersede these numbers but our methods and tools can easily be applied to this imaging data once it is available.

Chapter 5

New Narrator - Unsupervised Machine Learning with Convolutional Autoencoder for Strong Lensing Identification

This chapter is based on published material by ***Ting-Yun Cheng****, Nan Li, Christopher J. Conselice, Alfonso Aragón-Salamanca, Simon Dye, Robert B. Metcalf. Monthly Notices of the Royal Astronomical Society, Volume 494, Issue 3, May 2020, Pages 3750–3765.*

Abstract

In this chapter we develop a new unsupervised machine learning technique comprised of a feature extractor, a convolutional autoencoder (CAE), and a clustering algorithm consisting of a Bayesian Gaussian mixture model (BGM). We apply this technique to visual band space-based simulated imaging data from the Euclid Space Telescope using data from the Strong Gravitational Lenses Finding Challenge. Our technique promisingly captures a variety of lensing features such as Einstein rings with different radii, distorted arc structures, etc, without using predefined labels. After the clustering process, we obtain several classification clusters separated by different visual features which are seen in the images. Our method successfully picks up $\sim$63 percent of lensing images from all lenses in the training set. With the assumed probability proposed in this study, this technique reaches an accuracy of $77.25 \pm 0.48\%$ in binary classification using the training set. Additionally, our unsupervised clustering process can be used as the preliminary classification for future surveys of lenses to efficiently select targets and to speed up the labelling process. As the starting point of the astronomical application using this technique, we not only explore the application to gravitationally lensed systems, but also discuss the limitations and potential future uses of this technique.

5.1 Introduction

In previous chapters, we focused on the supervised machine learning applications; however, labelling data for the use of supervised methods can be extremely time expensive. Additionally, using the prior knowledge of labels defined by human is prone to result in classification bias during the training process. Unlike supervised machine learning, which requires a large amount of labelled data, unsupervised machine learning can be applied directly to observed data without labelling, this helps to reduce human bias while training a machine. Therefore, scientists have started to explore the application of unsupervised machine learning to, e.g., photometric redshifts (e.g., Geach, 2012; Way and Klose, 2012; Carrasco Kind and Brunner, 2014; Siudek et al., 2018a), as well as to classification using photometry or spectroscopy (e.g., D'Abrusco et al., 2012; Fustes et al., 2013; Siudek et al., 2018b).

The application of unsupervised machine learning becomes more challenging when using high dimensional data such as images. Hocking et al. (2018) and Martin et al. (2019) are amongst the first studies of unsupervised machine learning applications using imaging data applying the Growing Neural Gas algorithm (Fritzke, 1995). In this study, we explore a different technique from Hocking et al. (2018) and Martin et al. (2019) in which we apply a convolutional autoencoder (CAE, Masci et al., 2011) to do feature extraction before connecting with unsupervised machine learning algorithms.

We test this proposed unsupervised machine learning method in the task of identifying galaxy-galaxy strong lensing (GGSL) system. A GGSL system is a particular case of gravitational lensing in which the background source and foreground lens are both galaxies, and the lensing effect is sufficient to distort images of the source into arcs or even Einstein rings. Since the discovery of the first GGSL system in 1988 (Hewitt et al., 1988), they have been used in many valuable scientific applications, such as studying galaxy mass density profiles (e.g., Sonnenfeld et al., 2015; Shu et al., 2016a; Küng et al., 2018), detecting galaxy substructure (e.g., Vegetti et al., 2014; Hezaveh et al., 2016; Bayer et al., 2018), measuring cosmological parameters (e.g., Collett and Auger, 2014; Rana et al., 2017; Suyu et al., 2017), investigating the nature of high redshift sources (Bayliss et al., 2017; Dye et al., 2018; Sharda et al., 2018), and constraining the properties of the self-interaction physics of dark matter (e.g., Shu et al., 2016b; Gilman et al., 2018; Kummer et al., 2018).

Increasing the statistical power of these applications to gravitational lensing and improving sample uniformity requires a large increase in the number of known GGSL systems. Next generation imaging surveys arising from facilities such as Euclid, the Large Synoptic Survey Telescope (LSST), and the Wide Field Infrared Survey Telescope (WFIRST) are anticipated to increase the number of known GGSL systems by several orders of magnitude (Collett, 2015). These forthcoming datasets present a challenge for identifying new GGSLs using automated procedures that operate in an efficient and reliable manner. To this end, a number of algorithms have been developed to detect GGSLs in image

data by recognising arc-like features and Einstein rings (e.g., Gavazzi et al., 2014; Joseph et al., 2014; Paraficz et al., 2016; Bom et al., 2017). In addition, instead of recognising arc-like features, an alternative detection technique that has had some success is to attempt to fit lens mass models to candidate GGSLs and reject those systems that do not converge (Marshall et al., 2009; Sonnenfeld et al., 2018).

More recently, efforts to automate GGSL finding have turned to machine learning algorithms given their strong performance in the general field of image recognition, in particular, the CNN. These algorithms are widely used in categorizing galaxy morphologies (e.g., Dieleman et al., 2015; Huertas-Company et al., 2015; Domínguez Sánchez et al., 2018; Walmsley et al., 2020), measuring photometric redshifts (Cavuoti et al., 2017; Sadeh et al., 2016; Samui and Samui Pal, 2017), and classifying supernovae (Lochner et al., 2016, see also Section 1.2). Recent work has also shown that CNN can be used to perform lens modelling as a vastly more efficient alternative to traditional parametric methods (Hezaveh et al., 2017; Pearson, J. et al., 2019).

The application of CNN to the detection of GGSL systems has reached a high success rate in binary classification (Jacobs et al., 2017; Petrillo et al., 2017; Ostrovski et al., 2017; Bom et al., 2017; Hartley et al., 2017; Avestruz et al., 2019b; Lanusse et al., 2018). However, as mentioned previously, supervised methods suffer from some degree of classification bias from the labelled data which may not properly represent the diversity of real GGSL systems observed in future surveys. Additionally, GGSLs are rare events in the Universe so there are insufficient homogeneous data for training in supervised machine learning methods. Although simulated images can be used for training, they are generally lacking in the complexity of real observed data.

We use our unsupervised machine learning method to provide an alternative to humans identifying GGSLs without providing the labels needed in supervised methods. Our system can also be used for the preliminary selection of GGSL candidates in future imaging surveys. Furthermore, without human bias, we can explore unique GGSL systems that would not be found without unsupervised machine learning techniques.

This chapter is structured as follows. The unsupervised machine learning technique adopted in this study is introduced in Section 5.2. Details about the implementation, including the pipeline and dataset, are described in Section 5.3. Section 5.4 discusses our findings. Future work is discussed in Section 5.5. Finally, the conclusions are presented in Section 5.6.

5.2 Methodology

The application of unsupervised machine learning has achieved successes on one dimensional data in astronomy such as with spectroscopic data or photometric parameters (e.g., D'Abrusco et al., 2012; Geach, 2012; Way and Klose, 2012;

Fustes et al., 2013; Carrasco Kind and Brunner, 2014; Siudek et al., 2018a,b). However, the capability of unsupervised machine learning for high dimensional data such as imaging data has not been well explored.

The latest astronomical approaches of unsupervised machine learning application using imaging data made by Hocking et al. (2018) and Martin et al. (2019) apply the concept of deep clustering. Deep clustering (e.g., Hsu and Kira, 2015; Hershey et al., 2015; Xie et al., 2016; Caron et al., 2018) is a clustering method that groups together the features learned through a neural network. Both Hocking et al. (2018) and Martin et al. (2019) apply a neural network called 'growing neural gas algorithm' (GNG; Fritzke, 1995), which is a type of self-organizing map (Kohonen map; Kohonen, 1997), to create feature maps from imaging data. They then connect these feature maps with a hierarchical clustering technique (Hastie et al., 2009).

In addition to neural networks, studies in computer science also use an architecture of both supervised (CNN) and unsupervised convolutional neural networks (UCNN) (e.g., Dosovitskiy et al., 2014) to the process of feature learning (computer science: e.g., Dundar et al., 2015; Bautista et al., 2016; Borji and Dundar, 2017).

There are a variety of unsupervised approaching for deep clustering using the architecture of CNN. However, most of them use alternative unsupervised algorithms (e.g., k-mean) to calculate the weights between layers that reduces the power of CNN for capturing features fit with human judgement when using imaging data. Therefore, instead of variational CNN, we propose to use a convolutional autoencoder (CAE, Section 5.2.1) as the feature extractor (Masci et al., 2011) in this study. This preserves the intrinsic features of the images (Guo et al., 2017; Li et al., 2017; Dizaji et al., 2017). For the clustering part we apply the Bayesian Gaussian mixture model (BGM, Section 5.2.2) to images presented by the features extracted by the CAE to group the input features in a high-dimensional feature space.

5.2.1 Convolutional AutoEncoder (CAE)

The convolutional autoencoder (CAE) (Masci et al., 2011) is a kind of autoencoder (AE) which is mostly well known for denoising images (Vincent et al., 2010). The function of an AE is to learn a prior which features best represent the data distribution. With a limited number of features available, an AE intentionally captures significant features from images rather than the details of the background noise. The AE can then reconstruct images with this obtained prior.

The CAE improves the performance of an AE by considering the structures within two dimensional images that are ignored in the AE. Hence, the CAE preserves spatially localised features from image patches, while the AE can only obtain the global features.

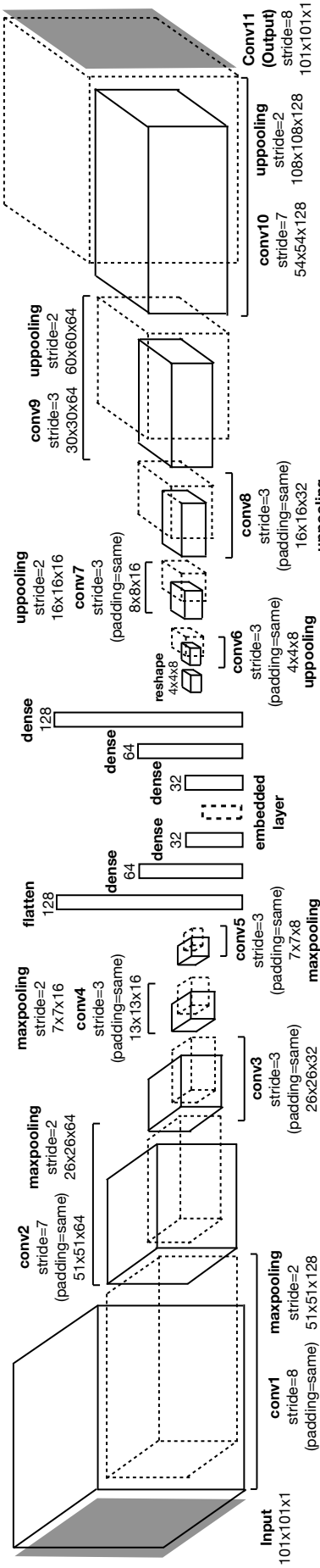

Figure 5.1: The schematic overview for the architecture of our convolutional autoencoder (CAE) which is composed of two parts, the encoder and the decoder. The encoder starts from an input image with a size of 101 by 101 pixels (leftmost side) which is then connected with 5 convolutional layers (filter size: 128, 64, 32, 16, and 8). Each convolutional layer is followed a pooling layer. Three dense layers (units: 128, 64, 32) follow the fifth convolutional layer. The central dense layer of the architecture is called the 'embedded layer'. We explore different number of units for this layer in this study (section 5.3.2). The decoder has similar structure to the encoder, and we use the units in the embedded layer to reproduce the input image as the output (rightmost side).

The architecture of the CAE used in this study is shown in Fig. 5.1. It includes two parts: encoder (left) and decoder (right). The encoder extracts the representative features from the input image. For an input x, the j-th representative feature map is given by

$$h^j = f\left(x * W^j + b^j\right), \tag{5.1}$$

where W are filters, $*$ denotes the 2 dimensional convolution operation, b is the corresponding bias of the j-th feature map, and f is an activation function. The encoder in this study is built with five convolutional layers (filter size: 128, 64, 32, 16, and 8) and three dense layers (units: 128, 64. 32). The activation function used in the convolutional layers is the Rectified Linear Unit (`ReLu`) (Nair and Hinton, 2010) such that $f(z) = 0$ if $z < 0$ while $f(z) = z$ if $z \geq 0$. Each convolutional layer is followed by a pooling layer with a size of 2 by 2 pixels. The pooling layer is also referred to as a downsampling layer which is to reduce the spatial size and reduce the parameters involved in the CAE.

The decoder then reproduces input images from the representative features; therefore, the architecture of the decoder is symmetric but reverse to that of the encoder. We invert the procedure of the encoder to reconstruct the representative feature maps back to the original shape of the input image by using the following formula:

$$y = f\left(\sum_{j \in H} h^j * \widetilde{W}^j + c\right), \tag{5.2}$$

where $\widetilde{W}$ is the flip operator that transposes the weights, $*$ denotes 2 dimensional convolution operation, c is the corresponding bias, f is an activation function, and H indicates the group of feature maps. The design for the number of filters in the convolution processes is based on the size of input images to form a symmetric structure between encoder and decoder.

We have three dense layers (units: 32, 64, and 128), five convolutional layers (filter sizes: 8, 16, 32, 64, and 128) using the `ReLu` activation function (Nair and Hinton, 2010), and an extra convolutional layer (filter: 1) using the `softmax` function (Bishop, 2006), $f(z) = exp(z)/\sum exp(z^j)$, as the output for the decoder. Each convolutional layer apart from the last layer (output) is followed with an upsampling layer which has the opposite function to the pooling layer that is used for recovering the resolution.

The central dense layer of the CAE is called the 'embedded layer (EL)' (see Fig. 5.1). This is composed of the final latent representation features used for the reconstruction of the input images. In section 5.3.2, we explore the number of units required for the EL.

The CAE extracts the latent representative feature maps by minimizing the reconstruction error. In this study, we use `binary_crossentropy` in the KERAS library[1] to calculate the loss function of the CAE which is given by the following

[1] https://keras.io

form,

$$L = -\frac{1}{N}\sum_{n=1}^{N}\left[y^n \log \hat{y}^n + (1-y^n)\log\left(1-\hat{y}^n\right)\right], \tag{5.3}$$

where N is the number of samples, y^n are targets, and $\hat{y}^n$ are the reconstructed images (equation 5.2). We build our CAE using the KERAS library and the TENSORFLOW backend [2] (Abadi et al., 2015a).

5.2.2 Bayesian Gaussian Mixture Model (BGM)

A Gaussian mixture model is a probabilistic model for either density estimation or clustering using a mixture of a finite number of Gaussian distributions to describe the distributions of data points on a feature map. Given K components, the algorithm uses `Kmeans` to initialise the weights, the means, and the covariances for the K Gaussian distributions which are given in the form:

$$p\left(x\right) = \sum_{k=1}^{K} w_k G\left(x|u_k, \varepsilon_k\right), \tag{5.4}$$

where $G\left(x|u_k, \varepsilon_k\right)$ represents k-th Gaussian, u_k denotes the mean of the k-th Gaussian distribution, ε_k is the covariance matrix of the k-th Gaussian, and w_k is the prior probability (weight) of the k-th Gaussian where,

$$\sum_{k=1}^{K} w_k = 1. \tag{5.5}$$

The algorithm then searches for the best fit of the K Gaussian distributions to the data distribution through an iterative process.

A two dimensional illustration of the BGM is shown in Fig. 5.2 (Equation 5.4). The input data are distributed on the feature map (black dots). We use 3 Gaussian distributions in this illustration (coloured ellipses), to fit the data distribution on the feature map.

In unsupervised learning, expectation-maximization (EM) (Hartley, 1958; Dempster et al., 1977; McLachlan and Krishnan, 1997) is used to find the maximal log-likelihood estimates for the parameters of the Gaussian mixture model by an iterative process. The log-likelihood of the Gaussian mixture model is calculated using the formula:

$$\ln\left[p\left(x|u, \varepsilon, w\right)\right] = \sum_{n=1}^{N}\left\{\ln\left[\sum_{k=1}^{K} w_k G\left(x|u_k, \varepsilon_k\right)\right]\right\}, \tag{5.6}$$

where N is the number of samples.

[2] https://www.tensorflow.org

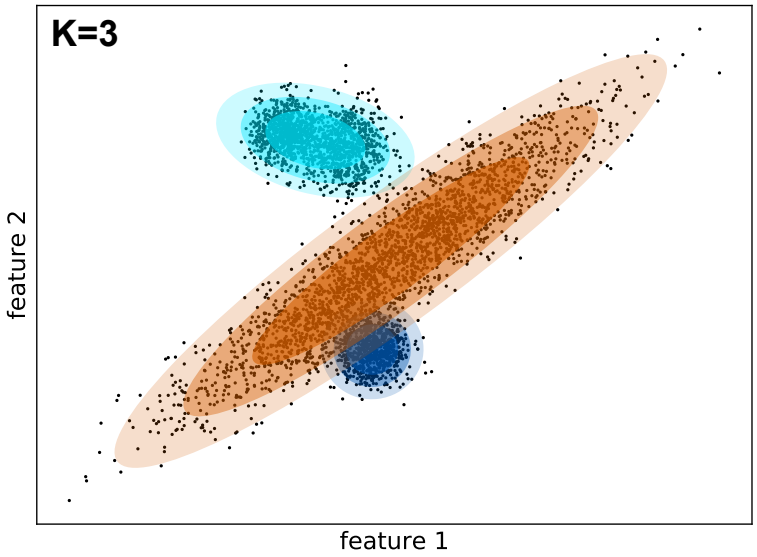

Figure 5.2: An illustration of the Gaussian Mixture model we use. The K value is the number of Gaussian distributions. The black dots show the data distribution on the feature map, and the coloured ellipses represent the three Gaussian distribution we applied here to fit the data distribution.

The Bayesian Gaussian mixture model (BGM) is a variational Gaussian mixture model (Kullback and Leibler, 1951; Attias, 2000; Bishop, 2006) which maximises the evidence lower bound (ELBO) (Kullback and Leibler, 1951) in the log-likelihood. In this study, we apply the BGM from the SCIKIT-LEARN library [3] (Pedregosa et al., 2011).

5.3 Implementation

In this section, we first introduce the datasets used in this study. The feature learning procedure is discussed in section 5.3.2. Section 5.3.3 presents the clustering and classifying phase which explains how to obtain the predicted lensing probability for each image. The tests for quantifying the performance of the classifications are described in section 5.3.4.

5.3.1 Data Sets

The strong lensing data are from the Strong Gravitational Lens Finding Challenge (Lens Finding Challenge; Metcalf et al., 2019b). The generation of mock images follows the procedures described in Grazian et al. (2004) and Meneghetti et al. (2008), and starts with a cosmological N-boby simulation, the Millennium simulation (Boylan-Kolchin et al., 2009). The background objects are modeled by the sources from the Hubble Ultra Deep Field (UDF). The detail of the simulation setup can be found in Metcalf et al. (2019b).

[3] https://scikit-learn.org/stable/index.html

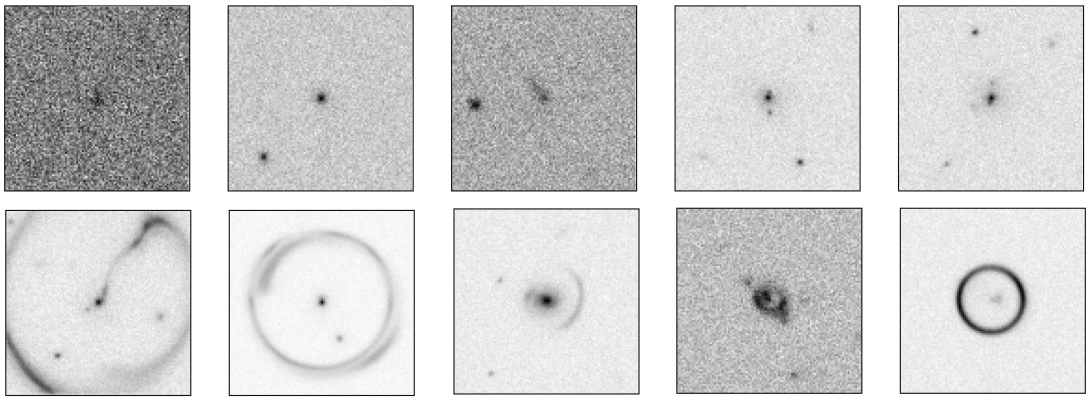

Figure 5.3: An example of the training set for Lens Finding Challenge *Top:* non-lensing image; *Bottom:* lensing image.

We use the datasets which mimic the data quality of observations that will be taken by the Euclid Space Telescope (Laureijs et al., 2011) in the visual (VIS) band. The pixel size is set to 0.1 arcsec and a Gaussian point spread function is applied to the images. Additionally, the noise follows a Gaussian distribution which is added to the final images (Metcalf et al., 2019b).

There are 20,000 labelled images with lenses for training (13,968 lensing images; 6,032 non-lensing images, see Fig 5.3) and 100,000 unlabelled images with lenses for testing in the Lens Finding Challenge.

We split the training set received from the Lens Finding Challenge into two parts, our own training set and testing sets. We randomly pick 12,800 lensing images out of 13,968 lensing images to obtain enough information for feature extraction. Additionally, we rotate a random set of 3,200 non-lensing images 4 times (0, 90, 180, 270 degrees) to obtain the same number of images as there are lensing images (12,800 images) for our training set. An extra insignificant Gaussian noise is added into the rotated images to enhance the difference between the rotated images and the original images. The ratio between lensing and non-lensing images is 1 in the training set to make the convolutional autoencoder (CAE) consider both types equally when extracting features.

The rest of the images are the candidates for the testing sets. In our own testing sets, we initially have 1,168 lensing and 2,832 non-lensing images, which are leftover from the selection of the training set. We rotate the non-lensing images 4 times (0, 90, 180, 270 degrees) and add Gaussian noise to increase the number of images to 11,328 non-lensing images.

We test several different ratios between the number of lensing and non-lensing images to mimic a more realistic case. To avoid a biased influence from lensing images, we use the same set of lensing images in the testing process. We generate different ratios by randomly and repeatedly picking samples from the set of rotated non-lensing images. The arrangement is shown in Table 5.1 and is based on

Labels	Ratios	Number of data in each type
1	1:1	lensing:1168/ non-lensing:1168
2	1:2	lensing:1168/ non-lensing:2336
3	1:20	lensing:1168/ non-lensing:23360
4	1:50	lensing:1168/ non-lensing:58400
5	1:100	lensing:1168/ non-lensing:116800
6	1:1000	lensing:1168/ non-lensing:1168000
7	1:10000	lensing:1168/ non-lensing:11680000

Table 5.1: The arrangement of the testing datasets in this study. The ratios between lensing and non-lensing images are shown in the second column and the content included in the datasets are shown in the third column.

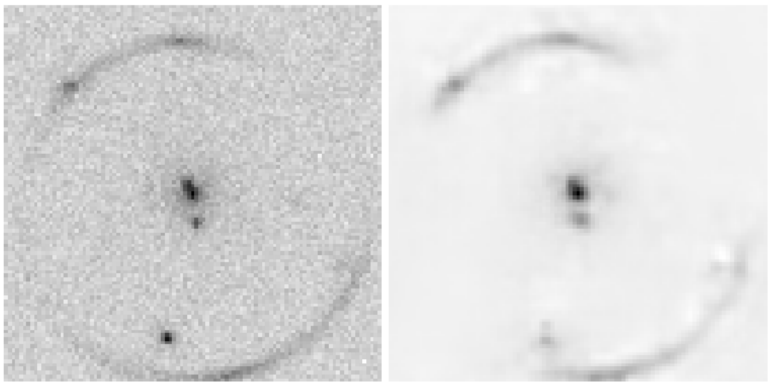

Figure 5.4: An example of the denosing process. *Left*: the original image. *Right*: the image after denoising by an alternative CAE architecture described in section 5.3.2

the prediction of Collett (2015) which forecasts 2,400, 120,000, and 170,000 detectable galaxy-galaxy strong lenses out of 11 million lenses from their model for lensing systems in the Dark Energy Survey[4], Large Synoptic Survey Telescope[5], and Euclid Space Telescope, respectively. This arrangement for the fractions of lensing images in the testing sets cover from 50 percent to 0.01 percent.

5.3.2 Feature Learning

There are three steps to take in the application of the techniques used in this study: (1) denoising the images by the convolutional autoencoder (CAE) with a simpler structure; (2) extracting the features of the images using the CAE (Fig. 5.1); (3) identifying clusters using the features extracted from the CAE by the Bayesian Gaussian mixture model (BGM).

We recognise that the background noise in images influences the result of feature extraction because the CAE can overfit to the noise. As mentioned in Section 5.2.1, an autoencoder learns the prior distribution from the input images (with noise) which preferentially captures the representatively strong features in

[4]https://www.darkenergysurvey.org/
[5]https://www.lsst.org

images, but ignores insignificant features such as noise. Therefore, the reconstruction based on the prior distribution learnt through an autoencoder generates noiseless reconstructed images. We apply a CAE with a simpler architecture without hidden layers in Fig. 5.1 to generate noiseless images at the first step.

This architecture contains five convolutional layers (filters: 128, 64, 32, 16, 8) with `ReLu` activation function for the encoder, five convolutional layers (filters: 8, 16, 32, 64, 128) with `ReLu` activation function for the decoder, an output layer with a `softmax` activation function. Each convolutional layer is followed with either a pooling layer or an upsampling layer in the encoder or decoder, respectively. The effect is shown in Fig. 5.4. The left panel is the original image, and the right panel is the image after denoising. Although the reconstructed images have lower resolution, they preserve and emphasize the features of lenses and sources that helps our CAE (Fig. 5.1) to capture meaningfully representative features from images in the second step.

Secondly, we apply the CAE to carry out feature extraction (Fig. 5.1). The final representative features are located within the embedded layer (EL) in the centre of the architecture. Finally, these extracted features are the input for the third step - clustering using the Bayesian Gaussian mixture model (BGM) utilising the representative features extracted by the CAE from the images.

The number of clusters, K, when using unsupervised machine learning is generally unknown and difficult to be determined as there is not yet a reliable optimisation process to decide this quantity in unsupervised machine learning.

In Guo et al. (2017), they suggest the number of extracted features to use should be the same as the number of clusters of datasets used (MNIST[6]). These number of clusters are however known in their case. This arrangement ensures that: (1) the dimension of the embedded layer was lower than the input data, and (2) the network could be trained directly in an end-to-end manner without any regularisations.

In contrast, the number of clusters is unknown in our work, and the number of extracted features is a hyper-parameter which can be controlled. Therefore, we decided to set the number of clusters, K, using the opposite concept from Guo et al. (2017), to be the same as the number of extracted features.

We can explain this decision using a simplified condition by assuming each feature decides one cluster; therefore, the number of features would be the intrinsic minimal number of clusters used.

The process of feature learning using the CAE is computationally expensive. Presently, it takes up to 5 days to train 100,000 images running on a NVIDIA GeForce GTX 1080 Ti GPU. In the future a more complex analysis of this issue can be carried out once computing power significantly improves.

[6]http://yann.lecun.com/exdb/mnist/

5.3.3 Clustering and classifying

After clustering by the Bayesian Gaussian mixture model (BGM), we obtain the probability of each image belonging to each cluster. These probabilities are used to calculate the overall probability of each image being a strong lensing system.

With the probability of the n-th image to the k-th cluster, given by P^{kn} and known fractions of lensing and non-lensing images in the k-th cluster, P^k_{len} and P^k_{non}, we are able to calculate the predicted probability of different types, lensing (P^n_{len}) and non-lensing (P^n_{non}) for the n-th image by the formulas:

$$\begin{cases} P^n_{len} = \sum_{k=1}^{K} P^k_{len} \times P^{kn} \\ P^n_{non} = \sum_{k=1}^{K} P^k_{non} \times P^{kn} \end{cases} . \tag{5.7}$$

However, our technique is meant to be unsupervised; therefore, P^k_{len} and P^k_{non} are unknown. Without the label information, the network has no prior knowledge regarding classes of lensing or non-lensing. Therefore, to be able to compare the performance of this work and others, we must involve human classification after the step of the feature learning.

Supervised machine learning methods applied to strong lens finding typically require tens of thousands of labelled images for training. This is of course too large for viable human classification and negates the whole purpose of using machine learning in the first place. Therefore, we propose a vastly streamlined way to calculate the predicted lensing and non-lensing probability for the n-th image by assuming the probability of each type for the k-th cluster through looking at the representative features of each cluster. We assume the lensing probability for the k-th cluster is 1.0, i.e. $P^k_{len} = 1.0$, if the representative features of this cluster have significant lensing features (e.g., Einstein rings, distorted arc, etc) (see the bottom of Fig. 5.5). If the features of this cluster are convincingly non-lensing features (e.g., singly isolated and oval object), the lensing probability of the k-th cluster is set to 0.0, i.e. $P^k_{len} = 0.0$ (see the top of Fig. 5.5). In the condition where it is difficult to classify such as those with multiple objects, the probability is assumed to be 0.5, i.e. $P^k_{len} = 0.5$ (see the middle of Fig. 5.5).

The summation of the lensing and non-lensing probabilities (equation 5.7) may not be 1.0 when using assigned probabilities for clusters because the assigned probabilities cannot accurately represent the distribution of lensing and non-lensing images in each cluster. Therefore, we unify the predicted lensing and non-lensing probabilities as follows: $P^{n'}_{len} = P^n_{len}/(P^n_{len} + P^n_{non})$ and $P^{n'}_{non} = P^n_{non}/(P^n_{len} + P^n_{non})$.

The combination of assigned probabilities within our unsupervised technique promisingly reduces the quantitative effort of human judgement on data labelling whereby experts classify a few images that are grouped based on features rather than derived by a machine using over 10,000 images. The comparison of the results using true fractions and assumed probabilities are discussed in section 5.4.1.

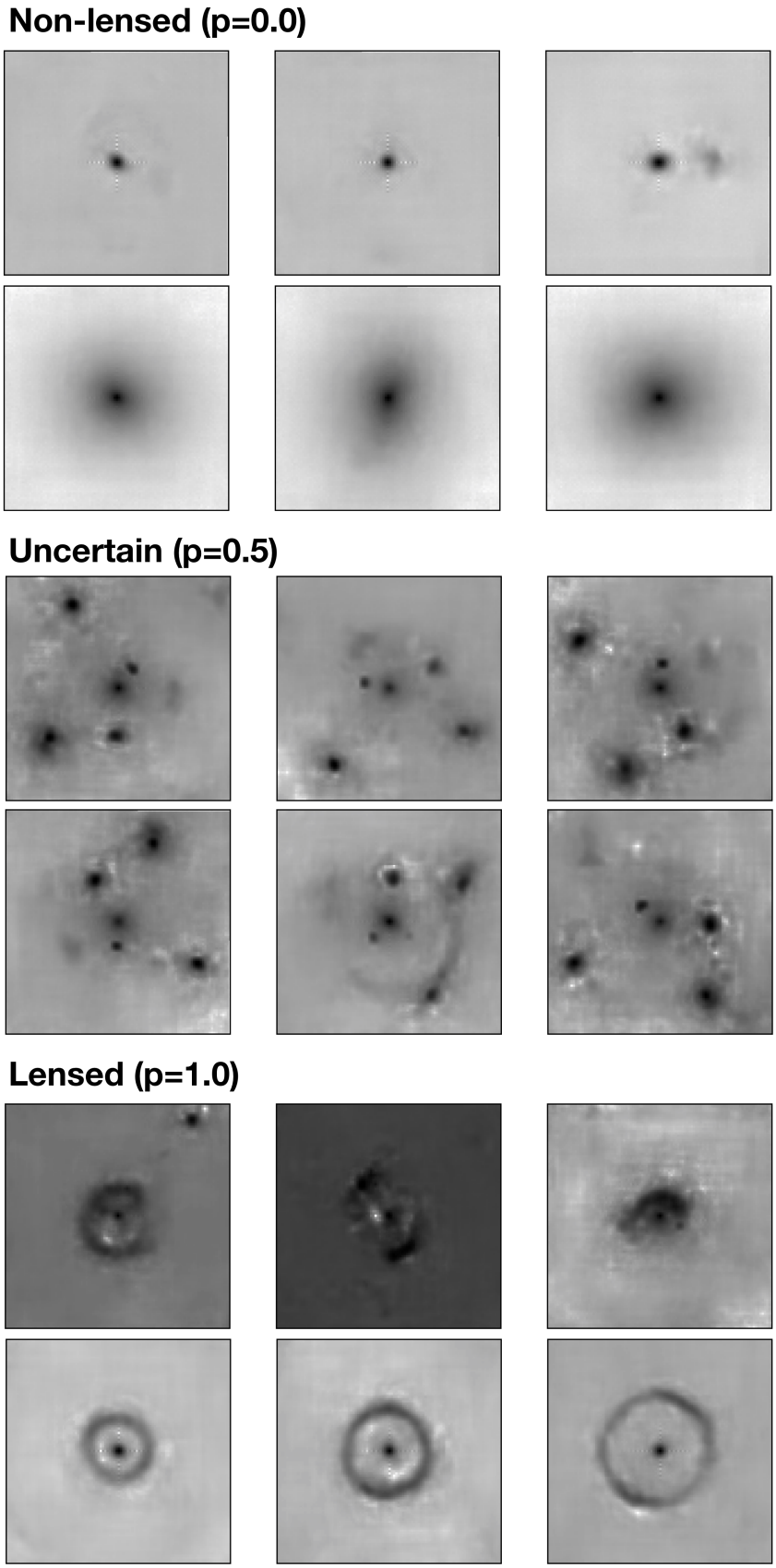

Figure 5.5: Examples of the denoised images from which we assume the lensing probability for clusters. The 'p' value represents the assumed lensing probability for clusters. *Top:* the examples of visually non-lensing images (p=0.0). *Middle:* the uncertain case (p=0.5). *Bottom:* the visually lensing images are presented (p=1.0).

5.3.4 Examinations

With the information on the lensing and non-lensing probability in each cluster, we can compare the performance of our technique with other supervised machine learning techniques using the Receiver Operating Characteristic curve (ROC curve; Fawcett, 2006; Powers, 2011, see details in Section 2.4.1). The definition of the true positive and the false positive are shown in Fig. 2.6 in terms of the confusion matrix. Different from Chapter 2, the '0' means negative as well as non-lensing type while '1' represents positive signal and lensing type in this study. The true positive rate (TPR) and false positive rate (FPR) are defined the same as Equation 2.4.

With the ROC curve, the 'area under the Receiver Operating Characteristic curve' (AUC; Bradley, 1997; Fawcett, 2006, also see Section 2.4.1) is measured to evaluate the performance of machine learning algorithms. In this study, the AUC is used for finding the most optimal number of extracted features within the EL in the CAE. In Fig. 5.6, the black solid line shows the results trained by the images in a logarithmic scale, and the lighter orange dashed line presents the one trained by the images within a linear scale. The lighter shadings show the variation in training defined by the maximum and minimum of three reruns.

Once the CAE model has been trained, the results of the clustering do not change as long as we use the same datasets. Therefore, the main uncertainty in the procedure is from the training process in the CAE. To determine the variation of results using different training we rerun our CAE three times for different numbers of features of the EL within the CAE, and use the maximal and minimal value of the AUC as the uncertainty for each number of features (Fig. 5.6).

We discover that the CAE cannot reproduce the input images if we have an insufficient number of neurons in the EL. However, too many neurons cause overfitting such that the CAE captures noisy features. We find that the highest value of the AUC is carried out from the training by using logarithmically scaled images and the optimal number of neurons in the EL is 24 according to Fig. 5.6. As such, we adopt this set up for all results presented in this work.

Apart from the ROC curve and the AUC value mentioned in section 5.3.2, we also use some other evaluation factors such as recall, precision, f1_score, and accuracy (also see Section 2.4.1), which are measured based on a probability threshold $p = 0.5$. The definition of 'recall' is identical to the TPR in statistics which represents the completeness that shows the fraction of true types correctly identified, while 'precision' indicates the contamination which means the fraction of true types in the list of candidates predicted. The 'f1_score' is a weighted average of the precision and recall which can be interpreted as the overall performance considering the contributions from both completeness and contamination. This is calculated by the formula (Powers, 2011):

$$\mathrm{f1} \quad = \quad 2 \times \frac{(\mathrm{precision} \times \mathrm{recall})}{(\mathrm{precision} + \mathrm{recall})}. \tag{5.8}$$

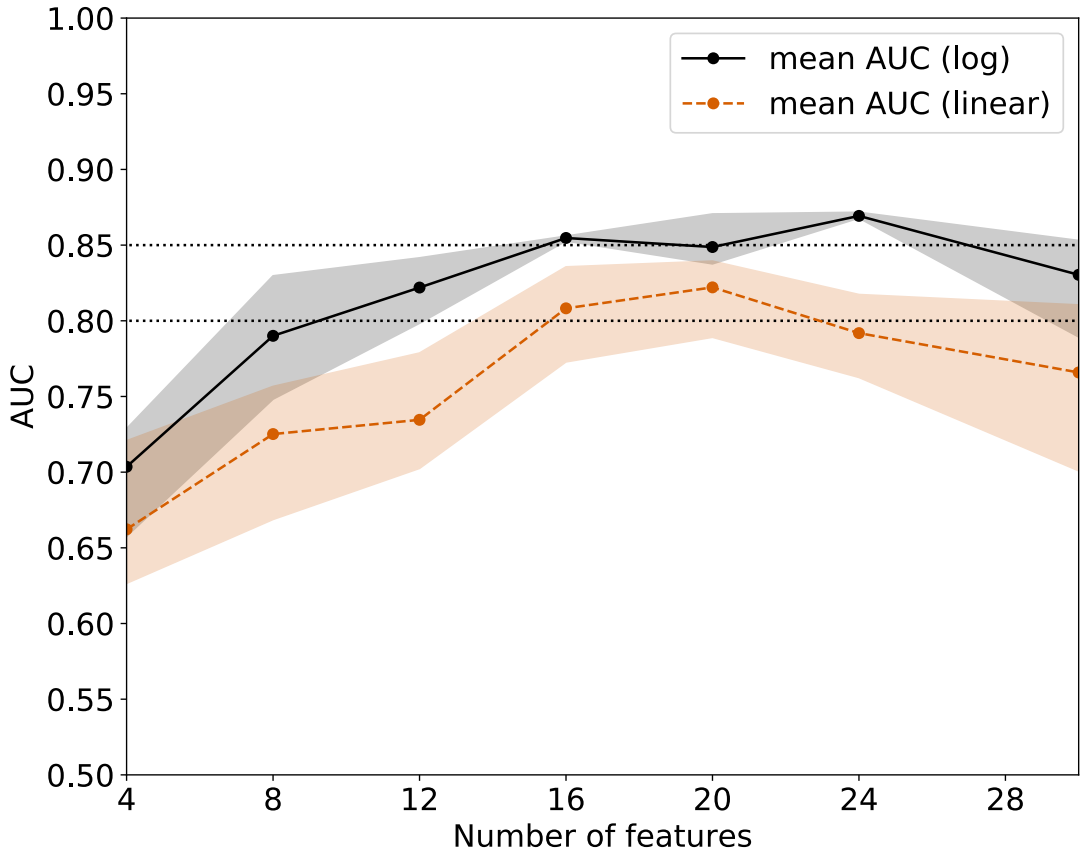

Figure 5.6: The graph of AUC versus the number of extracted features in the CAE (Section 5.2.1). The black solid line represents the mean value of the AUC trained by images with a logarithmic scale, and the orange dashed line is trained by images with a linear scale. The lighter shadings show the variation defined by the maximum and minimum of three reruns. The two dotted lines are locations of AUC = 0.80 and 0.85.

The accuracy is defined as Equation 2.7 such that the meaning of this is defined as how many successfully classified samples there are out of all the samples.

5.4 Results

In this section, we first compare the results using two different calculations of the lensing and non-lensing probabilities for each image (section 5.3.3) in Section 5.4.1. The capability of our unsupervised technique to distinguish different types of lenses, and the performance of classification are presented in Section 5.4.2.1. We also analyse our technique on the testing datasets with different fractions of lensing images; the result of this is shown in section 5.4.2.2. Finally, we revisit the Strong Gravitational Lens Finding Challenge; we present our comparison with other supervised machine learning methods and human inspection in Section 5.4.2.3.

5.4.1 Comparison of Known and Assumed Probabilities

The comparisons of results with a known fraction of lensing and non-lensing images and an assumed probability of lensing (P^k_{len}) and non-lensing (P^k_{non}) in the k-th classification cluster (Section 5.3.3) are shown in Fig. 5.7 using images with logarithmic scale and 24 units in the embedded layer (EL) of the convolutional autoencoder (CAE).

The left panel in Fig. 5.7 presents the Receiver Operating Characteristic curve (ROC curve); the right panel is a comparison of different factors between these two methods such as recall, precision, f1_score, and accuracy. In Fig. 5.7, the black solid line shows the mean value of the ROC curve using a known fraction of lensing images, and the orange dashed line represents the mean value of the results using an assumed probability. The colour shadings represent the variation defined by the maximum and minimum within three reruns.

Although the results of the 'assumed probability' show larger scatter and slightly worse performance than the results of the 'known fraction', the scatter of the 'assumed probability' method is consistent with the results of the 'known fraction' method. Additionally, the mean values of both methods are close to each other. Overall, these two methods show consistent results in their general performance, which is shown through the ROC curve, recall, precision, f1_score, and accuracy (calculated based on a probability threshold of $p = 0.5$).

This comparison confirms that the alternative calculation assigning an assumed probability to the classification clusters can be used to obtain promising lensing and non-lensing probabilities for each image. Furthermore, this indicates that the classification clusters obtained by our technique captures representative features from images and reflects the real lensing fractions in the clusters. Additionally, this result also shows an advantage of our technique for saving effort on data labelling by clustering the data before classifying it so that we can classify the feature of the small number of classification clusters instead of each image itself.

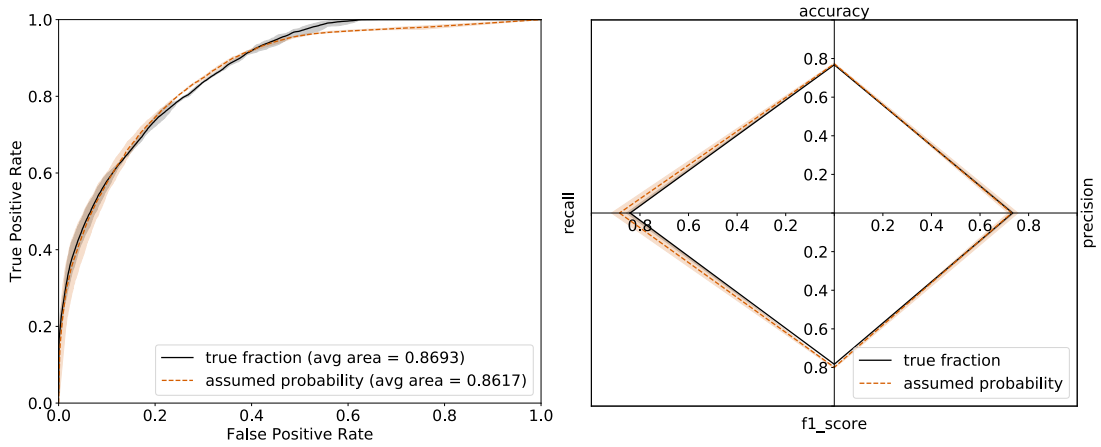

Figure 5.7: The comparison of two methods to obtain the predicted probability of each class for each image using a known fraction and an assumed probability (section 5.3.3). The black solid line represents the mean value using a known fraction, and the orange dashed line shows the mean value using an assumed probability of each class. The colour shadings are the variation defined by the maximum and minimum within three reruns. *Left*: the ROC curve. *Right*: the comparison of different statistic factors, e.g., recall, precision, f1_score, accuracy.

This can be used as a preliminary selection method for future surveys when using a large amount of data.

5.4.2 Identifying Lenses

5.4.2.1 Initial Results

We begin with the results of binary classification using the predicted lensing probability obtained using the 'assumed probability' method in Section 5.3.3. In Fig. 5.8, we present the confusion matrix of the training set. The accuracy of our technique reaches 0.7725 ± 0.0048 and the AUC reaches 0.8617 ± 0.0063 using a probability threshold of $p = 0.5$. The error estimation of the accuracy on the AUC is based on the standard deviation of 3 reruns.

This method promisingly separates features in a way similar to how a human would. Fig. 5.9 shows examples of the classification clusters with a high fraction of lensing images ($\geq$0.6). Every classification cluster shown in Fig. 5.9 has its own characteristic features, which indicates that our technique is able to capture the visual difference and similarity between images. Additionally, these classification clusters with a fraction of $\geq$0.6 contain $\sim$63 percent of lensing objects in the training set. The last row in Fig. 5.9 shows an example of the simulated data without lenses for the classification cluster. It is clear that our technique captures features such as Einstein rings with different radii, different strength, and distorted arc structures, etc, and images without lenses. The classification clusters with significant lensing features such as Einstein rings and arc structures are easily distinguishable (the fraction of lensing images in these groups is $\geq$0.8) in our results.

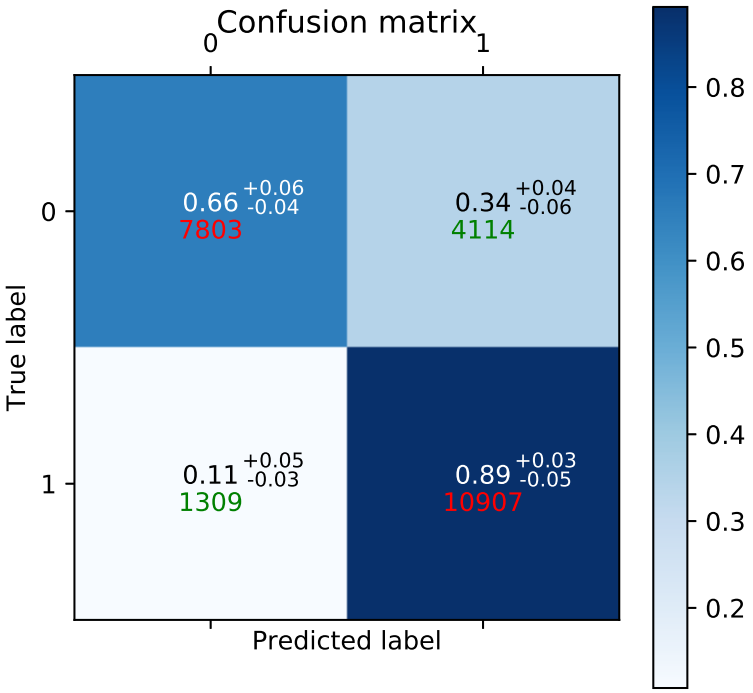

Figure 5.8: The confusion matrix of the training set trained with 24 features in the embedded layer (EL) of the convolutional autoencoder (CAE). The floating values show the mean of the three reruns and the deviation from the maximum and minimum. The red and green texts shown below the fraction are the actual number in the quadrant.

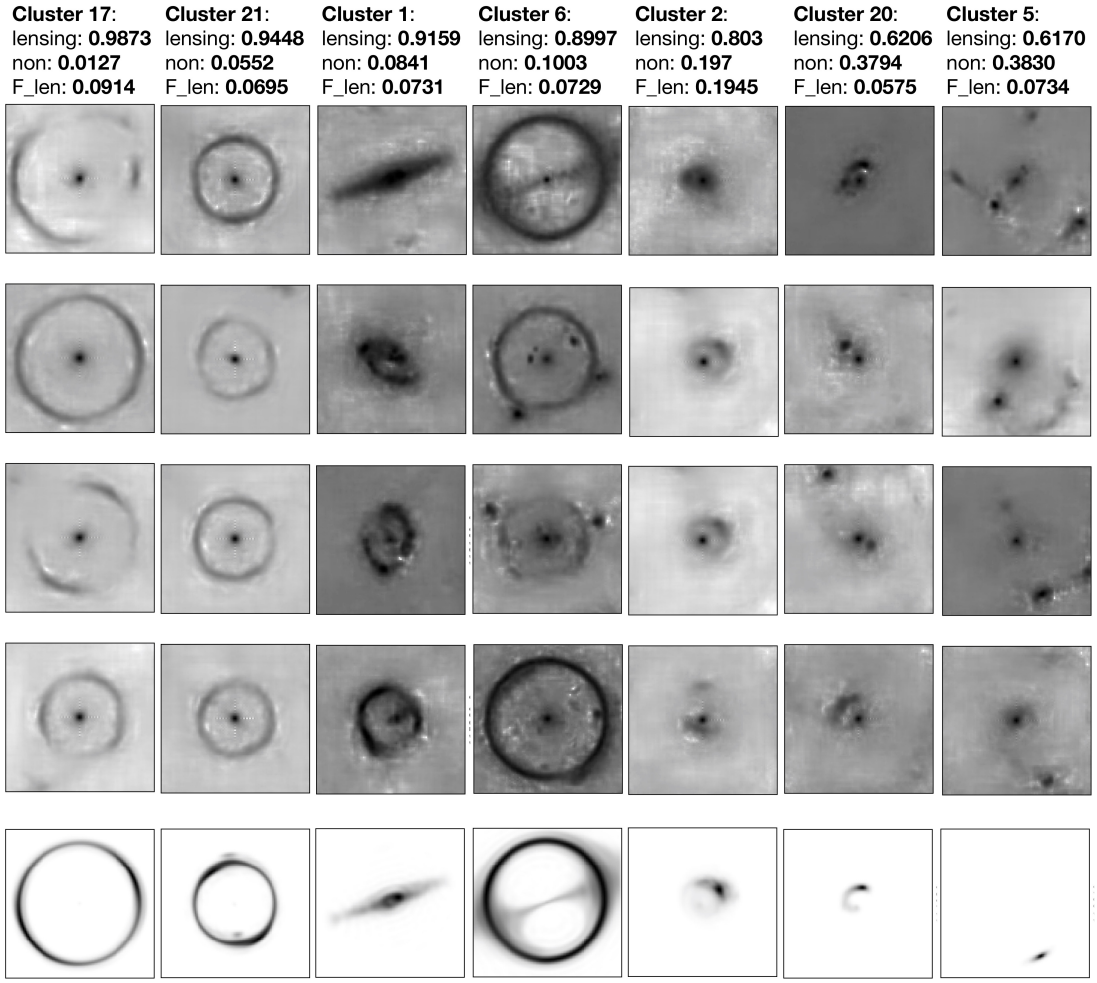

Figure 5.9: Examples of the classification clusters having a high fraction of lensing types in individual clusters (denosied images). The top of each column shows the classification cluster index, the fraction of lensing (lensing) and non-lensing (non) in the cluster, and the fraction of lensing in the cluster of all lensing images in the training set (F_len). The last row shows the simulated data without lenses within each column.

In the same run, there are 7 classification clusters which have a high fraction of non-lensing images ($\geq$0.7); 6 out of 7 clusters include $\geq$0.9 fraction of non-lensing images. The features of these classification clusters are round or oval and isolated objects (Fig. 5.10). The feature of cluster 0 looks oval and isolated, but has a relatively lower fraction of non-lensing images than others. It is produced by visually insignificant arc-like structures in the images that might also be created through the process of denoising.

The last four columns in Fig. 5.10 which contain images with a fraction of non-lensing images between 0.6 and 0.7 are visually multiple objects. It is difficult to distinguish the classification of these types of images without colour information; however, our data is limited to a single visual band (section 5.3.1) so the decrease of performance is unavoidable. Additionally, these four classification clusters are similar to each other, but they are in a different orientation which shows that our technique cannot take care of rotation invariance at the current stage (also see Appendix 5.A and the discussion in section 5.5).

The remaining 6 classification clusters are regarded as uncertain types because the fractions of lensing images in these groups are within the range from 0.4 to 0.6 (Fig. 5.11). Apart from clusters 15 and 23, the features of other classification clusters are single or double objects with filament or arc-like structures which might also be generated by the denoising process. The main features of cluster 15 is a round and single object with lenses surrounded by a halo-like structure, which can occur when the Einstein radius of lensing is equal to or smaller than the size of lenses. On the other hand, cluster 23 has similar features to clusters 9, 13, 18, and 19 which all show multiple object types in the images. As mentioned in the previous paragraph, the images shown in the clusters 15 and 23 cannot be easily distinguished without colour information; therefore their categories are ambiguous.

Overall, it is more challenging to correctly classify images of lensing and non-lensing types without significant lensing features, such as Einstein rings, and highly distorted arc structures seen using our technique with a single band. Our method obtains classification clusters with lensing features containing $\sim$63 per-cent lensed images from all lensed images in the training set (Fig. 5.9). The remaining lensed images are distributed in the classification clusters with diffi-cult features (e.g., the last four columns in Fig. 5.10 and Fig. 5.11).

We anticipate that the inclusion of colour will enhance the performance of this method on the basis that additional diagnostic information would be provided from other surveys with multiple broad-band filters rather than the single Euclid Space Telescope with VIS band.

As part of our investigation, we applied our pre-trained CAE on the simulated data without lenses (central galaxies; Appendix 5.A). Examples are shown in Fig. 5.16 which confirms that the CAE promisingly captures the structure of different lensing types: Einstein rings with different radii, incomplete Einstein

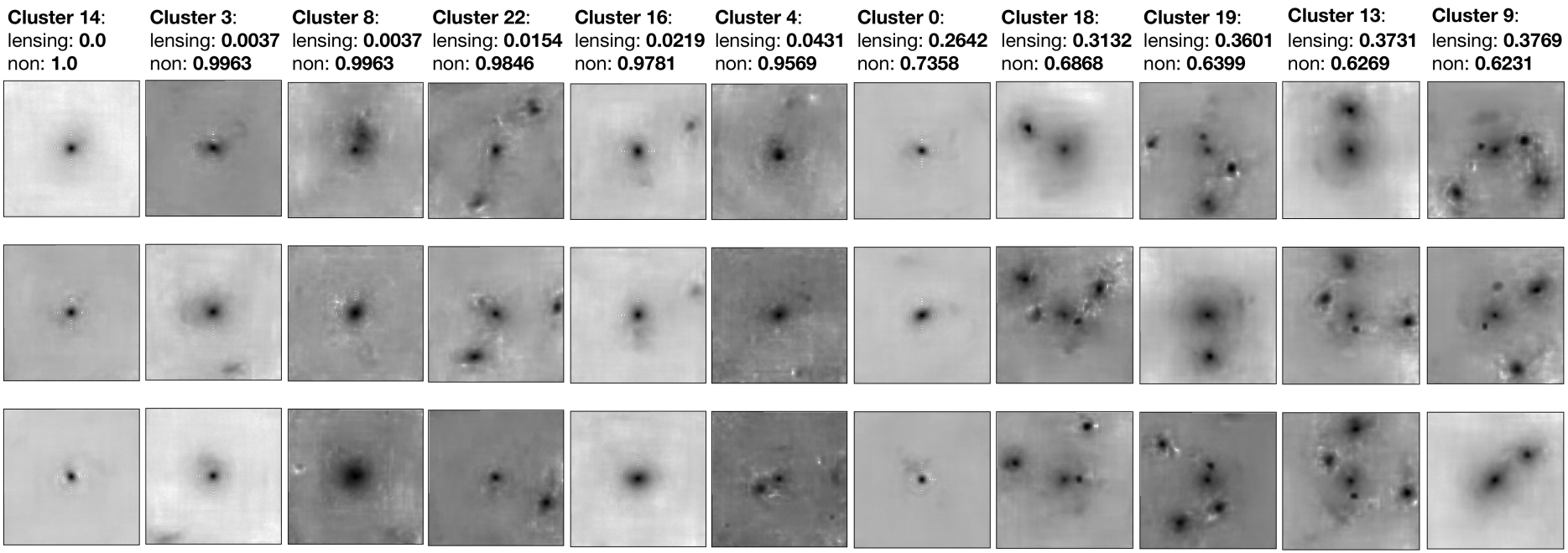

Figure 5.10: Examples of the classification clusters having a high fraction of non-lensing images (denoised images). The top of each column shows the number of the cluster and the fraction of lensing (lensing) and non-lensing (non) in that cluster.

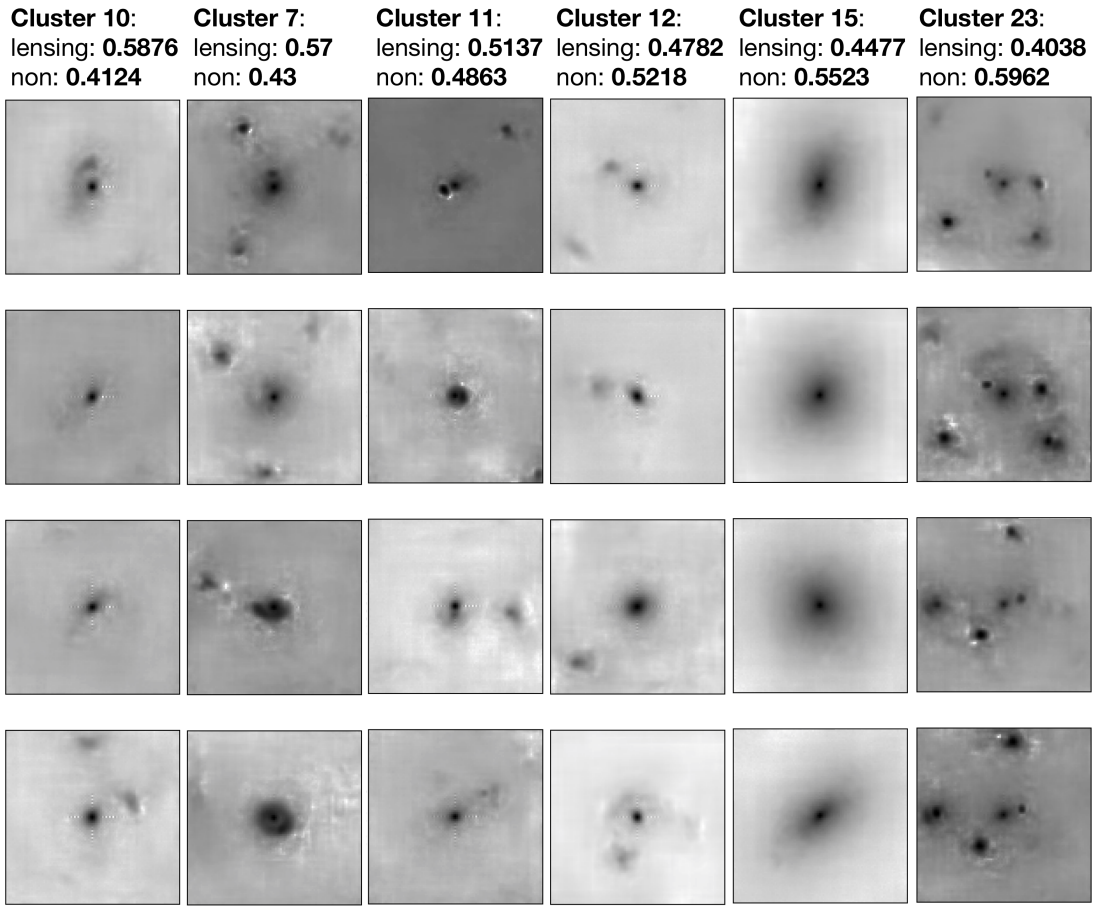

Figure 5.11: Examples of the classification clusters with uncertain classification (denoised images). The top of each column shows the number of the classification cluster and the fraction of lensing (lensing) and non-lensing (non) in the cluster.

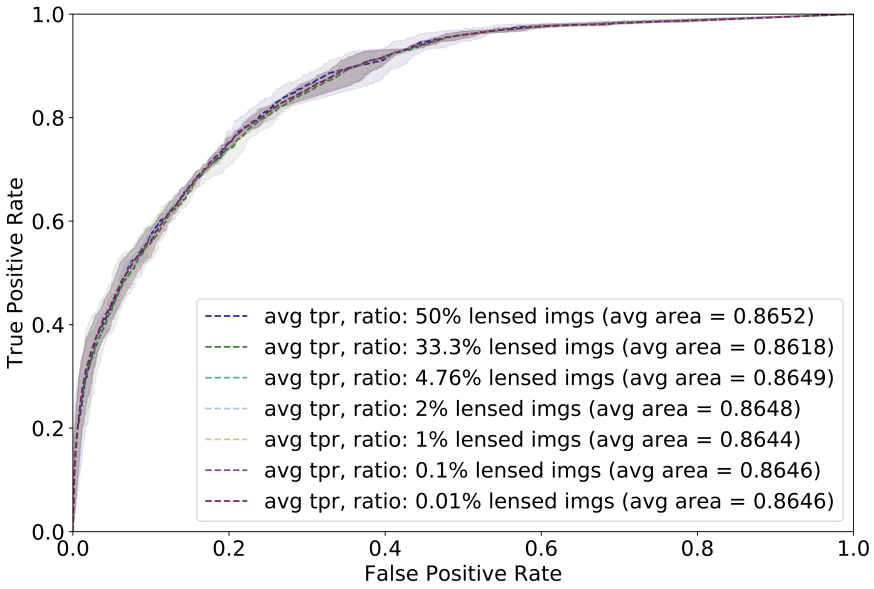

Figure 5.12: The ROC curve of the testing sets using different fractions of lensing images. Different colours represent different fractions (Table 5.1). The dashed lines show the average of the ROC curves within three reruns and the shading areas show the variation.

rings, arc structures with different lengths and positions, extended objects, etc, from these simulated images.

5.4.2.2 Test on datasets with different fractions of lenses

A detectable galaxy-galaxy strong lensing event is an extremely rare event in the universe, e.g., 0.05 percent of 640,000 early type galaxies in the Canada France Hawaii Telescope Legacy Survey are strong galaxy-galaxy lenses (Gavazzi et al., 2014). To be capable of a more realistic case, we test our CAE and pre-trained Bayesian Gaussian mixture model (BGM) on datasets using logarithmic images with different fractions of lensing images from 50 percent to only 0.01 percent of lensing images (Collett, 2015, Table 5.1).

The results are shown in Fig. 5.12. Here we always use the 'assumed probability' to calculate the predicted probability of each type for each image (section 5.3.3). Different colours represent testing sets with different fractions of lensing and non-lensing images. The dashed lines are the average of the ROC curves and the shadings are the variation within three reruns.

Fig. 5.12 clearly shows that there is not a significant difference between the performance of the testing sets with different fractions of lensing images using our technique. Secondly, Fig. 5.13 shows the accuracy of the classification in terms of a confusion matrix using the testing set with 0.01 percent of lensing images; this result is consistent with the results from training (Fig. 5.8).

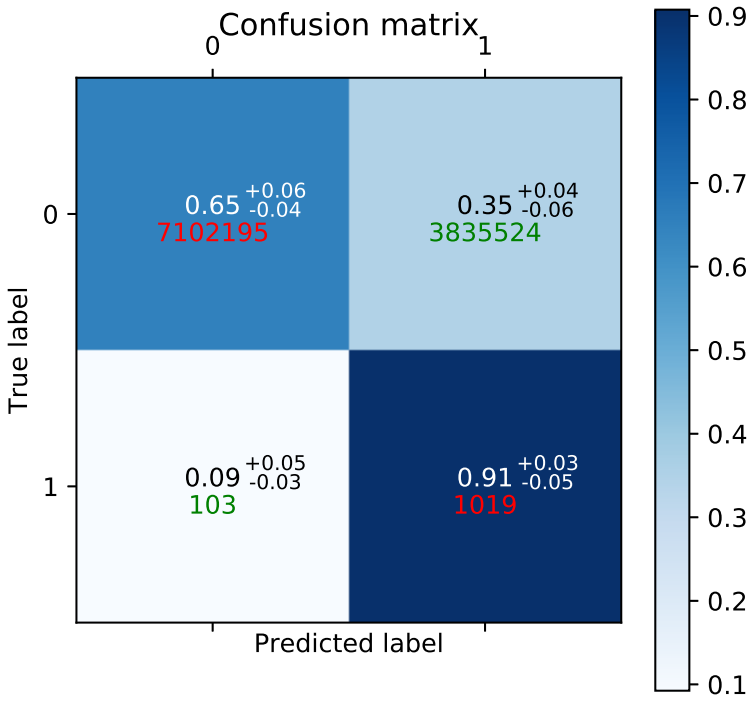

Figure 5.13: The confusion matrix of the testing set containing 0.01 percent lensing images using the pre-trained model with 24 neurons in the embedded layer (EL) of the convolutional autoencoder (CAE). The floating values show the mean of the three reruns and the deviation from the maximum and minimum. The red and green texts shown below the fraction are the actual number in the quadrant.

Both figures show that our unsupervised machine learning technique can maintain its performance even if the lensing events are rare in the data (to 0.01 percent of lensing images) when the model is well pre-trained.

5.4.2.3 Comparison with Other Methods

To further compare the performance of our technique with other supervised machine learning methods and human inspection, we revisit the Strong Gravitational Lens Finding Challenge (Lens Finding Challenge; Metcalf et al., 2019b). The final challenge testing data in the Lens Finding Challenge includes 100,000 images, which are ∼60 percent of non-lensing images and ∼40 percent of lensing images.

A visually detectable lensing feature generally has a high Signal-to-Noise Ratio (SNR) or has a low SNR but a larger number of correlated lensed pixels. Fig. 5.14 shows the comparison of the SNR and the number of lensed pixels above 1σ between the training set and the challenge testing data. The value of the SNR in Fig. 5.14 is calculated by $SNR = \frac{S}{\sigma\sqrt{N}}$, where $\frac{S}{\sigma}$ represents the intensity (flux) in

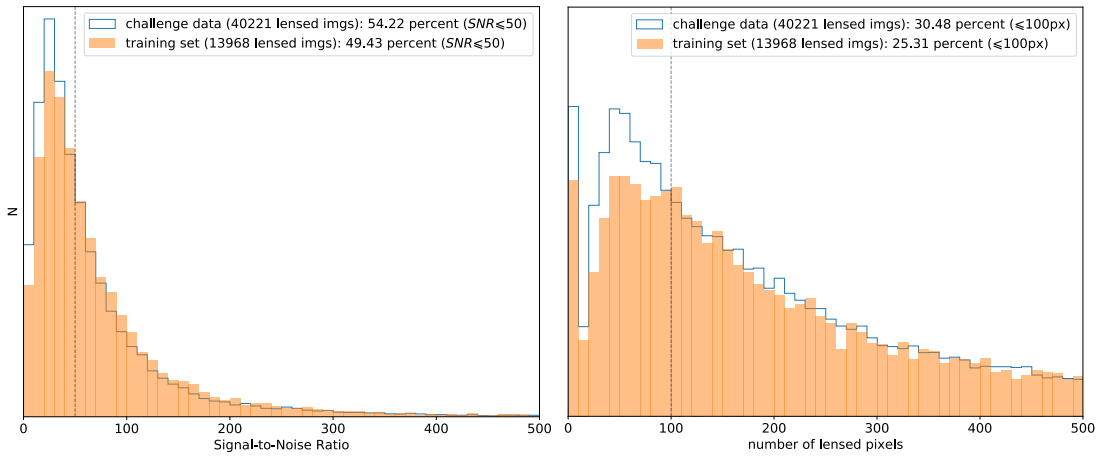

Figure 5.14: The comparison of the Signal-to-Noise Ratios (SNR) and the number of lensed pixels above 1σ comparing the training set and the challenge testing data. *Left:* the comparison of SNR. *Right:* the comparison of the number of lensed pixels above 1σ. The dashed lines represents the divide based on a visual assessment whereby the distribution on the left shows significant inconsistency between the training set and the challenge data set.

a sigma contributed by the N lensed pixels. This figure shows that the fraction of the images that are difficult to visually classify has increased from the training set to this challenge testing data.

In addition to the value of AUC, Metcalf et al. (2019b) apply two other factors: TPR_0 and TPR_{10} to score the performance of their techniques. The TPR_0 is defined as the highest TPR reached when the FPR=0 in the ROC curve. This quantity is used to recognise the classifiers whose highest classification levels are not conservative enough to eliminate all false positives; therefore, the TPR_0 of these classifiers are often equal to 0. The TPR_{10} is defined when TPR at the point where less than ten false positive are made.

We apply the same architecture for the CAE as we do for the training set (Fig. 5.1), followed by the training process shown in section 5.3.2, and the classifying process shown in section 5.3.3 whereby we are applying the 'assumed probability' to this challenge testing data. The results are shown in Table 5.2.

Our unsupervised machine learning technique using a single band is more sensitive to significant lensing features. However, the challenge testing data contains the most visually difficult images with lower SNR and fewer lensed pixels resulting in poorer performance ('Unsupervised technique' in Table 5.2) compared to the training set (labeled as * at the bottom row in Table 5.2).

To fully test our method, we make a cut at 100 pixel and 50 SNR to exclude visually difficult images. This cut is determined by Fig. 5.14 and a visual assessment to the images with these criteria. Applying this cut improves the performance of our technique from $AUC = 0.72$ to $AUC = 0.83$ that indicates that the difference

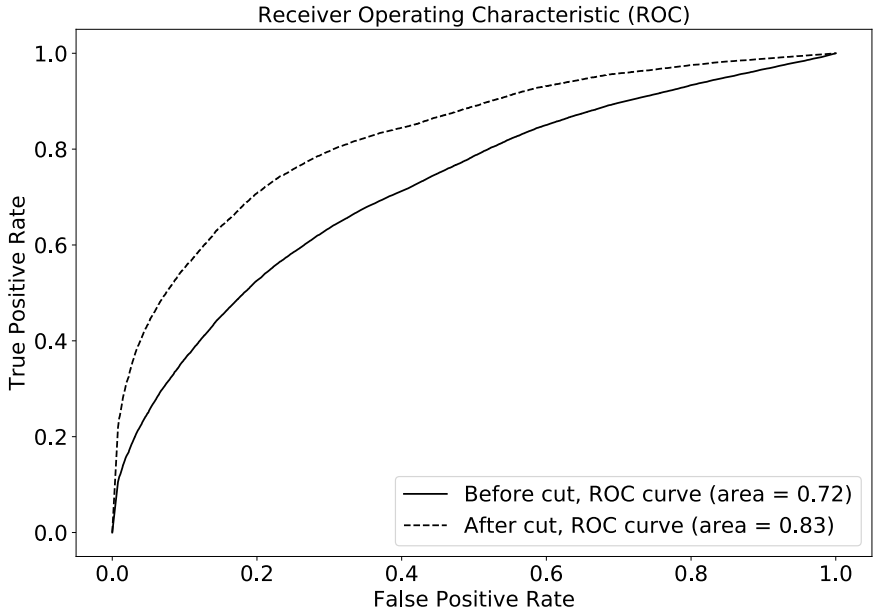

Figure 5.15: The comparison of the ROC curve between before and after a cut at images with sizes greater than 100 lensed pixels and with a Signal-to-Noise Ratio larger than 50.

in performance (i.e. AUC) between the two highlighted entries in Table 5.2 using our method is caused by the difference in the distribution of SNR and lensed pixels between the training and testing data. The comparison between applying the cut and not doing so is shown in Fig. 5.15.

As in most methods, both TPR_0 and TPR_{10} are equal to 0.00 using the challenge testing data in our results. However, in Fig. 5.15, both curves have a nearly vertical line at False Positive Rate $\sim$0 until True Positive Rate $\sim$0.1 (before) and $\sim$0.2 which means that although our technique is not able to eliminate all the misclassifications when the probability threshold is high (left), there are only a tiny number of images which were predicted incorrectly.

This comparison gives an idea for the feasibility of this unsupervised machine learning technique compared with supervised methods. However, unsupervised machine learning is a qualitatively different method than supervised methods, such that unsupervised methods can explore data without label limitations and addresses questions that current supervised methods cannot. Therefore, the performance of unsupervised machine learning methods cannot simply be compared to supervised methods where the true label information is used.

5.5 Future Work

In this chapter, we describe an unsupervised machine learning technique for the detection of galaxy-galaxy strong gravitational lensing using simulated data based on the Euclid Space Telescope from the Strong Gravitational Lens Finding Chal-

Name	Author	AUC	TPR_0	TPR_{10}	short description
LASTRO EPFL	Geiger, Schäfer & Kneib	0.93	0.00	0.08	CNN
CMU-DeepLens-Resnet	Francois Lanusse, Ma, C. Li & Ravanbakhsh	0.92	0.22	0.29	CNN
GAMOCLASS	Huertas-Company, Tuccillo, Velasco-Forero & Decencière	0.92	0.07	0.36	CNN
CMU-DeepLens-Resnet-Voting	Ma, Lanusse & C. Li	0.91	0.00	0.01	CNN
AstrOmatic	Bertin	0.91	0.00	0.01	CNN
CMU-DeepLens-Resnet-aug	Ma, Lanusse, Ravanbakhsh & C. Li	0.91	0.00	0.00	CNN
Kapteyn Resnet	Petrillo, Tortora, Kleijn, Koopmans & Vernardos	0.82	0.00	0.00	CNN
CAST	Bom, Valentín & Makler	0.81	0.07	0.12	CNN
Manchester1	Jackson & Tagore	0.81	0.01	0.17	Human Inspection
Manchester SVM	Hartley & Flamary	0.81	0.03	0.08	SVM / Gabor
NeuralNet2	Davies & Serjeant	0.76	0.00	0.00	CNN / wavelets
YattaLensLite	Sonnenfeld	0.76	0.00	0.00	Arcs / SExtractor
All-now	Avestruz, N. Li & Lightman	0.73	0.05	0.07	edges/gradiants and Logistic Reg.
Unsupervised technique	**This Work (Section 5.4.2.3)**	**0.72**	**0.00**	**0.00**	**Deep Clustering**
GAHEC IRAP	Cabanac	0.66	0.00	0.01	arc finder
***Unsupervised technique**	**This Work (Training, Fig. 5.7)**	**0.87**	**0.08**	**0.08**	**Deep Clustering**

Table 5.2: Edited based on the Table 3 in Metcalf et al. (2019b). The AUC, TPR_0 and TPR_{10} for the entries in order of AUC. The highlighted entry without a * is the result of the challenge testing data (this Section). The bottom row with * shows the result obtained by using the training set (Fig. 5.7), which is used for comparing with the result of the testing data (the highlighted entry above without a *). The difference in AUC using our method between these two entries is due to the difference in the distribution of signal-to-noise ratio and lensed pixels between two datasets (Fig. 5.14).

lenge (Lens Finding Challenge; Metcalf et al., 2019b). This technique uses feature extraction provided by a convolutional autoencoder (CAE) and a Bayesian Gaussian mixture model (BGM) clustering algorithm.

This is an initial step in the use of convolutional autoencoders for astronomical unsupervised learning problems and as such there are many further explorations and improvements for this technique. For instance, there are other types of autoencoders e.g., variational autoencoder (Kingma and Welling, 2013) for feature learning, and other kinds of clustering algorithms to explore the features and the properties of the obtained groups e.g., hierarchical clustering such as Agglomerative Hierarchical Clustering (Bouguettaya et al., 2015) and density-based clustering such as DBSCAN (Ester et al., 1996), etc.

In addition to other approaches that could be taken with different autoencoders and different clustering algorithms, some other future improvements are discussed here. First of all, we use the simulated data with a single VIS band in the optical region for the Euclid Space Telescope from Lens Finding Challenge. As shown in Section 5.4.2.1, the lack of multiple bands causes difficulty in classifying certain types of images (Fig. 5.11). In the future, we will apply our pipeline to surveys with multiple filters, which is expected to improve the performance further.

Secondly, the current state of this technique cannot preserve rotation invariance which means it categorises images differently when we rotate the images (see the last four columns in Fig. 5.10 & Fig. 5.16). This condition does not affect the current results negatively in distinguishing lensing or non-lensing feature. However, considering the rotation invariance may help to reduce the number of classification clusters we obtain from this method when applying this technique on real data.

On the other hand, using an alternative autoencoder, the 'variational autoencoder' (Kingma and Welling, 2013) which applies Gaussian distributions to map the extracted features of each images is another potential approach to solve the issue of this rotation variance of clustering results. Preservation of rotation invariance in this way will be left for future work.

Thirdly, in our Appendix 5.A, we show a perfect separation between lensing and non-lensing using the simulated data without lenses (i.e. central galaxies) within our technique. Although it is an unrealistic result considering we cannot perfectly deblend lenses and sources in real data, it is an indication of the improvement we might see without lenses through a pre-processing procedure of removing central galaxies.

One of the main issues of this technique is that we need a certain amount of data with strong features (e.g., lensed images, merger events, feature galaxies, etc) to let a CAE capture a variety of features from these objects. If the data with strong features is rare, the CAE would fail to capture the features and reproduce an inaccurate image.

The galaxy-galaxy strong lensing systems are relatively rare events in the universe. We have therefore had to use an amount of simulated data to train on. This situation could be potentially improved upon by further modification of the CAE architecture and possible data pre-processing. However, this technique is likely suitable for the astronomical objects with a relatively balanced distribution of features, such as the classification of galaxy morphology (Chapter 6). However, few-shot learning (Li et al., 2006) can be used when the labelled data is very limited. This could be one direction for improving the issue of having an extremely imbalanced data set within strong lensing detection scenarios.

On the other hand, the true power of an unsupervised machine learning technique is to find the hidden patterns or unrevealed characteristics in imaging data rather than just improving the efficiency or the performance for a known classification. To reveal the power of this unsupervised technique, we need to reconsider the selection method to determine the optimal number of the neurons in the embedded layer (EL) of the CAE to replace the value of AUC (Fig. 5.6) in the future. Additionally, a forecast for the minimum number of features needed when using real observed data will be investigated in future work by improving the quality of the simulations and by adding more categories with realistic contamination. The ultimate determination for the optimal number of extracted features is also crucial for future usage when applying this unsupervised technique to observed data.

5.6 Conclusion

The purpose of this chapter is to introduce an unsupervised machine learning technique that differs considerably from previous related works on the application to astronomical data. The unsupervised machine learning technique adopted in this study is composed of the feature extraction by a convolutional autoencoder (CAE) and a clustering algorithm - a Bayesian Gaussian mixture model (BGM). We go beyond previous unsupervised work such as Hocking et al. (2018) and Martin et al. (2019) who applied Self-Organised Map (neural networks; Kohonen, 1997) and hierarchical clustering to carry out feature extraction and clustering, respectively.

We use the spaced-based simulated data from the Euclid Space telescope with a visual band (VIS) from the Strong Gravitational Lenses Finding Challenge (Lens Finding Challenge; Metcalf et al., 2019b) and revisit this challenge. To compare our result with other lens-finding approaches, we propose a simple way to calculate the predicted probability of an image to be within each type - lensing and non-lensing by classifying the features of each cluster (Section 5.3.3). This method, which promises to save an extensive effort need for data labelling in supervised machine learning, reaches an AUC value of 0.8617 ± 0.0063 and an accuracy of 0.7725 ± 0.0048 on the classification of galaxy-galaxy strong lensing events using the training set of the space-based survey from the Lens Finding Challenge.

The main accomplishment of this study is that our technique captures meaningful features which follow human visual assessment from images without any initial label information. Additionally, this technique distinguishes a variety of lensing types (e.g., Einstein rings with different radii, different appearance of arcs) (Fig. 5.9 & Fig. 5.16) and potentially can detect unusual lensing features. The discriminating ability is highlighted in Appendix 5.A using a pre-trained CAE model on the simulated data without lenses.

We then revisit the Lens Finding Challenge by applying our technique on their challenge testing data (section 5.4.2.3). The results show a degradation in performance from the training set to the challenge testing data which is due to the difference in the distribution of the Signal-to-Noise Ratios (SNR) and the number of lensed pixels above 1σ in the lensed images in the challenge testing data. Therefore, we applied a cut at 100 pixels and 50 SNR to the challenge testing data, with the results shown in Fig. 5.14. As can be seen, by removing these systems we improve the performance of our technique.

Another advantage of our technique is that it also retains its discriminating ability when the fraction of lensing images varies. As is shown in Section 5.4.2.2, the performance is consistent for the cases of the data holding $\sim$0.01 percent or $\sim$50 percent of lensing images, once the unsupervised model is well pre-trained.

The most promising advantage of this technique is the pre-selection in the process of searching for strong lenses in upcoming large scale imaging surveys. It reduces the sample size of the dataset needed for the classification by cleaning up apparent non-lensing systems. Also, our approach can identify rare lensing systems with unusual characteristics such as multiple Einstein Rings, which can be identified as non-lenses with a high probability by supervised finders if the training sets do not contain these features.

In the future, as discussed in Section 5.5, we will try to improve the competitiveness of our approach by adopting different architectures of neural networks, alternative autoencoders or clustering algorithms. Combining unsupervised and supervised techniques is another direction we plan for increasing the performance of the identification of strong lenses. Finally, the development of a quantitative validation tool for unsupervised machine learning techniques such as the Receiver Operating Characteristic curve (ROC curve) for supervised machine learning techniques is of great importance for future work. Without such diagnostics, it is not possible to objectively compare unsupervised machine learning approaches.

5.A A Test on Simulated Data without Lenses

As part of our investigation, we test our pre-trained convolutional autoencoder (CAE; section 5.3.2) on our simulated data without lenses (i.e. central galaxies) in this study. The result is shown in Fig. 5.16. The purpose of this test is to

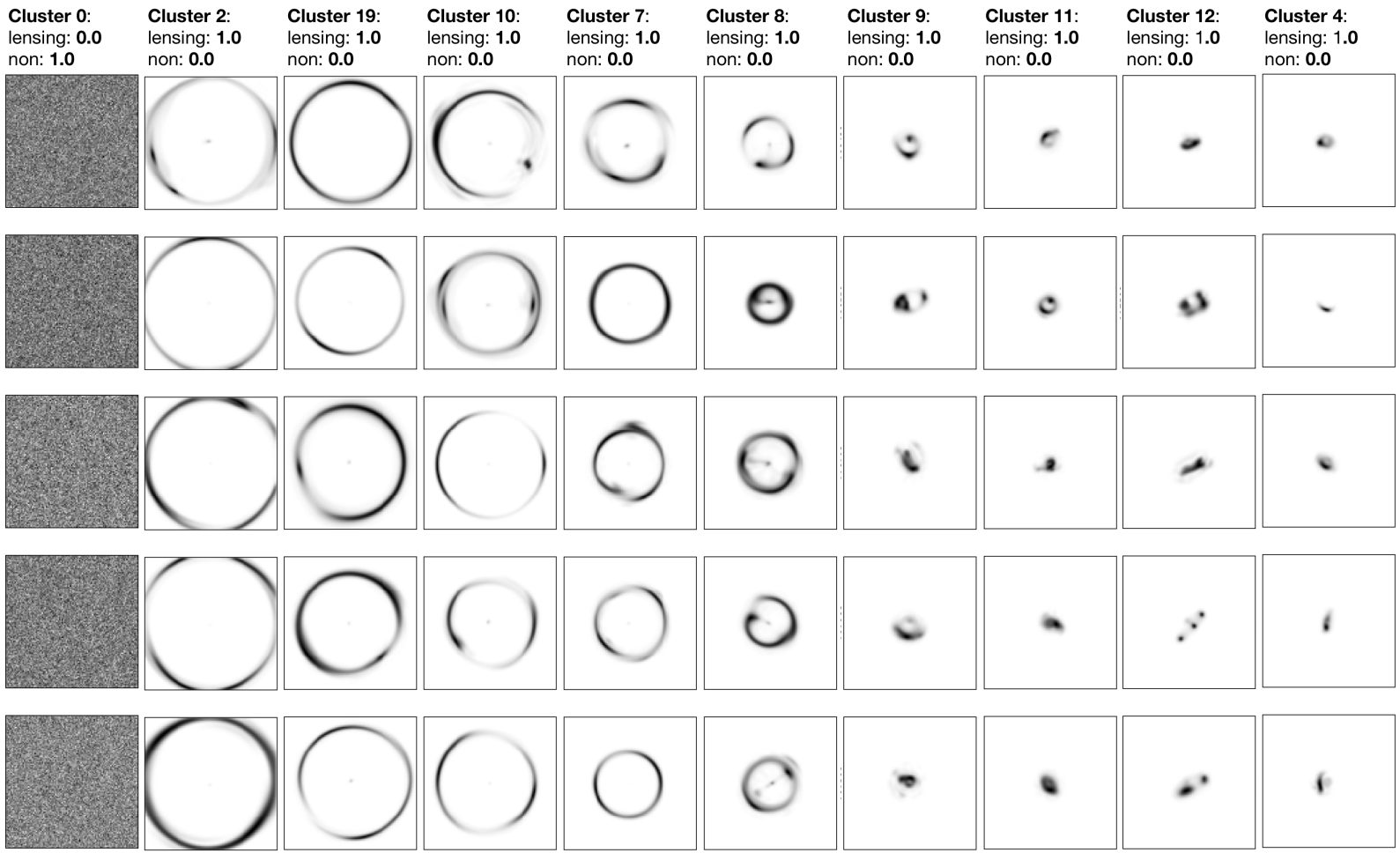

Figure 5.16: Examples of classification clusters using the simulated data without lenses (central galaxies). The top of each column shows the number of the cluster and the fraction of lensing (lensing) and non-lensing (non) in the cluster. The figure is continued in Fig. 5.17.

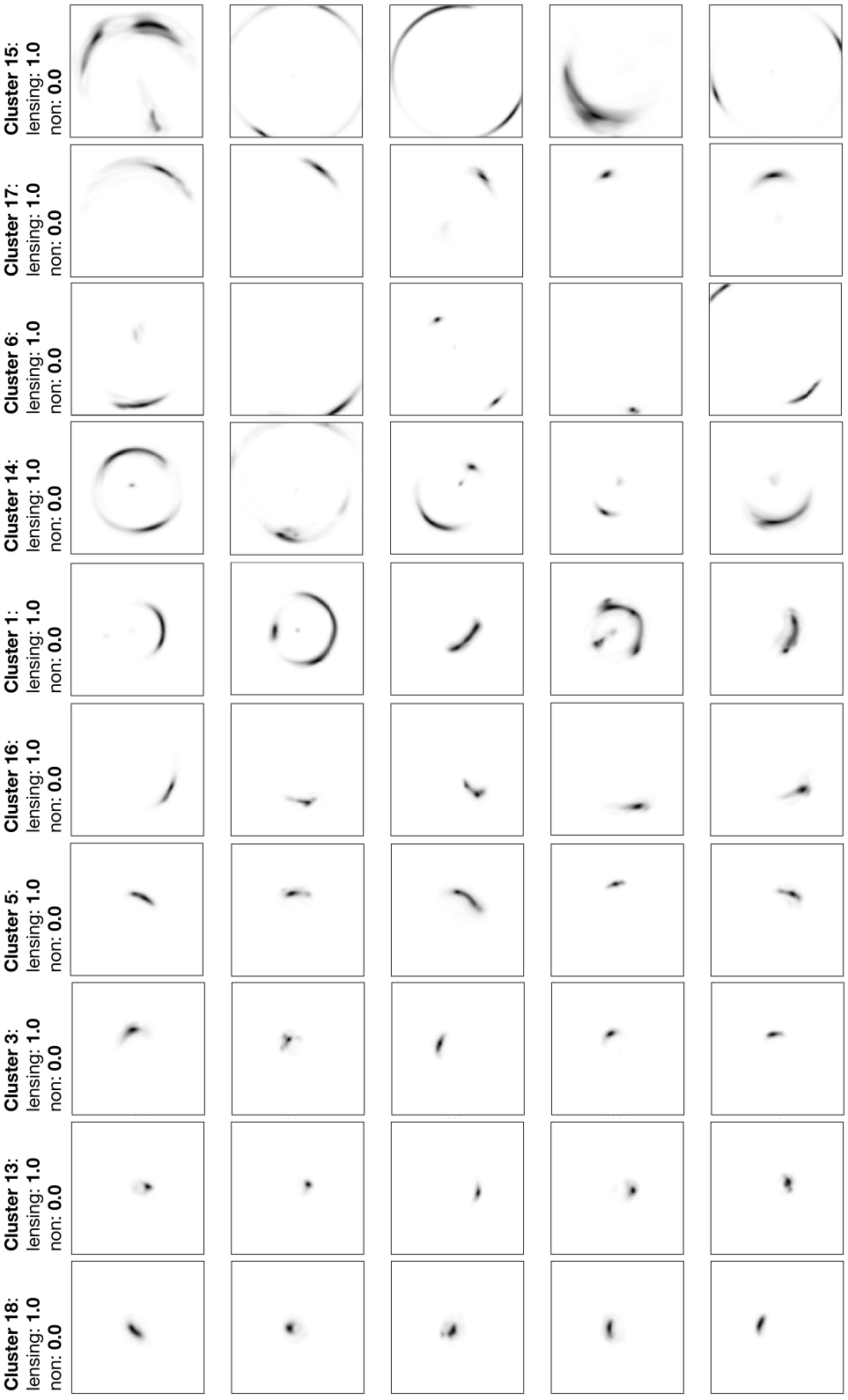

Figure 5.17: The continued figure of Fig. 5.16.

explore the potential usefulness for this technique when deblending of the lenses from the sources is possible.

The simulated data we used is the training set from the Strong Gravitational Lenses Finding Challenge (Lens Finding Challenge; Metcalf et al., 2019b). This challenge offered participants images with all possible image types (lenses, sources, and background noise), images with lenses only, and images with sources only. The simulated data without lenses (central galaxy, i.e. with source only) emphasizes the features of the images, thus, we use the pre-trained model trained by images with linear scale using 20 features (Fig. 5.6) in the embedded layer (EL) of the CAE.

The result reconfirms our results in section 5.4.2.1. We ordered the clusters based on the appearance of the images in the cluster in Fig. 5.16 such that it is easier to see the trend. Above the first row in Fig. 5.16 shows the cluster ID and the fraction of both lensing (lensing) and non-lensing (non) in the cluster.

The first column (cluster) contains all the non-lensing images, which are shown as empty images when there are no lenses in the images. From the second to the eighth column in Fig. 5.16 show the structure of Einstein rings with different radii and from the ninth column in Fig. 5.16 to Fig. 5.17 show the arcs structure with different features such as positions, lengths, or the radii of arcs.

We also reconfirm that the rotation invariance cannot be preserved using our current technique (the last four columns of Fig. 5.10 in section 5.4.2.1). The characteristic of the CAE is to minimize the difference between input and output images; therefore, arcs with similar radii and lengths but located at different positions are identified as different clusters by our unsupervised technique at the current stage. Although this rotation variant has no significant effect on the final result, the improvement on considering rotation invariance might be helpful to reduce the complexity of extracted features when applying this technique to real data.

Additionally, the lensing and non-lensing images are perfectly separated in this test. Although it is unrealistic, we might be able to significantly improve the performance and strengthen the usefulness of this technique by approaching the condition of the images in this test through a pre-processing procedure of removing central galaxies which is possible.

Chapter 6

Beyond the Hubble Sequence - Exploring Galaxy Morphology with Unsupervised Machine Learning

This chapter is based on unpublished material by ***Ting-Yun Cheng****, under the supervision of Marc Huertas-Company, Christopher J. Conselice, and Alfonso Aragón-Salamanca.*

Abstract

In this chapter, we apply an unsupervised machine learning technique composed of a feature extractor with a vector-quantised variational autoencoder (VQ-VAE) and a hierarchical clustering algorithm (HC) to explore unsupervised deep learning classifications of galaxy morphology. We propose a new methodology including: (1) consideration of the clustering performance simultaneously when learning features from images; (2) to allow different distance thresholds used in the HC algorithm; (3) to transform the feature of galaxy orientation in the dataset into a cut to determine the number of clusters. This setup provides 27 clusters which are separated based on galaxy shape and structure (e.g., Sérsic index, concentration, asymmetry, Gini coefficient). The given clusters are well correlated with physical properties such as the colour-magnitude diagram, and show an evolution of the mass-size relations between different machine-defined galaxy morphologies. When we merge the given clusters into two preliminary clusters to provide a binary classification, an accuracy of $\sim 87\%$ is reached using the imbalanced dataset which includes 22.7% early-type galaxies and 77.3% late-type galaxies. Comparing the given clusters with the Hubble types (ellipticals, lenticulars, early spirals, late spirals, and irregulars), we conclude an intrinsic vagueness existed in visual classification systems, in particular galaxies with transitional features such as lenticulars and early spirals. Based on this, the main result in this work is not how well the unsupervised method can match visual classification, but that the method provides an independent classification that may be more physically meaningful than the visual one.

6.1 Introduction

As introduced in Section 1.3, galaxy structure and visual morphology display a strong connection with their stellar population properties, such as surface brightness, colour, and the formation history of galaxies (Holmberg, 1958; Dressler, 1980). The dominant visual morphological classification system in use today was first constructed by Hubble (1926, 1936, Fig. 1.1). Since then, a number of detailed classification systems were proposed such as ones including the notation for the inter and outer ring structure (de Vaucouleurs, 1959) and different arm classes (Elmegreen and Elmegreen, 1982, 1987), among others.

However, visual classification systems can be intrinsically biased due to the subjective judgement of different human classifiers. These human errors are unavoidable and sometimes cannot be reproduced for carrying out a statistical analysis. This greatly limits the ability to use galaxy classification in a formal quantitative way. These issues led astronomers to search for a quantitative description of galaxy structure based on the shape, structure, and physical properties of galaxies which can in principle be connected with visual morphology. For example, the Principal Component Analysis (PCA) was applied to determine the number of dominant features to reproduce the variance shown in observation in Whitmore (1984) as well as to provide an objective procedure for analysing galaxy properties (also see Conselice, 2006). Other studies such as non-parametric methods, e.g., concentration, asymmetry, smoothness/clumpiness, and gini coefficient (Conselice et al., 2000; Bershady et al., 2000; Abraham et al., 2003; Conselice, 2003; Lotz et al., 2004; Law et al., 2007), and parametric methods, e.g., Sérsic profile (Sérsic, 1963, 1968) for measuring galaxy structure were also proposed to provide a more objective and quantitative classification systems than visual assessment alone.

Even though quantitative measures of galaxy structure are extremely useful for measuring properties such as the merger history (e.g., Conselice, 2003), morphological 'classifications' into types is still an important and complementary process. However, it is not clear if indeed we know what these best 'types' are. Thus, in this study we build a galaxy morphological classification system that does not involve human bias through a machine learning approach. For this purpose, we use unsupervised machine learning which is trained without any prior knowledge (e.g., galaxy labels, such as Hubble types). This approach is able to give us the classifications from the machine's perspective based upon input features. However, with an unsupervised machine learning technique it becomes more challenging to have a 'sensible' classification, that is one with more consistency with human opinion, when the dimensionality of a feature space becomes high (curse of dimensionality, Bellman, 1954; Keogh and Mueen, 2017). In astronomical studies, unsupervised machine learning applications have been mostly used in the studies of spectroscopic data which is less dimensional than applying to imaging data (e.g., Geach, 2012; Krone-Martins and Moitinho, 2014; Carrasco Kind and Brunner, 2014; Siudek et al., 2018a).

There are currently several types of astronomical studies that apply unsupervised machine learning techniques to images which reach reasonable results, including: galaxy morphology (Hocking et al., 2018; Martin et al., 2019), strong lensing identification (Cheng et al., 2020b, Chapter 5), and anomaly detection (Xiong et al., 2018; Margalef-Bentabol et al., 2020). For example, Hocking et al. (2018) and Martin et al. (2019) apply a technique called Growing Neural Gas algorithm (Fritzke, 1994), which is a type of Self-organising Maps (SOMs, Kohonen, 1997), to extract features from images. These features are then connected with a hierarchical clustering algorithm (Hastie et al., 2009). On the other hand, Cheng et al. (2020b, Chapter 5) use a fundamentally different approach by using a convolutional autoencoder (Masci et al., 2011), which includes an architecture of convolutional neural networks, for feature extraction. This method connects the extracted features with a Bayesian Gaussian mixture model from which a clustering analysis can be done.

In this study, we apply an architecture consisting of a convolutional autoencoder, considering convolutional neural networks have demonstrated their capability for capturing representative and meaningful features from images (Krizhevsky et al., 2012). We do not use the same convolutional autoencoder as Cheng et al. (2020b, Chapter 5), but we apply a newly developed technique by Google DeepMind (van den Oord et al., 2017; Razavi et al., 2019) called 'Vector-Quantised Variational Autoencoder (VQ-VAE)'. This technique includes a vector quantisation method that accelerates the time-consuming process of feature extraction when using a convolutional autoencoder, as explained in Cheng et al. (2020b, Chapter 5). On the other hand, for clustering algorithms, we decide to apply a modified hierarchical clustering method to group the data (see details in Section 6.2) in order to explore connections between the distances amongst extracted features in feature space, and the number of classification clusters.

In this chapter, we use this unsupervised machine learning technique to develop a galaxy morphology classification system defined by a machine, and compare it with traditional visual classification system such as the Hubble sequence. We furthermore also compare our machine developed classification with galaxy physical properties, such as stellar mass, colour, and physical size of galaxies. We use monochromatic images throughout to focus only on the impact of galaxy shape and structure on morphological classifications in this chapter. The methodology we develop is introduced in Section 6.2, while the detailed description of how to approach using our method and the data used in this study are shown in Section 6.3. Section 6.4 presents the results in this study. Finally, we conclude the work in Section 6.5.

6.2 Methodology

In this section we explain our unsupervised machine learning methodology that is used throughout this chapter. We give a brief overview here, before going into detail in the following subsections.

Our unsupervised machine learning technique includes a feature learning phase with a vector-quantised variational autoencoder (VQ-VAE; Section 6.2.1 and Section 6.2.2) and a clustering phase using a hierarchical clustering algorithm (HC; Section 6.2.3). Several novel approaches for unsupervised machine learning applications are made in this work: (1) the VQ-VAE considers both reconstruction and preliminary clustering results in the feature learning phase (Section 6.2.2 and also see Section 6.3.3); (2) multiple different distance thresholds are used to draw the decision lines on the merger tree in the clustering process (see details in Section 6.2.3).

6.2.1 Vector-Quantised Variational Autoencoder (VQ-VAE)

The vector-quantised variational autoencoder (hereafter, VQ-VAE) was built by Google DeepMind (van den Oord et al., 2017; Razavi et al., 2019) and was originally used for high-fidelity image emulation. The task of image emulation is to learn the distribution of the data given a set of training images, and then to reproduce the images with the learnt distribution. In details, the structure of an autoencoder (Fig. 6.1) contains an encoder with a posterior distribution $q(z|x)$ and a prior distribution $p(z)$ where x is the input data and z represents latent variable, and a decoder with a distribution $p(x|z)$ for reproducing the input data.

The VQ-VAE is a type of autoencoder which includes the structure of convolutional neural networks and applies a vector quantisation process (van den Oord et al., 2017) to make the posterior and prior distribution become categorical. By using a categorical distribution, the computational time for training an autoencoder is significantly reduced compared to other machine learning methods. For example, in Cheng et al. (2020b, Chapter 5), it takes up to 5 days to train 100,000 images by a convolutional autoencoder running on a NVIDIA GeForce GTX 1080 Ti GPU, while a VQ-VAE takes up to a few hours to train the same amount of data with the same device. This is an enormous difference and shows the power of the VQ-VAE method.

Following the top coloured area in Fig. 6.1, the posterior categorical distribution $q(z|x)$ is defined as (van den Oord et al., 2017; Razavi et al., 2019):

$$q(z=k|x) = \begin{cases} 1 & for \quad k = \text{argmin}_j \|z_e(x) - e_j\|_2 \\ 0 & otherwise \end{cases}, \tag{6.1}$$

where $z_e(x)$ is the output of the encoder (the blue part at the left in the figure), the value e_j represents a vector in the codebook which is used for vector-quantising the $z_e(x)$, and k is the index for the vector used in the selected codebook (the top box of the yellow part in the figure). We then measure the vector-quantised representation $z_q(x)$, which is the input of the decoder (the blue shading at the right side in the figure), through Equations 6.1 and 6.2.

$$z_q(x) = e_k, \quad \text{where} \quad k = \text{argmin}_j \|z_e(x) - e_j\|_2. \tag{6.2}$$

The vector quantisation process is shown as the yellow part in Fig. 6.1. The output of an encoder, $z_e(x)$ can be represented by a combination of the index of different vectors, k, in the codebook (the square in the middle of the yellow part). For example, in Fig. 6.1, a three dimensional 'pixel' in the output of an encoder is represented by a vector, e_3, after the vector quantisation. We then use the index of these vectors to build a two dimensional index map. For the pixel used in our example the value is 3. With this index map, we can rebuild the distribution, $z_q(x)$, with the same dimension as $z_e(x)$ but in this case each 'pixel' in $z_q(x)$ is quantised to one of the vectors shown in the codebook. For our example, the vector e_3 is used for the pixel. The distribution of $z_q(x)$ is then used as the input for the decoder to reconstruct the images.

The loss function of the original VQ-VAE contains three parts: reconstructed loss, codebook loss, and commitment loss. An additional penalty is considered later in the modified version of the VQ-VAE (see Section 6.2.2). The reconstructed loss is measured by comparing the reconstructed images with the input images. The codebook loss is used to make the selected codebook, e_j, approach the output of the encoder, $z_e(x)$, while the commitment loss is applied to encourage the $z_e(x)$ to be as close as possible to the chosen codebook from the previous epoch. With these definitions, the loss function, L, for the VQ-VAE is defined as (Razavi et al., 2019):

$$L = \log p\left(x|z_q(x)\right) + \|sg\left[z_e(x)\right] - e\|_2^2 + \beta\|z_e(x) - sg\left[e\right]\|_2^2, \tag{6.3}$$

where the value sg is the stopgradient operator and β is used for adjusting the weight of the commitment loss. The study of van den Oord et al. (2017) found that these results correlate with the value of β, and no apparent change occurs when β ranges from 0.1 to 2.0. Therefore, we set $\beta = 0.25$ in this study which follows the setting in van den Oord et al. (2017).

The details of the VQ-VAE architecture is shown in Table 6.1. Four convolutional layers are used in both the encoder and decoder, and residual neural networks (ResNets, He et al., 2016) are used in this architecture to create a deeper neural network with less complexity. The activation function applied in the convolutional layers is the Rectified Linear Unit (`ReLu`) (Nair and Hinton, 2010) such that $f(z) = 0$ if $z < 0$ while $f(z) = z$ if $z \geq 0$. The VQ-VAE code is based upon the example provided in SONNET library (DeepMind, 2018)[1] which is built on top of TensorFlow (Abadi et al., 2015b)[2]. To train the VQ-VAE, we apply the `Adam` Optimiser (Kingma and Ba, 2014) and the learning rate is set to 0.0003 which is used in Razavi et al. (2019).

6.2.2 Modified VQ-VAE

In this study, we apply a modification to our original VQ-VAE to consider both image reconstruction and a preliminary clustering result when extracting the representative features from images (Fig. 6.1). To achieve this goal, a penalty

[1] https://github.com/deepmind/sonnet
[2] https://www.tensorflow.org

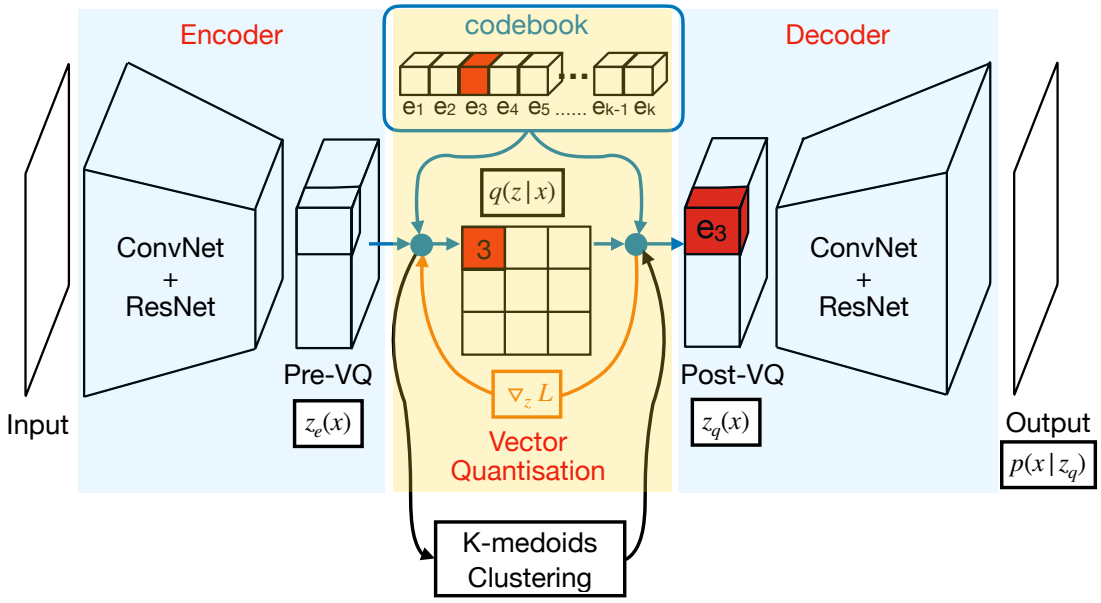

Figure 6.1: The schematic architecture of the **modified** VQ-VAE used for feature extraction of images. The top aspect with a coloured background is the main architecture of the VQ-VAE, which is then modified to consider the silhouette score calculated using the two preliminary clusters given by k-medoids clustering as a part of the loss function when training VQ-VAE (see details in Section 6.2.2). The blue shading at the left and right represents the encoder and the decoder, respectively while the yellow part shows the vector quantisation process. The details of each layer are shown in Table 6.1

Type	#channel	kernel size	stride size	activation function
		Encoder		
Conv2D_1	64	4×4	2×2	ReLu
Conv2D_2	128	4×4	2×2	ReLu
Conv2D_3	128	4×4	2×2	ReLu
Conv2D_4	128	3×3	1×1	ReLu
ResNets				
		Pre-VQ		
Conv2D_4	64	1×1	1×1	
		Decoder		
Conv2D_5	128	3×3	1×1	ReLu
ResNets				
Conv2DTranspose_1	128	4×4	2×2	ReLu
Conv2DTranspose_2	64	4×4	2×2	ReLu
Conv2DTranspose_3	1	4×4	2×2	
		ResNets		
Conv2D_res1	32	3×3	1×1	ReLu
Conv2D_res2	128	1×1	1×1	ReLu

Table 6.1: The hyper-parameters for the setup of the VQ-VAE used throughout this study.

defined by silhouette score (Rousseeuw, 1987) is added (Equation 6.5). The silhouette score indicates how well clusters are separated from each other and is defined by the formula,

$$s = \frac{b - a}{\max(b, a)}, \tag{6.4}$$

where a represents the mean intra-cluster distance while b is the distance between a cluster and its nearest neighbour cluster. Therefore, a larger silhouette score indicates a better separation between clusters in feature space. To train our VQ-VAE, we minimise the final loss function combining the loss described in Equation 6.3 and the penalty defined as,

$$L_s = (1 - s)\,\lambda, \tag{6.5}$$

where s represents the silhouette score and λ is a constant used for making the magnitude of this penalty of the same order as other losses used in the VQ-VAE (Section 6.2.1). The value of λ is equal to 0.1 in our case.

As shown in Fig. 6.1, during the training of the VQ-VAE, we interpolate an instance-based clustering algorithm called 'k-medoid clustering' (Maranzana, 1963; Park and Jun, 2009) to obtain two preliminary classification clusters using a flattened index map. The two clusters are then used for measuring a silhouette score to evaluate the performance of the clustering. The Hamming distance (Hamming, 1950) is used as the distance metric as our data is represented by the indices of the vectors in the codebook whereby the number itself only represents a category rather than a real value of the vector (more description in Section 6.2.3). The 'k-medoid clustering' is used here for a fast evaluation; in the main clustering process after feature extraction, we appy hierarchical clustering algorithms (Section 6.2.3).

6.2.3 Uneven Iterative Hierarchical Clustering

In this section we describe our hierarchical clustering procedure for identifying different types of clusters. Hierarchical Clustering (HC; Johnson, 1967; Hastie et al., 2009), in particular agglomerative HC (called sometimes 'bottom-up'), first assigns each input as an individual group, then merges two nearest (the most similar) groups together based upon the measured pair distance in the feature space, recursively. The 'bottom-up' HC structure allows a different number of datapoints in clusters because it starts with individuals (Fig. 6.2). Other kinds of clustering such as 'top-down' HC and K-medoid clustering used in Section 6.2.2 start with clusters themselves, which are more difficult to provide a starting point for an uneven number of datapoints for the initial clusters.

The distance (similarity) measured in this study is the Hamming distance (Hamming, 1950). As stated in Section 6.2.2, our data is represented by the index of the vectors selected from the codebook. This is such that an index indicates a category rather than the real value of a vector. We compare two data sets represented by a set of features labelled with indices. The Hamming distance is defined as the number of mismatched indices between the pair over the number

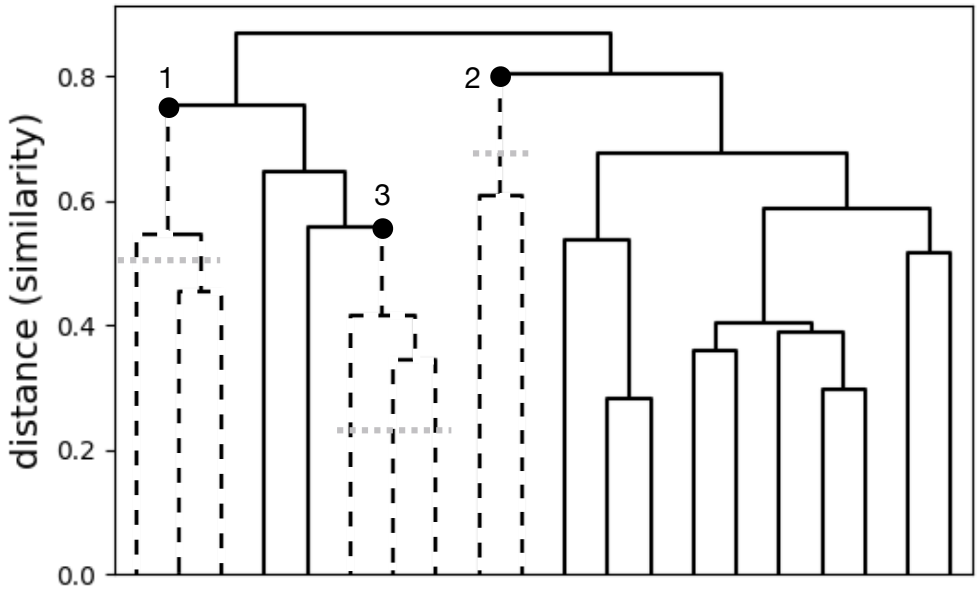

Figure 6.2: The schematic dendrogram of the HC process. Datapoints are shown on the x-axis, and gradually merge with each other based on the distance (similarity) at the y-axis. Each solid line represents a branch and each black circle indicates a stopping point for the corresponding branch (see Section 6.3.4). The dashed lines represents the leaves (clusters) after the stopping points. The gray dotted line indicates a cut suggesting the number of clusters in a branch (also see Section 6.3.4; the results are shown in Section 6.4.2).

of features used to represent the data. For example, assuming that an image can be presented by four different features labelled with the indices: 1, 2, 3, 4, after VQ-VAE; in this case the Hamming distance is 0 if the other image is represented as 1, 2, 3, 4 as well, and the Hamming distance is 1 if it is represented by 4, 3, 2, 1.

For further clarification, Fig. 6.2 illustrates the clustering process. Each solid line is a 'branch' while each black circle is a stopping point for the branch (Section 6.3.4). The dashed lines below circles are leaves (clusters) where the gray dotted lines indicates the distance for the number of clusters in a branch (Section 6.3.4).

Within this study, we realise that when all the data are considered, the merging point can be less accurate due to the mixture of blindly measured distances from a great variety of extracted features in images. Therefore, we carry out an iterative clustering process with a reverse concept that we control the data used for doing HC from the top to bottom. We first make the HC merge all data into two top parent branches, then apply the second round of HC to the data of a parent branch to obtain two children branches, and apply the same procedure again to the sub-data of a child branch to get two grandchildren branches, and so on. The iterative action stops when it reaches a certain condition (the black circle in Fig. 6.2; see Section 6.3.4).

In a typical HC, a uniform distance is used to determine the final clusters. However, a uniform distance threshold is not appropriate considering that galaxies' appearance in different morphological types have different complexity, such that spiral galaxies have a larger diversity in appearance than elliptical galaxies. Therefore, in this study, we propose to allow a different stopping point/distance threshold for each branch depending on the complexity of the objects in the branch (see Section 6.3.4). For example, a branch which consists of galaxies which can look very different within a class may continue for many iterations, while others may reach the stop criteria with fewer iterations due to a relatively monotonous structure within the data of the branch. For example, spiral galaxies can have a variety of spiral arms appearances, e.g., different number of arms, different positions of arms, etc. Therefore, the distance between spiral-like galaxies are generally larger than the distance between two elliptical-like galaxies. This consideration is sensible and is of great importance in morphological classification of galaxies; however, this is neglected in a typical HC algorithm. Therefore, to distinguish it from a typical HC algorithm, we call this setup 'uneven clustering' which provides us with a more precise distinction in galaxy shape, structure, and morphology.

6.3 Implementation

The pipeline of this study includes three main steps: (1) feature selection; (2) feature learning (using the modified VQ-VAE); and finally (3) clustering process. The data used in this study are introduced in Section 6.3.1. The feature selection

This work	E	S0	eSp	lSp	Irr
	E	$S0^-$, S0	S0/a - Sab	Sb - Sdm	Irr
DS18	-3	-2, -1	0 - 2	3 - 8	10

Table 6.2: The classification scheme used in this work and in Domínguez Sánchez et al. (2018, DS18; presented in T-Type). In DS18, they define the T-Type of -3 for ellipticals (E), -2 for lenticulars at the early stage ($S0^-$), -1 for lenticulars at the intermediate to late stages ($S0$), 0 for $S0/a$, and the positive values of T-Type are for different stages of spirals. Finally the T-Type of 10 represents irregular galaxies (Irr).

is described in Section 6.3.2, and the setup for the feature learning process using the modified VQ-VAE (Section 6.2.2) is discussed in Section 6.3.3. Finally, in Section 6.3.4 we explain the details of the clustering process we use to classify galaxies.

6.3.1 Data Sets

The imaging data used throughout this work is from the Sloan Digital Sky Survey (SDSS) Data Release 7 (York et al., 2000; Abazajian et al., 2009) with a redshift cut of $z < 0.2$. In order to focus on the impact of galaxy shape and structure to morphological classifications, we utilise monochromatic r-band images. An extension including colour and other factors is some to consider for the future. Here we are focused on single-band morphological classification on features seen and not in general a physical classification that might result from considering galaxy colours and colour distributions.

To examine what types of systems our classification clusters contain, as well as to have the flexibility within the data distribution in our datasets, we use morphology labels defined by T-Type (de Vaucouleurs, 1964) and the probability of being a barred galaxy (P_{bar}). The two quantities are both obtained using deep learning techniques from Domínguez Sánchez et al. (2018, hereafter, DS18). We define eight labels including barred galaxies that contain significant features shown in the Hubble morphological system: ellipticals (E), lenticulars (S0), early spirals (eSp), late spirals (lSp), irregulars (Irr), barred lenticulars (SB0), early barred spirals (bar eSp), and late barred spirals (bar lSp).

The comparison of the classification scheme is shown in Table 6.2; in which, S0, eSp, and lSp are separated into barred and non-barred galaxies based on the value of P_{bar}. We additionally include labels of irregular galaxies from three other works: Fukugita et al. (2007), Nair and Abraham (2010), and Oh et al. (2013) to provide more irregular galaxies in our database. The morphological labels in our datasets are not used for training our machine, but to prepare an appropriate dataset with a specific data distribution, and as a way to examine the obtained clusters in terms of these types.

To investigate the differences in the classification systems defined by humans and those from a machine, as well as potential application within our unsupervised

machine learning technique in future surveys, we prepare two different datasets: which are 'balanced' and 'imbalanced'. In the balanced dataset, we artificially allocate the same number of galaxy images to each morphological type. The eight human defined morphological types have visually distinctive differences from each other; therefore, the purpose of this arrangement is to allow our VQ-VAE consider fairly the characteristics of each morphology type when extracting the representative features from input images. Otherwise it is possible that some type of bias would result if the distribution of the types we select are input into our VQ-VAE in the same fraction as they are found in the nearby universe. In this case we would find that the late-type disks would dominate over early disks and ellipticals (e.g., Conselice, 2006).

On the other hand, it is of great importance to know how an unsupervised machine learning technique can be applied in future surveys to explore a large scale of unknown galaxies' morphology in an 'as is' situation. That is, we need to know how our VQ-VAE performs when galaxies are inputted from imaging observations of the real universe with no balancing. For this goal, we set up the 'imbalanced dataset' with a realistic distribution in terms of galaxy morphological types which follows the distribution of nearby galaxies at z=0.033-0.044 as presented in Oh et al. (2013). The type distributions of the balanced and imbalanced dataset are shown in Fig. 6.3.

6.3.2 Feature Selection

In this section we discuss a preprocessing procedure to reject irrelevant information from images. The feature selection procedure is used to select the pixels in images that are significant and which reflect the shape or structure of the targets. Cheng et al. (2020b, Chapter 5) showed that the background noise can result in an overfit to the noise when training the convolutional autonencoder. To solve this, Cheng et al. (2020b, Chapter 5) applied a simplified convolutional autoencoder to denoise the images and emphasise the pixels from the targets themselves before the main task is computed. However, a denoising process by another autoencoder is time-consuming and could potentially add artificial structure when reconstructing the images. Therefore, in this study, we simply use a one sigma clipping of pixel values measured through the background noises as our selection threshold. Any pixel value is below this criterion the pixel value is set as 0 (Martin et al., 2019). Whilst this will remove noise, it will also potentially remove outer fainter portions of the galaxies themselves. However, this will retain the brighter portions of the inner parts of galaxies where classification is done in any case. Removing this fainter light does not have an effect on our measurements as it would if we were measuring for example surface brightness profiles.

6.3.3 Feature Learning

As described, in this study, we apply a modified vector-quantised variational autoencoder (VQ-VAE) (see Section 6.2.2) to carry out our unsupervised learning. Our VQ-VAE basically learns the representative features from our images. It considers a preliminary clustering result by including an additional penalty

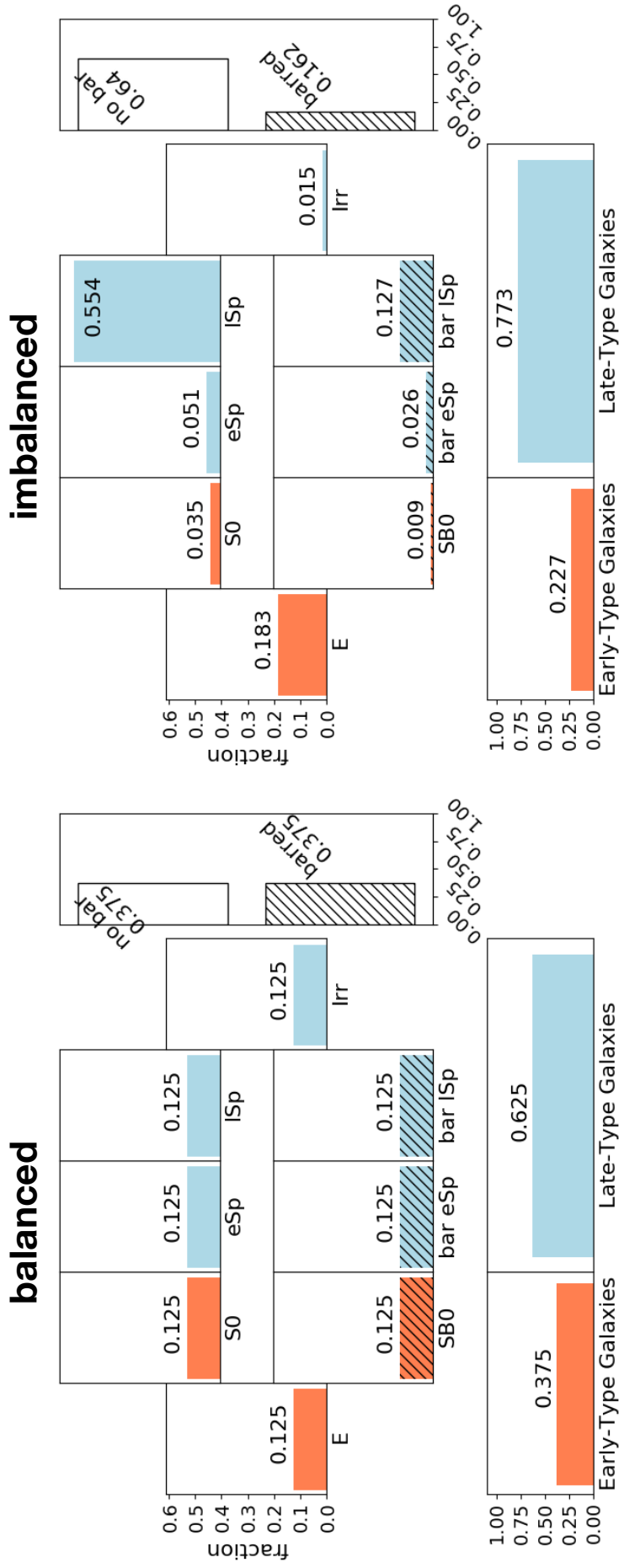

Figure 6.3: The type distributions of the balanced (left) and imbalanced (right) datasets. The latter follows the distribution of nearby galaxies (Oh et al., 2013). The number shown above the coloured bar represents the fraction of the type in all data. The fraction of barred galaxies are highlighted with hashed lines general on the bottom of the 'tunning fork'. The orange and light blue colouring represent early-type galaxies and late-type galaxies, respectively.

(Equation 6.5) in the VQ-VAE (Section 6.2.2). This modification helps to find not only better representative features for image reconstruction, but also the features that can be well separated into two initial groups in feature space.

The main advantage of the VQ-VAE technique is to accelerate the unsupervised feature extraction process which is over 30 times faster than using a typical convolutional autoencoder (e.g., Cheng et al., 2020b) without a significant trade-off to the reconstruction accuracy (Razavi et al., 2019). This is achieved by quantising the values used for reconstruction (Section 6.2.1).

The hyper-parameters setting used in this study follows the setup described in Razavi et al. (2019) except for the codebook size. It determines the number of vectors available in the quantisation process (Section 6.2.1). This number of vectors decides the 'resolution' of the reconstructed images. Namely, the more available vectors, the more details can be presented in images. Razavi et al. (2019) use 512 vectors in their codebook to generate high-fidelity emulated images of different animals, e.g., dogs, cats. However, with a different goal from emulation in our study, we realised through out analysis that a larger codebook size leads to a worse clustering result. This is because the machine with a larger codebook uses too many details of the images into account when carrying out the clustering. These details help to complete the puzzle when emulating images but they blur the boundary in the feature space when doing clustering. In this study, after a series of tests, we choose a size of 16 for our codebook, which forces the machine to use the provided vectors on the most significant features while still retaining a certain level of the reconstruction quality. This number of 16 was determined through experimental method, and is not based on any basic principles related to galaxies or machine learning. It may, and probably does, differ within different instances of use.

6.3.4 Clustering

Within the clustering task, we apply an uneven iterative hierarchical clustering (Section 6.2.3) on the data represented by a set of vector-quantised features obtained after the VQ-VAE.

In this study, we propose a new approach to decide the number of clusters within unsupervised machine learning applications. This approach can be used in other instances beyond using a VQ-VAE. Part of this is inspired by the fact that the clusters can be highly sensitive to galaxy orientation. The concept we use is to take the threshold measured by the features of galaxy orientation on the merger tree to find where the effect of galaxy orientation in a branch starts to appear (e.g., gray dotted lines in Fig. 6.2). In other words, this threshold also provides the number of classification clusters that are not separated based on the galaxy orientation. This threshold is defined by the average distance between the artificially rotated images in a branch ($\overline{d_{rot}}$),

$$\overline{d_{rot}} = \frac{\sum_i^N \sum_j^N d_{ij}}{N\,(N-1)}, \tag{6.6}$$

where N is the number of datapoints in the branch, and d_{ij} represents the distance between an image i and image j. The distance, d_{ij}, is measured through the Hamming distance.

In this process we stop a branch and decide the number of clusters within that branch when one of two criteria is satisfied: (1) the $\overline{d_{rot}}$ suggests fewer than two clusters (≤ 2) in a branch; (2) the difference between the $\overline{d_{rot}}$ measured using the data of a parent branch and the data of a child branch are smaller than 0.015: that is, $\overline{d_{p,rot}} - \overline{d_{c,rot}} \leq 0.015$.

The first criterion indicates that galaxy orientation is considered when having more than two clusters (> 2) in this branch (e.g., circle 1 and 2 on Fig. 6.2). Two clusters are the minimal number to split; therefore, we stop the iterative clustering in a branch when this criterion is satisfied. On the other hand, the second criterion is used to decide whether a branch (the parent branch) should have more sub-branches (the child branches). The variation between branches is less significant when the difference in the distance between the data of a parent branch and a child branch is small (≤ 0.015). Therefore, there is no need to split a parent branch when the second criterion is satisfied. The suggested number of clusters by the $\overline{d_{rot}}$ of the parent branch is then the number of clusters in the branch without having the effect of galaxy orientation. For example in Fig. 6.2, the branch stops at the circle 3 by satisfying the second criterion, and the $\overline{d_{rot}}$ (gray dotted line) suggests three clusters without showing the effect of galaxy orientation in this branch.

6.4 Results and Discussion

6.4.1 Unsupervised Binary Classification

Starting with a simple examination, we enforce our machine to merge all galaxies in the balanced dataset into two preliminary clusters. Examples of galaxies within the two clusters are shown in Fig. 6.4. Galaxies in one cluster have clearly more features (featured group; e.g., arm structure) than the galaxies of the other cluster (less featured group; more elliptical). We examine the morphological distribution in both clusters (left column in Fig. 6.5); one cluster has $\sim 96\%$ late-type galaxies (LTGs) and the other one has $\sim 60\%$ early-type galaxies (ETGs).

Due to an unequal number between the ETGs and the LTGs in the balanced dataset (Fig. 6.3), the fraction of ETGs and LTGs in each cluster might be biased. We examine another quantity, 'dominance', which represents the ratio between the fraction of a certain type in a given cluster to the fraction of this type within the dataset (right column in Fig. 6.5). This quantity removes the statistical influence from different number of types used in the input datasets; hence, it shows a better representation of the galaxy features emphasised in the cluster. Through the dominance distribution, we observe that the featured and less featured group are clearly dominated by the features of LTGs and ETGs, respectively.

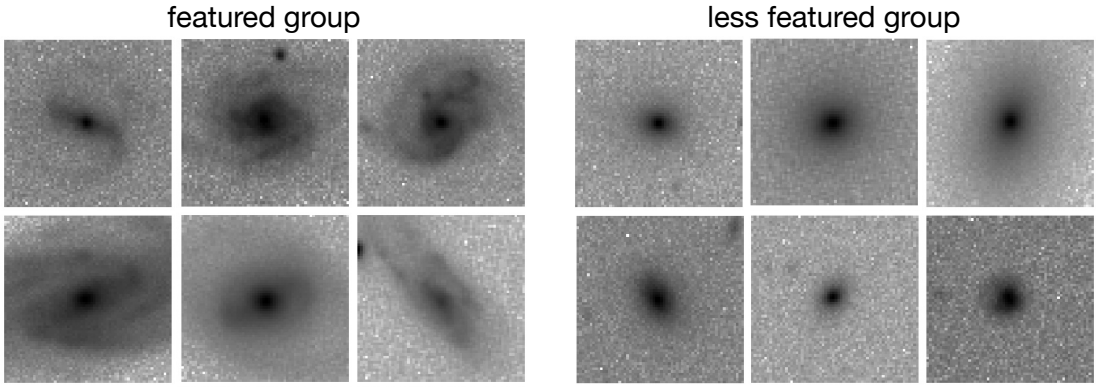

Figure 6.4: Examples of galaxies found within our two preliminary clusters using the balanced dataset. Galaxies in one cluster have more features (left left), and galaxies in the other group have relatively fewer features (right).

We further investigate the potential structural factors considered when separating the two clusters. With the analysis of the two clusters, we can decide what are the major structural factors in the clustering process. First of all it is clear that with our unsupervised learning we obtain a separation into two main clusters where one correlates with late-type galaxies and the other with early-type galaxies. This verifies with a machine this basic dichotomy which has existed in classification schemes for over 100 years.

However, we also want to compare our clusters with more quantitiative measures. In Fig. 6.6, we compare a variety of structural measurements such as concentration, asymmetry, smoothness/clumpiness, Sérsic index, Gini coefficient, M20, apparent half-light radius (R_e, arcsec), and r-band apparent magnitude (m_r) between the two clusters. These measurements, except for the r-band magnitude, are provided from the catalogue of Meert et al. (2015), and the r-band magnitudes are from Simard et al. (2011). Within these measurements, the asymmetry, Sérsic index, Gini coefficient, and M20 show a clear separation between the two clusters in Fig. 6.6. This indicates that our machine takes galaxy structure which correlates with measurable strcutural parameters (asymmetry, Gini coefficient, M20) and light distribution (Sérsic index) into account rather than the apparent size and the apparent brightness of galaxies, when categorising galaxies into the two clusters. This is good, as it shows that our method does not depend on distance or the apparent sizes of galaxies but on the inherent morphologies and structures of the galaxies themselves.

Note that the concentration and smoothness distributions show fewer differences between the two clusters. These two quantities also do not have apparent differences between the LTGs and ETGs in our dataset, because the galaxies in our datasets are relatively faint ($\sim 74\%$ galaxies fainter than $m_r = 16$) and the image resolution is limited by the ground-based seeing (> 1 arcsec; the image sampling is 0.396 arcsec per pixel). Although we cannot straightforwardly confirm the correlation between the two clusters and the concentration parameter, the Gini coefficient and M20 provide a connection with the concept of concentra-

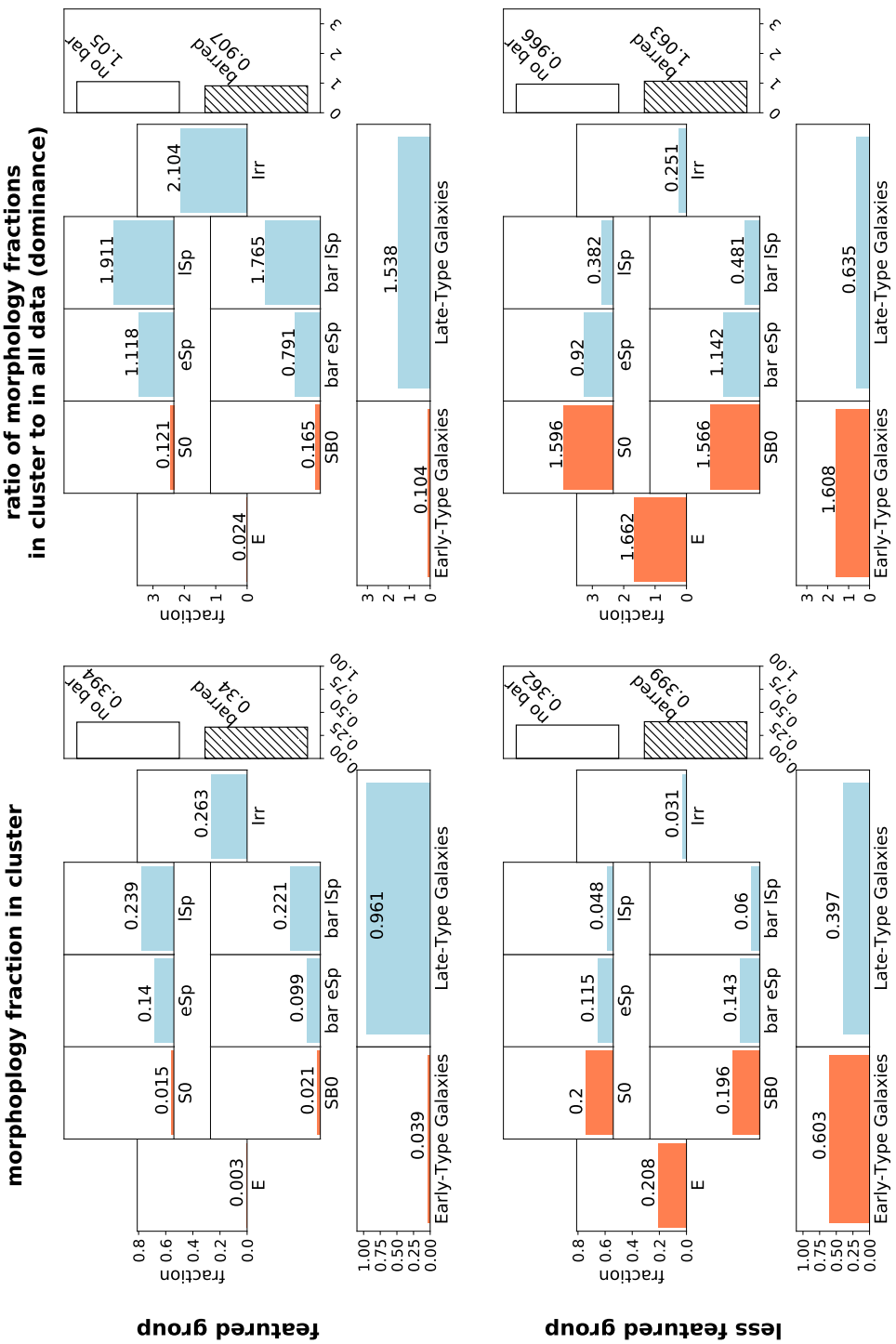

Figure 6.5: The distribution of visual galaxy morphology in each cluster obtained using the balanced input dataset. The left column shows the fraction of each morphology type in the clusters while the right column presents the dominance of each type. The 'dominance' is defined by the fraction of a certain morphology type in the cluster divided by a fraction of this type within the dataset. The top row shows the distribution of the 'featured group' while the bottom row presents the statistics for the 'less featured group'.

tion.

Based on our visual assessment, we proceed to associate the featured group to LTGs and the less featured group to ETGs in order to compare these machine-predicted labels with the catalogue labels. Using the balanced dataset, the machine-predicted and the catalogue labels agree with an accuracy of ~ 0.75 in this binary classification. The accuracy is defined as the number of the correct matches between the machine labels and the catalogue labels from all galaxies in the dataset.

In Fig. 6.7, we present the T-Type distribution between the two clusters. It shows that the main confusion in binary classification by our machine happens when classifying early spirals into either ETGs or LTGs, in particular, *Sab* galaxies (T-Type=2). When we exclude early spirals from the balanced dataset, the accuracy increases to $\sim$0.87 for binary classification.

We discuss some plausible reasons for this 'failed' classification by our machine. For example, one uncertainty originates from the provided labels which combine the uncertainty of both visual classifications and machine learning predictions. Second, from our machine's perspective, in addition to the potential machine learning uncertainty, another possible uncertainty is caused by the reconstruction inaccuracy in the VQ-VAE, particularly within spiral galaxies with insignificant arm structures. However, although these causes are unavoidable, these conditions exist only in a fairly small fraction of the data in the input imaging dataset. The main reason for the mixture of early spirals in both clusters is due to the intrinsic difficulty of classifying this type into either ETGs or LTGs based only on galaxy structure. The 'early spirals' in fact include a wide range of transitional features which are difficult to accurately define. The separation may become better when including colour information; however, with our method, we state the difficulty to discriminate early spirals when considering only galaxy appearance/structure in a unsupervised machine learning methodology.

6.4.2 Machine Classification Scheme

In the previous section, we enforce our machine to provide two initial clusters for a preliminary examination. However, the main motivation for this study is to investigate the classification system a machine would suggest when 'looking' at galaxies and classifying them through machine learning. We use the proposed method in Section 6.3 with the balanced dataset to let the machine explore freely and suggest the number of clusters to categorise the galaxies in the dataset. Galaxies in our dataset are categorised into 27 classification clusters by our machine. Comparing with previous work on unsupervised learning which produced 160 clusters (Martin et al., 2019). Our method suggests significantly fewer number of galaxy morphology classifications which is more in line with what one would surmise is a more accurate number of classes for galaxies. In addition to the different implementations applied in both works, the difference in the number of obtained clusters might be due to the fact that we only consider monochromatic images to investigate the impact of galaxy structure in this study, while Martin et al.

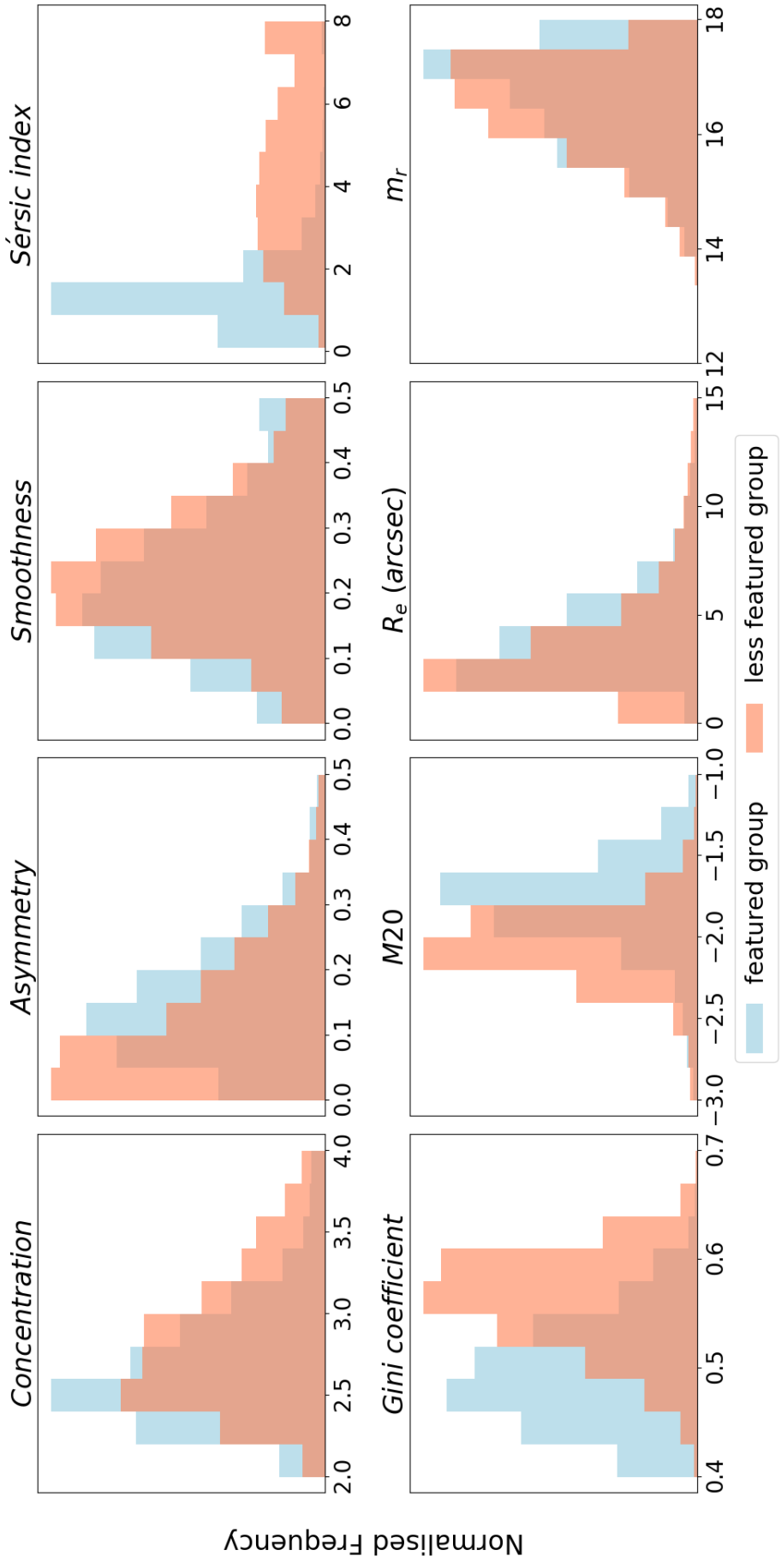

Figure 6.6: The comparison of structural measurements including: concentration, asymmetry, smoothness/clumpiness, Sérsic index, Gini coefficient, M20, half-light radius (R_e), and r-band apparent magnitude (m_r) between the two initial clusters. The blue shading represents the featured group while the orange shading is for the less featured group.

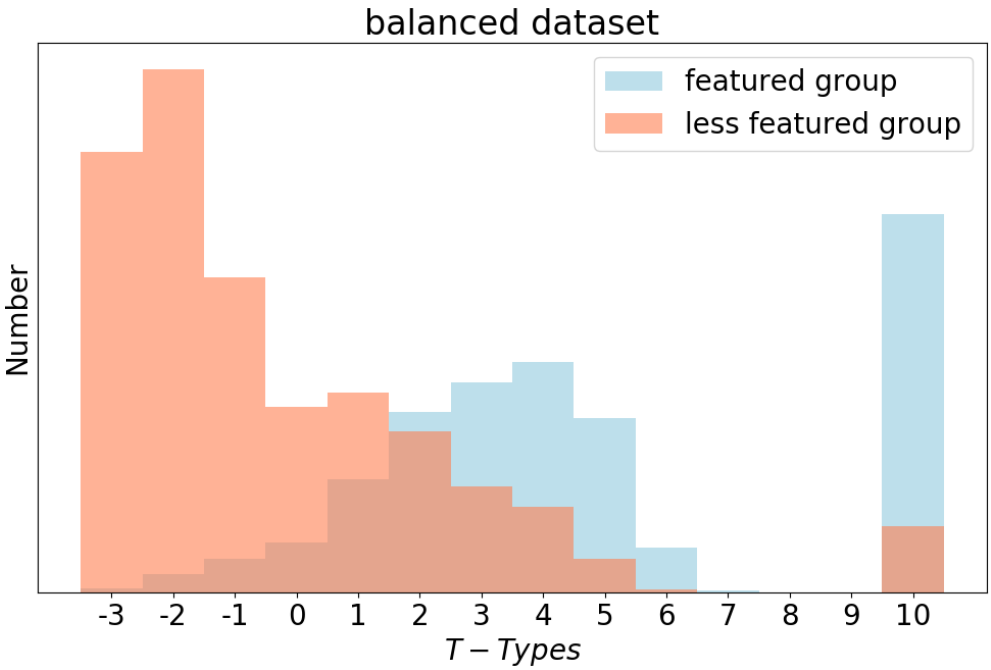

Figure 6.7: The T-Type distribution between the two preliminary clusters within the balanced dataset. The corresponding visual morphology class is shown in Table 6.2. The blue shading shows the distribution of the featured group, while the light orange colour represents the less featured group.

(2019) used coloured images. Additionally, to have more available measurements of galaxy structure and properties, we choose to use the imaging data from the Sloan Digital Sky Survey (SDSS; York et al., 2000; Abazajian et al., 2009) which has a worse resolution and image sampling (0.396 arcsec per pixel) than the one used in Martin et al. (2019, 0.168 arcsec per pixel). This may be a reason for the resulting fewer number of clusters obtained in our work. To further investigate galaxy morphology classifications, the colour information and images with better resolutions will be considered in future work.

Examples of images from each of the 27 clusters are shown in Fig. 6.8. The number shown on the bottom left is the average value of the T-Type in the clusters and the identification number of the cluster is shown on the top right. The identification numbers of groups are generated on the merger tree from left to right; therefore, they are simply labels without physical meanings. Table 6.3 lists the characteristics of each cluster in structural measurements, galaxy properties, and statistics. This can be used to co-examine the figures shown from this section to Section 6.4.4. Through visual assessment in Fig. 6.8, we observe that galaxies in some clusters show bars (e.g., g15 and g16 in Fig. 6.8) or show more elongated in shape than in others.

In Fig. 6.9, we re-examine the influence of the major structural parameters such as the Sérsic index, asymmetry, Gini coefficient, and M20 (Section 6.4.1), in separating clusters. Each coloured circle represents one cluster and is coloured by the average value of the T-Type in the cluster. We confirm again a clear

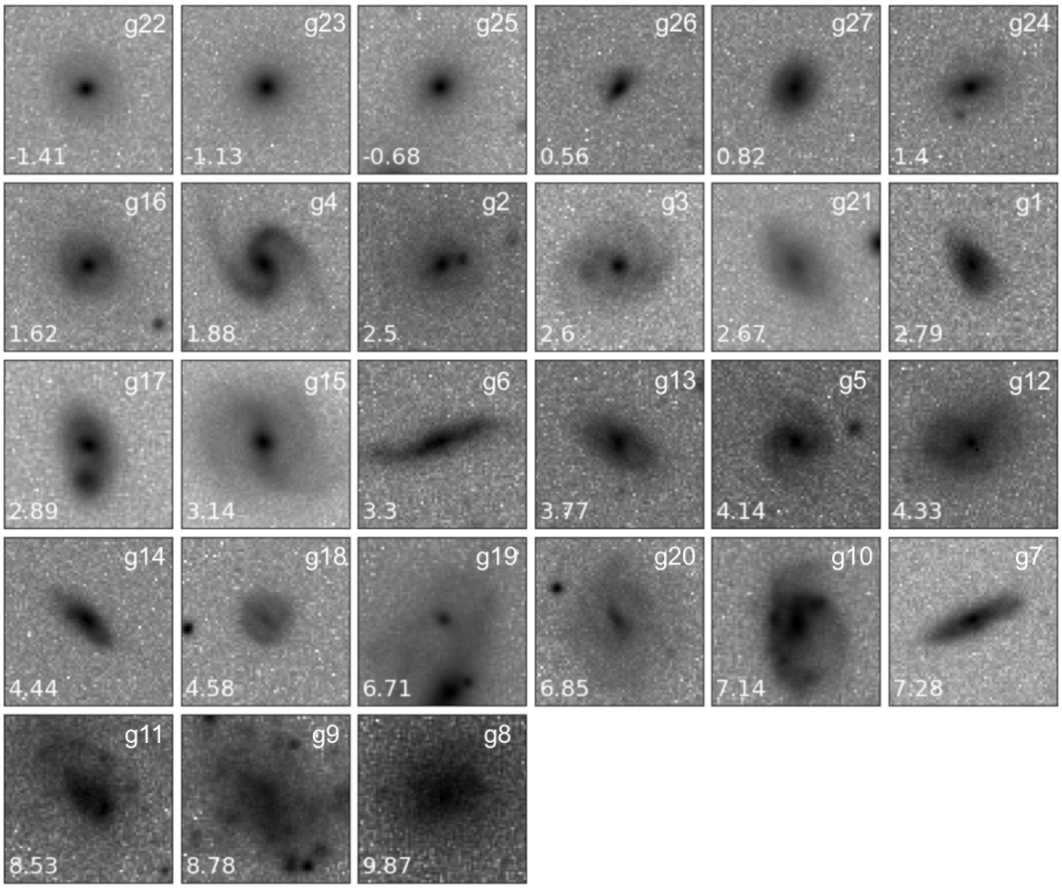

Figure 6.8: Examples of images from each cluster listed in the order of the average value of the T-Type within that cluster (Table 6.2). The number shown at the left bottom corner is the average value of the T-Type in the cluster. At the right top corner, the identification number of the belonging cluster for the image is presented.

correlation between our machine classification clusters and major structural features. Additionally, the given clusters show a transition along with the T-Type. This suggests the clusters are correlated with the visual morphology roughly from early-types to late-types.

6.4.3 Machine Classifications versus Human Visual Classifications

It is important to note that the goal of this work is not to find a perfect agreement between our machine-based classification and the visual morphologies. Our goals are to understand the features used by our method to categorise galaxy images, and to introduce a novel classification scheme 'proposed' by our machine. That is, we want to develop a scheme whereby galaxies are classified by a reproducible and scientific computational way and not by human opinion.

Group ID	<Sérsic n>	<Gini>	<M20>	<A>	$< g-r >$	$< Mag_r >$	<logM_*> ($M_\odot$)	$< R_e >$ (kpc)	N_g (F_g)	D_g ($F_{g,D}$)	$F_{g,bar}$	$D_{g,bar}$ ($D_{g,nobar}$)
g1	1.3	0.48	-1.84	0.16	0.63	-21.16	10.31	6.98	896 (1.4%)	eSp/lSp (0.97)	0.54	1.45 (1.22)
g2	1.6	0.47	-1.91	0.16	0.71	-21.5	10.47	9.48	441 (0.69%)	eSp/lSp (0.93)	0.68	1.82 (0.83)
g3	1.68	0.46	-1.85	0.15	0.71	-21.61	10.56	9.83	287 (0.45%)	eSp/lSp (0.87)	0.74	1.97 (0.7)
g4	1.63	0.5	-1.92	0.14	0.73	-21.32	10.46	6.92	2924 (4.57%)	eSp/lSp (0.79)	0.34	0.91 (1.75)
g5	1.17	0.46	-1.84	0.13	0.52	-20.19	9.79	6.52	2141 (3.35%)	lSp (0.76)	0.46	1.22 (1.3)
g6	1.08	0.5	-1.85	0.14	0.63	-20.53	10.12	6.06	2463 (3.85%)	eSp/lSp (0.8)	0.14	0.37 (2.17)
g7	1.35	0.51	-1.73	0.19	0.46	-20.31	9.8	5.05	3055 (4.77%)	lSp/Irr (0.78)	0.16	0.42 (0.67)
g8	0.82	0.44	-1.55	0.14	0.38	-19.45	9.37	3.98	510 (0.8%)	Irr (0.97)	0.02	0.04 (0.03)
g9	1.26	0.47	-1.64	0.16	0.36	-19.82	9.49	5.26	1291 (2.02%)	lSp/Irr (0.94)	0.16	0.43 (0.13)
g10	1.13	0.48	-1.65	0.19	0.42	-20.31	9.75	5.15	946 (1.48%)	lSp/Irr (0.94)	0.29	0.78 (0.47)
g11	1.27	0.48	-1.66	0.18	0.36	-19.88	9.49	5.2	1130 (1.77%)	lSp/Irr (0.88)	0.17	0.44 (0.29)
g12	1.33	0.46	-1.73	0.15	0.55	-20.99	10.22	7.32	1054 (1.65%)	lSp (0.85)	0.74	1.99 (0.5)
g13	1.01	0.46	-1.75	0.14	0.51	-20.43	9.92	6.01	941 (1.47%)	lSp (0.81)	0.51	1.37 (1.27)
g14	1.39	0.52	-1.83	0.14	0.63	-20.62	10.16	5.7	2079 (3.25%)	eSp/lSp/Irr (0.86)	0.12	0.32 (1.76)
g15	1.85	0.48	-1.87	0.14	0.69	-21.64	10.61	8.9	1397 (2.18%)	eSp/lSp (0.87)	0.73	1.94 (0.64)
g16	2.87	0.51	-2.02	0.15	0.83	-22.04	10.81	11.5	776 (1.21%)	S0/eSp/lSp (0.8)	0.8	2.12 (0.51)
g17	1.47	0.48	-1.8	0.15	0.65	-21.43	10.46	7.15	989 (1.55%)	eSp/lSp (0.93)	0.65	1.72 (0.87)
g18	1.82	0.53	-1.79	0.18	0.65	-20.95	10.2	6.51	553 (0.86%)	eSp/lSp/Irr (0.79)	0.27	0.72 (0.98)
g19	1.43	0.5	-1.69	0.13	0.57	-20.59	10.0	6.4	1013 (1.58%)	Irr (0.59)	0.17	0.46 (0.64)
g20	1.53	0.5	-1.69	0.15	0.54	-20.63	9.96	6.76	982 (1.53%)	lSp/Irr (0.71)	0.22	0.58 (0.53)
g21	2.56	0.53	-1.9	0.12	0.76	-21.29	10.46	7.8	2138 (3.34%)	S0/eSp/lSp/Irr (0.68)	0.29	0.76 (1.39)
g22	4.64	0.57	-2.09	0.1	0.94	-22.03	10.94	7.32	12733 (19.9%)	E/S0 (0.78)	0.3	0.81 (0.87)
g23	4.71	0.57	-2.09	0.11	0.94	-21.93	10.87	7.18	8474 (13.24%)	E/S0 (0.8)	0.4	1.07 (0.67)
g24	3.17	0.53	-2.04	0.13	0.81	-21.82	10.73	9.14	6420 (10.03%)	S0/eSp/lSp (0.69)	0.69	1.85 (0.56)
g25	3.81	0.56	-2.05	0.12	0.94	-21.67	10.78	6.26	3485 (5.45%)	S0 (0.62)	0.23	0.61 (1.77)
g26	2.62	0.53	-2.02	0.13	0.85	-21.52	10.62	7.36	2056 (3.21%)	S0/eSp/lSp (0.88)	0.27	0.72 (1.89)
g27	2.53	0.52	-1.99	0.14	0.85	-21.64	10.69	8.08	2826 (4.42%)	S0/eSp (0.71)	0.53	1.41 (1.21)

Table 6.3: The table lists the average values of structural measurements [Sérsic index, Gini coefficient, M20, Asymmetry (A)] and galaxy properties [$g-r$, r-band absolute magnitude (Mag_r), stellar mass (logM_*), physical size (R_e, kpc)] in each machine-defined cluster. Additionally, the statistics of each cluster are presented in the last four columns where N_g shows the number of galaxies in the cluster and F_g indicates the percentage of total samples. The D_g lists the dominated types in each cluster, which are selected based on the dominance of each morphology type, and $F_{g,D}$ shows the fraction of the dominated types in a cluster. The $F_{g,bar}$ is the fraction of barred galaxies in a cluster. Finally, $D_{g,bar}$ and $D_{g,nobar}$ is the dominance of barred galaxies and non-barred galaxies in a cluster, respectively. The ordering follows the group IDs which are simply labels for convenience.

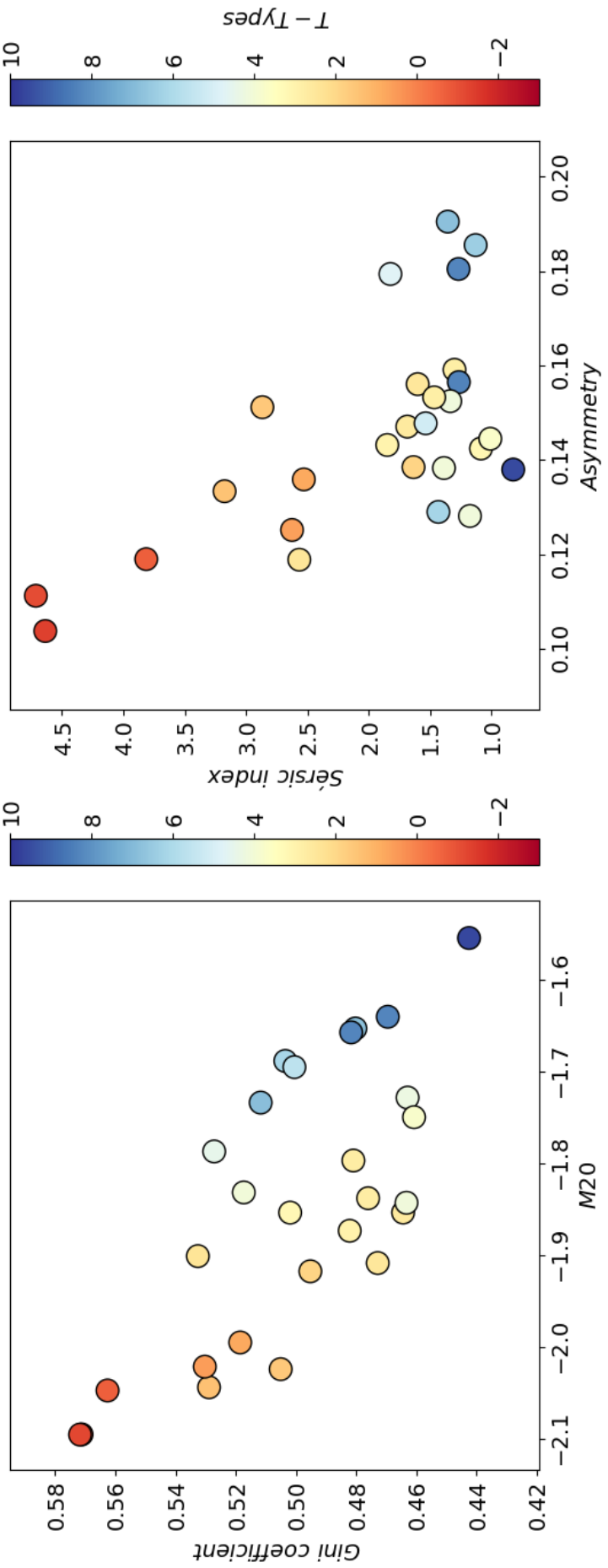

Figure 6.9: The comparison of the major structural features such as the Gini coefficient, M20, Sérsic index and the asymmetry as a function of each cluster from Section 6.4.1. Each circle represents one classification cluster from our unsupervised machine learning process which is coloured by the average T-Type in the cluster. The average value of the data in the clusters are used for each structural feature value. Clearly strong trends can be seen that varies along the clusters.

To better understand our machine-based classes, we compare them with visual morphological classes such as the Hubble sequence, and discuss the visual features extracted by our machine. To do this comparison, we associate each cluster with one or a mix of Hubble types based on the dominance of each type within each of the clusters (Fig. 6.10). As mentioned in Section 6.4.1, the 'dominance' of each type is the ratio between the fraction of a given morphology type in the cluster to the fraction in the dataset. We associate a given cluster with one or several morphology types when the dominance of a certain type is > 1. This selection indicates which kinds of visual features considered in a visual morphology type are dominated in a cluster.

In Fig. 6.10, we show the accumulated distribution of the classification clusters to one or a mix of visual morphology types. Each coloured bar represents one cluster and the deeper bluer colours indicate more barred galaxies than non-barred galaxies within that given cluster. In Fig. 6.10, the darkest blue represents a cluster with the strong bar dominance, $D_{g,bar} \geq 1$ and the non-bar dominance, $D_{g,nobar} < 1$ (see the last column in Table 6.3; e.g., g16 in the table). The medium blue is for a cluster with both bar and non-bar dominance ≥ 1 (weak bar dominance; e.g., g27 in Table 6.3). This criterion indicates that the features of a barred galaxy are not distinctive in a cluster. The lightest blue is used when the bar dominance is $D_{g,bar} < 1$ (no/less dominance; e.g., g14 and g19 in Table 6.3). Through the highlight of the bar dominance in clusters in Fig. 6.10, our machine is shown to successfully discriminate between barred and non-barred galaxies. Examples of clusters with different bar dominance are shown in Fig. 6.11.

We observe in Fig. 6.10 that no cluster is dominated by either elliptical galaxies or early spirals only. The features of elliptical galaxies are recognised to have a great similarity to some lenticular galaxies by our machine. Visually, we separate ellipticals and lenticulars mainly based on the disk structure. However, compared to the cluster dominated by only lenticulars (the g25 in Table 6.3) in Fig. 6.12, the galaxies in the two clusters dominated by $E/S0$ (g22; g23) lack significant disk structure, whereas 'g22' represents the 22th cluster, and so on (also see Fig. 6.8 and Table 6.3). However, clusters with more disky galaxies, such as g27 (blue solid line in Fig. 6.12), are dominated by a mix of $S0$ and eSp. This is likely an indication for an uncertainty in distinguishing ellipticals, lenticulars, and early spirals in the visual classification system we use and not a defect of our unsupervised learning. Only the lenticulars with a moderate range of Sérsic index (peaks at ~ 3; yellow solid line in Fig. 6.12) can be separated from other morphology types.

Additionally, as stated in Section 6.4.1, early spirals are difficult to categorised into either ETGs or LTGs, and as such it is difficult to have a distinctive cluster dominated by only this morphology type (Fig. 6.10) due to the broad transitional features in this type. This again indicates the intrinsic difficulty of visually separating early spirals from other morphology types, such as lenticulars and late spirals.

Most of our clusters have a mixture of different Hubble types within them which indicates galaxies with similar features in appearance can be visually classifying into a variety of morphology types (see examples in Fig. 6.13). In other words, a mix of galaxy structure in fact exists in a visually defined morphology type. This result reveals an intrinsic vagueness of the visual classification systems such that they are not always accurately defined, with many galaxies not optimally classified as a certain T-Type due to the diversity of properties beyond a guess at morphology.

One exception from the above discussion is our cluster 21 (g21 in Table 6.3 with a mix of four morphology types (S0, eSp, lSp, Irr). This cluster is shown to have galaxies with bright companions which overwhelms the brightness of the central objects (the 'g21' row shown in Fig. 6.13). After the feature selection and normalisation in Section 6.3.2, the central objects might become negligible to the machine learning compared to the companions. This can result in difficulty for our machine to capture the structure of the central objects as well as group these galaxies correctly. On the other hand, galaxies with companions are more likely to experience galaxy mergers, and thus this cluster can be used as an indication to find potential merger events or compact groups of galaxies.

6.4.4 Machine Classifications versus Physical Properties

In previous sections, we show that our machine learning classifications trained with monochromatic images are categorised based on structural features (Section 6.4.2) and visual features (Section 6.4.3). In this section, we use the machine classification scheme to study the correlation of galaxy physical properties and galaxy morphology using the colour-magnitude diagram and the mass-size relation of galaxies.

In Fig. 6.14, we examine our the machine classification clusters plotted on the colour-magnitude plane (left) and the mass-size plane (right). The colours and physical sizes are again taken from Simard et al. (2011) while the stellar mass originates from Mendel et al. (2014). Each circle represents one cluster, coloured by the average value of the stellar mass of the galaxies in the cluster for the colour-magnitude diagram and by the average value of colour for the mass-size relations. These two plots show that each galaxy cluster as defined by the machine has distinctive physical properties in galaxy colour, absolute magnitude, stellar mass, and physical size. Additionally, our machine classes show a clear transition between galaxy morphology and galaxy properties on both the colour-magnitude diagram and the mass-size relations. Each star shows the average value of the data with a certain visual morphology type (written in black) for comparison. The machine-defined morphology types fill in the gap within the correlation of galaxy morphology and galaxy properties along with the Hubble types. This indicates that the machine classification scheme can complete the missing morphologies in the visual classification systems without involving human potential bias. It will be interesting to investigate the correlation of these machine-defined classifications with galaxy environment and other galaxy properties, but this will be left to study in a future work.

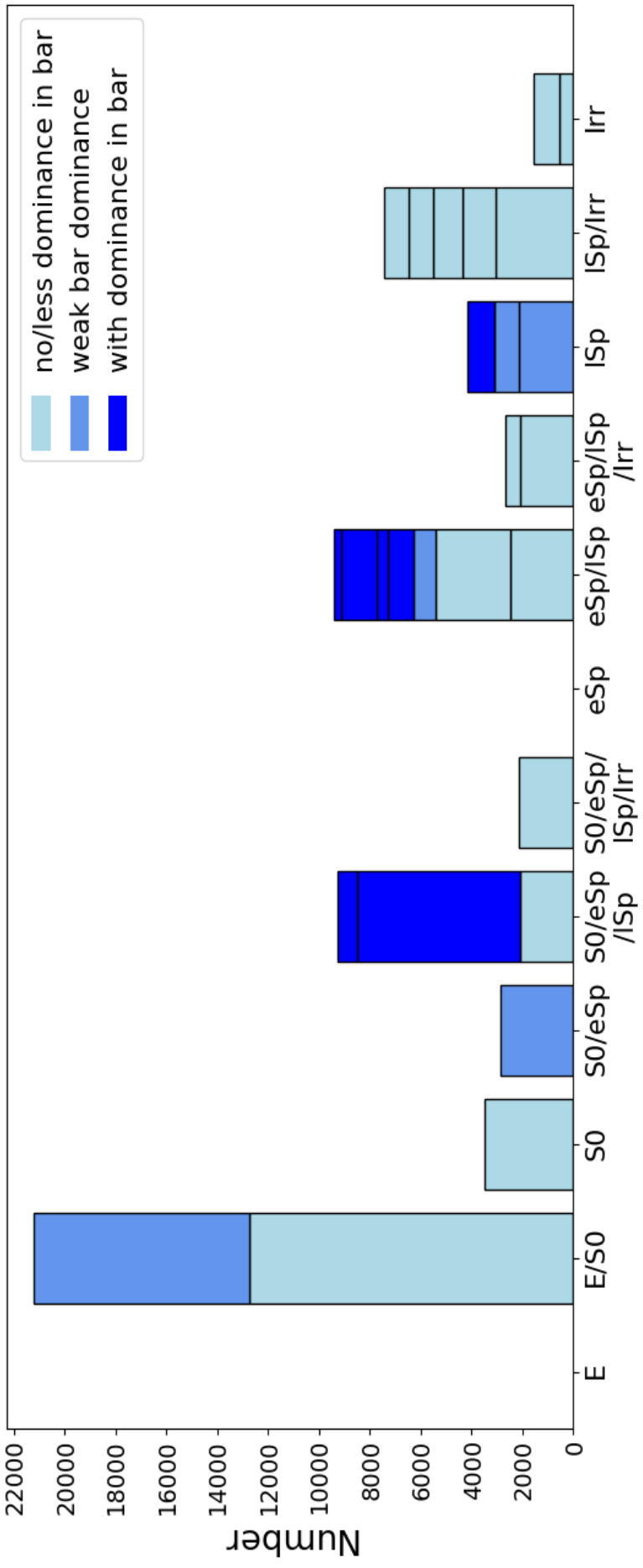

Figure 6.10: The accumulated distribution of the classification clusters compared with Hubble sequence morphological types. The x-axis shows one or a mix of visual morphology types which dominates the clusters listed in Table 6.3. All 27 clusters are plotted here, and each coloured bar represents one cluster. The different colours of the bars show different dominance levels of barred galaxies in the cluster, such that from deep to light blue represent more barred galaxies to no/fewer barred galaxies in the cluster.

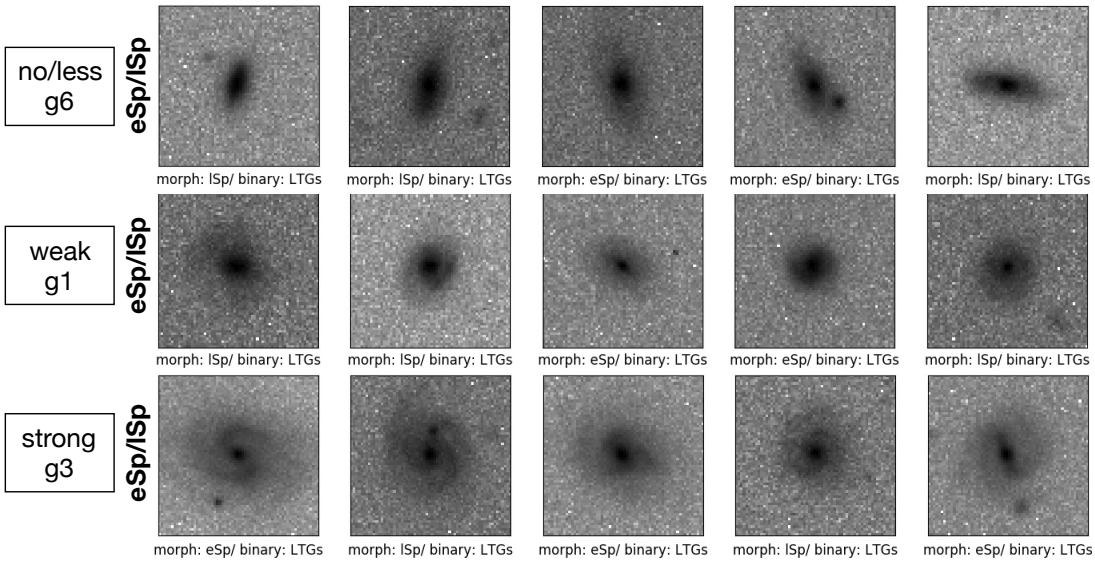

Figure 6.11: Examples of the clusters with different bar dominance levels. Each row shows five randomly picked examples in the cluster, where 'g6' represents the 6th cluster, and so on. From top to bottom, examples of no/less, weak, strong bar dominance are presented, respectively. The galaxy morphology information is shown below each image.

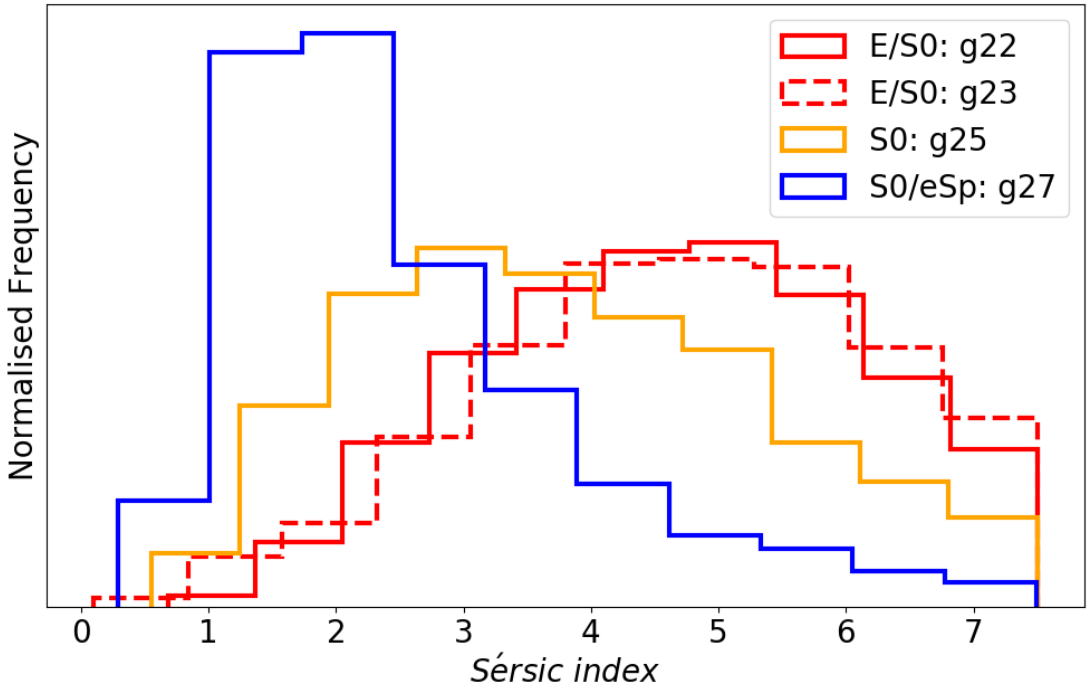

Figure 6.12: The Sérsic index distribution for the clusters dominated by $E/S0$ galaxies (g22: red solid line; g23: red dashed line), $S0$ (g25: yellow solid line), and $S0/eSp$ (g27 : blue solid line), where 'g22' represents the 22th cluster, and so on.

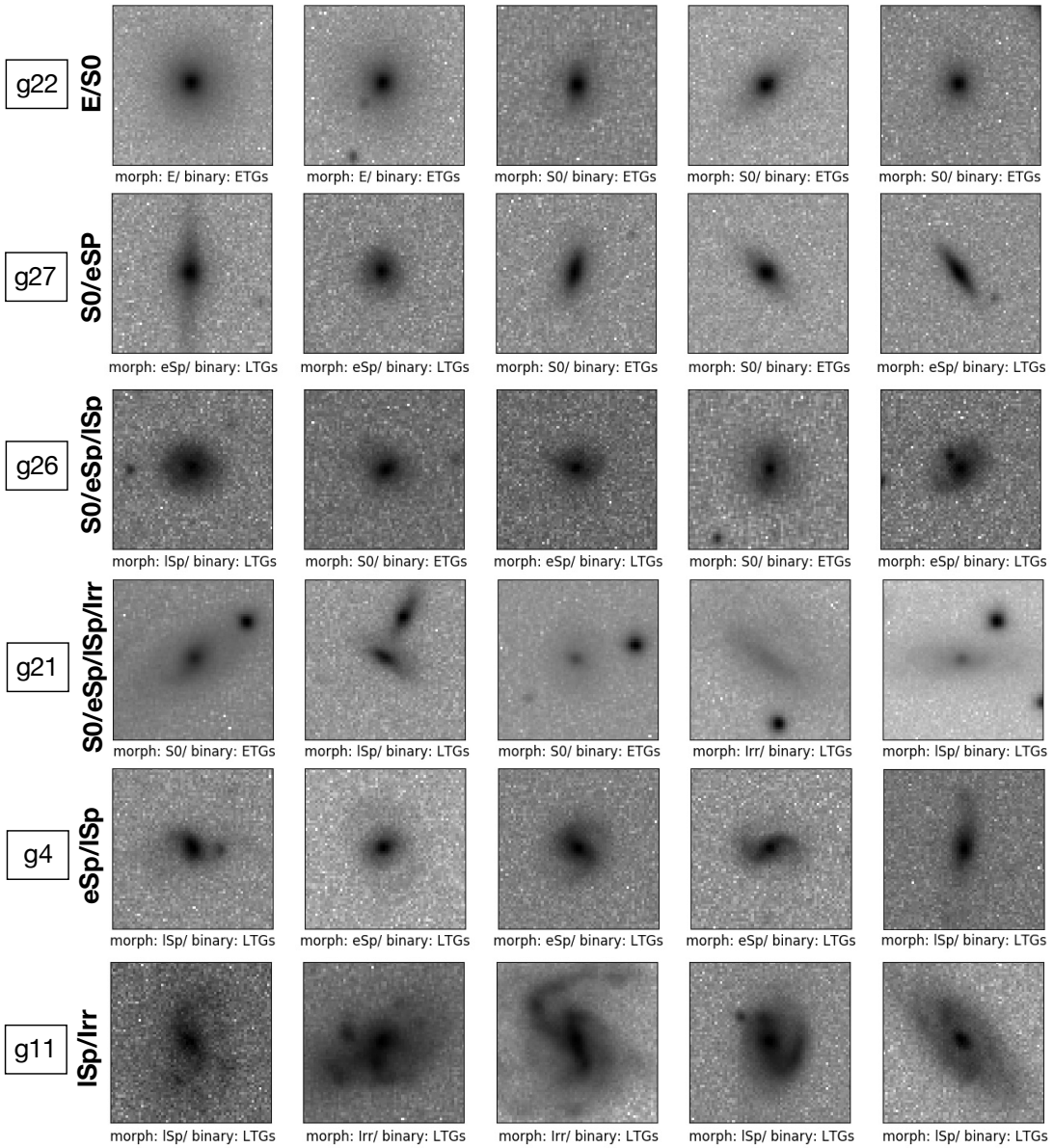

Figure 6.13: Examples of images of galaxies from clusters with a mix of many visual morphology types. Each row shows five randomly picked examples within the cluster, where 'g22' represents the 22th cluster, and so on. The morphology information is shown below each image.

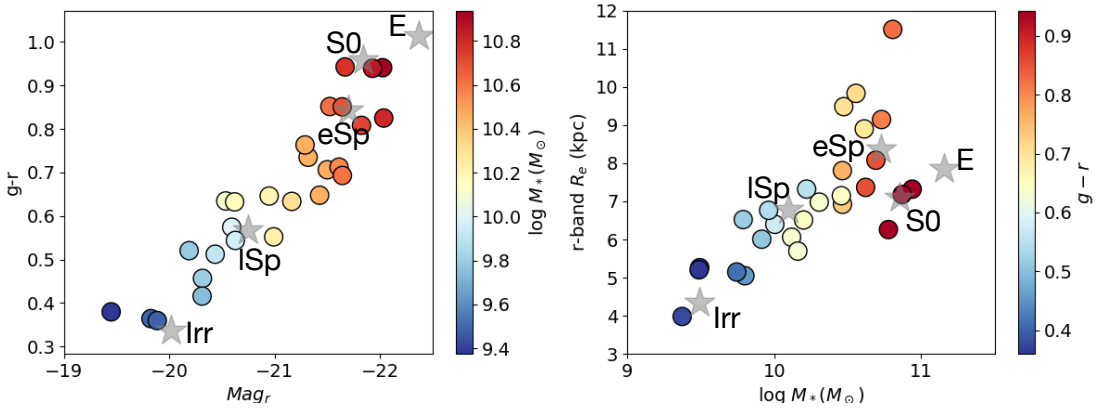

Figure 6.14: *Left:* the colour-magnitude diagram of the classification clusters where the x-axis is the average values of the r-band absolute magnitude (Mag_r) and the y-axis represents the average value of the galaxy colours ($g-r$) within each plotted cluster. Each circle represents one classification cluster from our unsupervised machine and coloured by the average value of the stellar mass (M_*). *Right:* the mass-size relation of the given clusters where the x-axis and y-axis is the average values of the stellar mass (M_*) and the average values of the galaxy physical sizes (R_e, kpc), respectively. Each circle is coloured by the average value of galaxy colour ($g-r$). In both graphs, each star shows the average values of these quantities for the traditional Hubble types for comparison, where the type of each is written in black.

Additionally, we notice on the mass-size diagram (right in Fig. 6.14) that the five orange clusters above the *eSp* star-label are dominated by barred galaxies, in particular, the top cluster with the largest average size has $\sim 80\%$ barred galaxies in the cluster (g16 in Table 6.3). Galaxies in this cluster have larger sizes, larger stellar masses, and are redder in colour than other clusters with a mix of typical spiral galaxies.

6.4.5 Dataset with a realistic distribution

To test the capability of our method on a realistic data distribution, we apply our method to the imbalanced dataset (Fig. 6.3) which follows the distribution of intrinsic morphology for nearby galaxies (Oh et al., 2013, Section 6.3.1). In this section, we examine the performance using this dataset for: (1) binary classification (Section 6.4.5.1) and (2) multiple classification clusters (Section 6.4.5.2) using the imbalanced dataset, and compare the results with the one using the balanced dataset.

6.4.5.1 Unsupervised binary classification

Similar to Section 6.4.1 for the balanced dataset, we merge the imbalanced dataset into two preliminary clusters (Example of galaxies in each is shown in Fig. 6.15). Although the imbalanced data has a significantly different distribution in galaxy

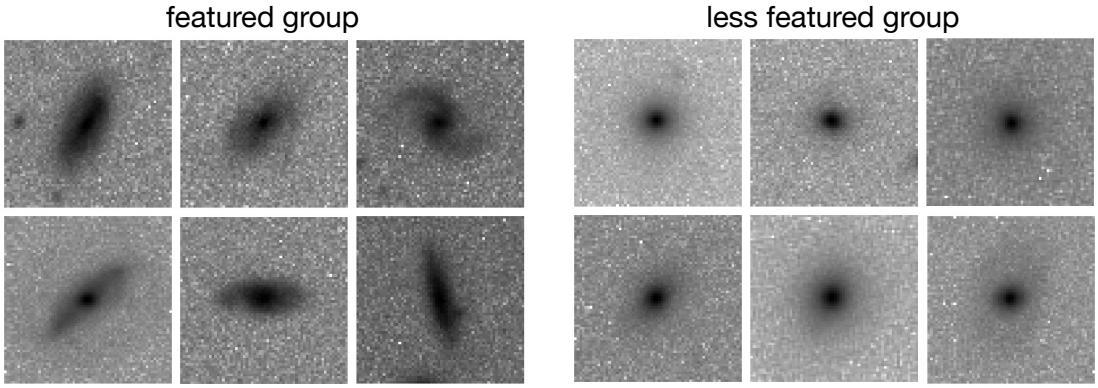

Figure 6.15: Examples of galaxies within the two preliminary clusters using the imbalanced dataset. Galaxies in one cluster are with more features (left), and galaxies in the other group are with relatively fewer features (right).

types from the balanced dataset, our machine obtains two preliminary clusters with similar features to the two clusters provided using the balanced dataset (Fig. 6.4). As before, one cluster is dominated by galaxies with many distinct features while the other has galaxies with significantly fewer features.

Fig. 6.16 shows the morphological fractions of different types (left column) and the dominance of each morphology type in each cluster (right column). The dominance is, again, the ratio between the morphological fraction in the cluster to the fraction in the dataset. This quantity removes the impact of the imbalanced numbers between each type, and indicates the visual features emphasised in a cluster. The two clusters are clearly dominated by LTGs and ETGs, respectively. Additionally, the dominance distribution of the imbalanced dataset is completely consistent with that of the balanced dataset (Fig. 6.5). This confirms that no matter which data distribution is used, our machine is capable of separating the two clusters based on the specific features existing in the corresponding morphology types.

Additionally, applying our method to the imbalanced dataset we get an initial accuracy of $\sim$0.87 in separating ETGs from LTGs. The accuracy is again defined as the number of correct matches from the total samples. The reason for a higher accuracy compared with the balanced dataset is due to a lower fraction of early spirals in the imbalanced dataset ($\sim$ 8%) than the balanced dataset ($\sim$ 25%). When we exclude the early spirals from the imbalanced dataset, the accuracy barely changes, and it is consistent with the accuracy obtained when using the balanced dataset (accuracy: $\sim$0.87; Section 6.4.1). These results show the ability of our method to achieve reliable binary morphological classification for large surveys with unknown morphological mixes.

6.4.5.2 Multiple classification clusters

Following Section 6.3.4, and using the imbalanced dataset, we obtain the same number of clusters, 27, as when we used the balanced dataset through our method

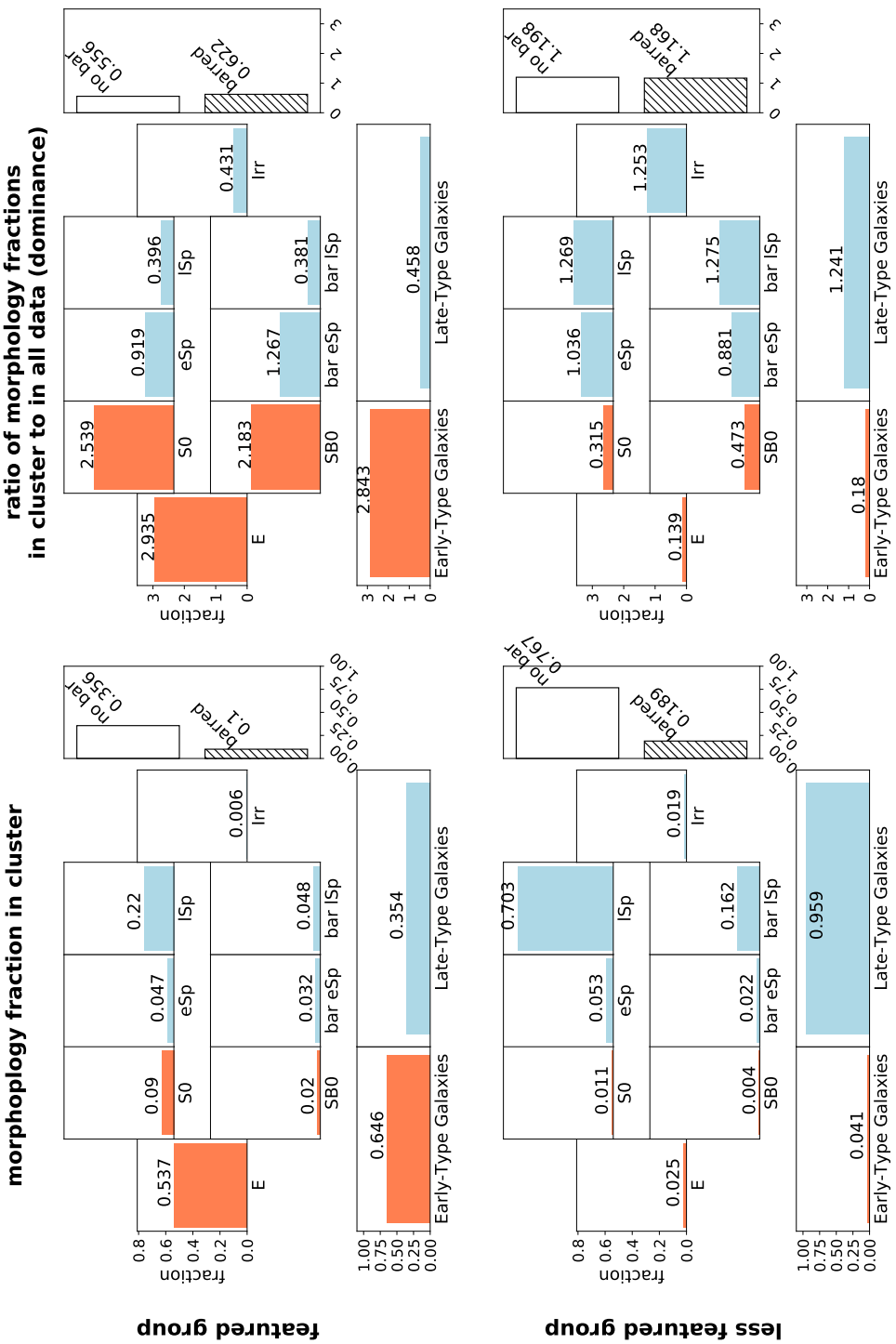

Figure 6.16: The distribution of visual galaxy morphology in each cluster obtained using the imbalanced dataset. The left column shows the fraction of each morphology type in the clusters, while the right column shows the dominance of each type. The top row shows the distribution of the 'featured group' while the bottom row presents the one of the 'less featured group'. This can be compared to the same distribution when using the balanced dataset shown in Fig. 6.5.

of determining the number of clusters (Section 6.4.2). The clustering results for both datasets are very close to each other, with only very minor differences. For example, 7 clusters are separated under the less featured group using the balanced dataset while 8 clusters are obtained using the imbalanced dataset. Conversely, we obtain 20 clusters for the featured group using the balanced dataset, and 19 using the imbalanced dataset.

In Fig. 6.17, we associate the classification clusters for the imbalanced, realistic, data set with Hubble types based on the dominance of each type. We find no clean clusters for ellipticals (E), lenticulars (S0), early spirals (eSp), irregulars (Irr) when using the imbalanced dataset. The lack of clusters for E and eSp is due to the same reasons for the balanced dataset discussed in Section 6.4.2: these two visual morphologies are intrinsically difficult to separated from other morphology types. Additionally, in Section 6.4.2, we conclude that to get a clean $S0$ cluster, galaxies have to show a moderate disk structure (Fig. 6.12). However, there is not a sufficient number of lenticulars with the relevant features due to the low fraction of this type in the imbalanced dataset (Fig. 6.3). It is impossible for the machine to classify a galaxy that does not exist in some abundence within the dataset; therefore, we miss the pure $S0$ cluster when using the imbalanced dataset. On the other hand, irregular galaxies do not have a specific structure; therefore, it is easy to be confused with some late spirals with less structured appearances by our machine, based on only galaxy structure and without the prior knowledge of 'late sprials' or 'irregulars'. They also suffer from the similar cause of the missing $S0$ cluster: the insufficient number of irregular galaxies in our imbalanced set decreases the possibility of the distinctive irregulars to be picked out by our machine.

Similar to the results of the balanced dataset, the separation between clusters might 'improve' in terms of being closer to a more physical classification when we consider colour information in our machine. Therefore, this will be an important part in future work.

6.5 Conclusion

In this chapter, we present an unsupervised machine learning technique by applying a combination of a feature extractor - a vector-quantised variational autoencoder (VQ-VAE) and a hierarchical clustering algorithm (HC). This method involves a vector quantisation process which provides a rate of classification with a feature extractor in the learning phase at least 30 times faster than a typical convolutional antoencoder used in Cheng et al. (2020b, Chapter 5) on the same device.

To sensibly explore galaxy morphologies and investigate the number of galaxy morphological classes, we propose some novel modifications to the machine learning algorithms used in this work (Section 6.2). First, we include a preliminary clustering result in the VQ-VAE architecture during the feature learning process. This helps to extract features that can not only reproduce the input images but

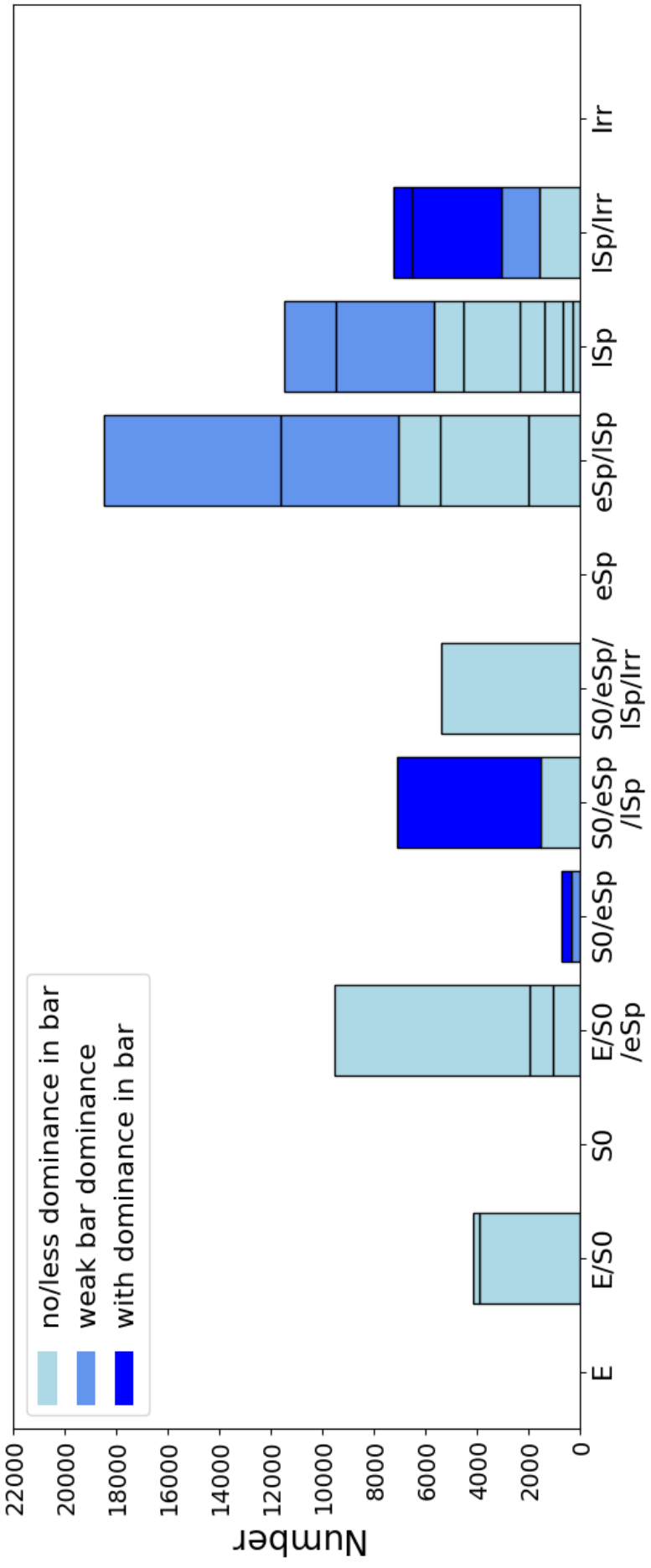

Figure 6.17: The accumulated distribution of the classification clusters compared with the Hubble sequence for the imbalanced dataset. The x-axis shows one or a mix of visual morphology types. Each coloured bar represents one cluster. Different colours are different dominance of barred galaxies in the cluster, such that from deep to light blue represent more barred galaxies to no/fewer barred galaxies in the cluster.

also be well separated into two preliminary clusters in feature space. Second, different distance thresholds are used within each branch in the merger tree in the HC process rather than a single distance threshold for a whole tree. This flexibility prevents the creation of unnecessary clusters separating galaxies with few features, while allowing more clusters for galaxies that show larger variation. Another innovation is to use galaxy orientation (a potential problem when classifying galaxies) to our advantage, helping to decide the number of clusters (Section 6.3.4).

Using the monochromatic images from the Sloan Digital Sky Survey (SDSS), we first explore galaxy classifications using a dataset with a balanced number of galaxies in each morphological class (Section 6.3.1). This is done to reduce potential biases associated with number imbalances. Using this method we obtain 27 clusters within this balanced dataset. We find that our method separates the classification clusters based on galaxy shape and structure (e.g., Sérsic index, asymmetry, Gini coefficient, M20). We then associate our classification clusters with the Hubble sequence based on the dominance of each type in a given cluster (Section 6.4.2). Clusters with barred, weak-barred, and non-barred galaxies are well distinguished by our machine. However, when using the balanced dataset, no clean clusters are found for ellipticals or early spirals (Fig. 6.10). Additionally, most clusters are associated with a mixture of Hubble types. We thus conclude that there is a fundamental difficulty in separating accurately galaxies with transitional features such as lenticular galaxies and early spirals with a machine. This applies both to visual and machine classifications.

In addition, we find that each machine classification cluster has characteristic galaxy properties (e.g., colours, masses, luminosities, sizes) that transition smoothly along the Hubble sequence.

Overall, the machine classification clusters provide a reasonable and detailed scheme for galaxy morphological classification based on a combination of multiple structural parameters, avoiding human errors and biases. The dominated features in our classification clusters can be used as the foundation of an objective alternative to the Hubble sequence. Since our system separates well galaxies with different shape, structure, and physical properties, it may prove useful in generic galaxy formation and evolution studies. The system may be improved by including multi-colour imaging and velocity maps. It will also be interesting to apply our technique to higher redshift galaxies to see how the classification cluster change.

To test the performance of our method with realistic morphological distributions, we also apply it to an imbalanced dataset which follows the morphological distribution of nearby galaxies. The results are very similar to the ones obtained with the balanced dataset, showing that our system is able to deal with large galaxy samples with more realistic morphological mixes. It also shows that our set up is not sensitive to different distributions of input galaxy morphologies, but can handle different distributions.

In the future, we intend to apply the techniques developed here to multi-colour images with better resolution such as the data from the Dark Energy Survey and the Euclid Space Telescope. Velocity maps from integral-field spectroscopic surveys could also be included. The resulting classification system(s) should prove very useful to better understand galaxy properties, their formation and evolution. We also hope the future development of this work will result in a fundamental change in how we approach galaxy morphological classification - both visually and when using machine learning.

Chapter 7

Conclusions and Future Work

In this thesis, we have demonstrated the use of both supervised and unsupervised machine learning techniques on galaxy morphological classification using imaging data. As mentioned in Section 1.2, machine learning techniques in astronomical applications can be categorised into three different stages: (1) before observation, (2) raw data, and (3) after calibration. In this thesis, we focus on applying machine learning techniques to calibrated imaging data, addressing two main topics:

- classification - we discuss an optimal machine learning technique in terms of accuracy, efficiency, and inclusiveness using imaging data for large-scale surveys;
- exploration - we explore galaxy morphological classification without human bias, and investigate a novel classification system defined by machine learning.

Through our approaches to these two topics, we provide a complete overview of galaxy morphological classification using both supervised and unsupervised machine learning methods. The conclusions for the two topics investigated in this thesis are shown in Section 7.1 and Section 7.2. Finally, future plans are presented in Section 7.3.

7.1 Automated Classifications

Along with the fast development of computational capability, astronomical observations will reach data rates of over terabyte-scale per night in the near future (e.g. Ivezić et al., 2019), and simulations output complex information also on terabyte scales (e.g. Springel et al., 2005). This officially declares that astronomical studies have stepped into the so-called 'Big Data era' (Section 1.1). To analyse such a vast amount of complex data produced in astronomical surveys and simulations, machine learning techniques are introduced to a variety of astronomical analyses. In this thesis, we concentrate on an accurate, efficient, and inclusive classification task.

Classification tasks are commonly approached by supervised machine learning where we train machines with labels involving human judgement. In Chapter 2, we introduced a variety of supervised machine learning techniques (Table 2.1) to

classify galaxies from the Dark Energy Survey (DES) Year 1 (Y1) imaging data into either Ellipticals or Spirals. A complete comparison in the accuracy and efficiency of each supervised method was carried out in that chapter. Additionally, we inspected the impact of rotated images, the balance in number between the target types, and the number of images used in training (Chapter 3). We concluded that (1) using rotated images for the data augmentation does not cause biases; (2) the balance in the number of data between the target types is important, and a balanced dataset shows a better performance than an imbalanced one; (3) more training data helps the performance, but the relative improvement decreases when the number increases. Meanwhile, we conclude that there is a significant improvement when using gradient images (specifically, the Histogram of Oriented Gradient technique) in most supervised methods with imaging data.

Convolutional neural networks (CNN) are the most optimal method within the ten supervised methods tested using imaging data. In Chapter 3, we further investigate the CNN trained with a combination of linear and gradient images from the Dark Energy Survey (DES) Year 1 (Y1) data and the labels provided from the Galaxy Zoo 1 (GZ1) catalogue. The better resolution (0.″263 per pixel) and greater depth ($i = 22.51$) of DES reveal a few incorrect GZ1 classifications based on data from the Sloan Digital Sky Survey (SDSS). After correcting these labels, our CNN reaches an accuracy of over 0.99 in the binary classification for Ellipticals and Spirals. We then apply this setup to the DES Year 3 (Y3) data and provide one of the largest galaxy morphological classification catalogue to date which includes over 20 million galaxies (Chapter 4).

In our studies, supervised methods can reach great accuracy with high efficiency (Table 2.3). However, supervised machine learning has a potential inclusivity issue: it may encounter problems in classifying galaxies that are significantly different from the galaxies in the training set or are not clearly defined in the training set. We notice in Chapter 4 that regardless of the image quality (e.g., signal-to-noise ratio and resolution), the CNN predictions show a better performance on classifying a certain type of galaxies that exists in the training set (Fig. 4.10). For example, a CNN model trained with bright galaxies at a low redshift shows a better prediction for fainter galaxies than for bright galaxies at a higher redshift. Our result resonated with the *The Elephant in the Room* from Rosenfeld et al. (2018) that supervised machine learning fails to correctly classify objects if the test conditions do not exist in the training set. Second, supervised machine learning classifies based on the provided labels; therefore, it lacks the flexibility to fairly distribute weights to unknown patterns. In Chapter 3, we notice that for difficult galaxies such as lenticulars, which have no clear definition provided by GZ1 labels, our CNN generally give low predicted probabilities. Although this result indicates that supervised machine learning can be used to provide a label that does not exist in the training by giving an appropriate probability threshold, supervised methods cannot broadly explore unknown features. In future surveys, we cannot guarantee that the current human knowledge of galaxy morphology has covered well what we might obverse when more and more galaxies are revealed. Therefore, unsupervised machine learning techniques which have no (or less) need for humans' involvement are applied to classification tasks.

As a bridge, we start an unsupervised machine learning application in identifying galaxy-galaxy strong lensing systems (GGSLs) using the simulated data with the image quality similar to that of the Euclid from the Strong Gravitational Lens Finding Challenge (Lens Finding Challenge; Metcalf et al., 2019a) in Chapter 5. This project has three main advantages for investigating the capability of unsupervised machine learning: (1) the GGSLs have distinctive features such as Einstein rings and arc structures; (2) simulated data are less complex than observed data; (3) the Lens Finding Challenge provides a complete comparison between our work and other supervised methods. We are not only the first research group introducing an unsupervised machine learning technique to classify lensing systems, but also the first one in astronomy proposing the method used in our work that combines a convolutional autoencoder for feature learning and a Bayesian Gaussian mixture model for clustering.

In Chapter 5, we prove the ability of this setup to capture representatively structural features from images and separate the images into several sensible clusters (24 clusters in this work). They distinguish different types of lensing systems such as different Einstein ring sizes and different arc structures (Fig. 5.9 and Fig. 5.16 to Fig. 5.17). With fewer clusters, separated using visual structures, our unsupervised machine can be used as a preliminary classification process to group images with similar features for large surveys. Compared with supervised methods, our method picks up ~ 63 percent of lensing images from all lenses in the training set. Additionally, with the assumed probability proposed in Chapter 5, we reach an accuracy of $77.3 \pm 0.5\%$ in the binary classification of lensing and non-lensing systems. Although our unsupervised method shows less accuracy than most supervised methods (Table 5.2), it is of great importance to note that (1) an unsupervised machine is fundamentally different from a supervised machine in a variety of aspects, e.g., weights allocation. Without the assistance (or contamination) of human labelling, an unsupervised machine provides a less accurate but more objective judgement to classify target objects. (2) the 'accuracy' here is measured by comparing the predictions with human-defined labels; therefore, there is an intrinsic unfairness to compare the performance between unsupervised and supervised methods. (3) Furthermore, human-defined labels have an intrinsic bias and broader inclusion due to the subjective judgement and enormous background knowledge taken into account beneath the given decision.

Due to the reasons discussed above, we suggest to apply unsupervised machine learning techniques (1) to a simple classification task (e.g., binary or ternary); (2) to give a preliminary categorisation based on visual features for large-scale data; (3) to explore data without human bias. The first two approaches are discussed in Chapter 5, and the last task is carried out in Chapter 6 for galaxy morphological classification.

7.2 Galaxy Morphology without Human Bias

In Chapter 6, we improve the technique proposed in Chapter 5 and apply it to explore galaxy morphology using the Sloan Digital Sky Survey imaging data.

The improved unsupervised machine includes a vector-quantised variation autoencoder (VQ-VAE) for feature learning and a hierarchical clustering (HC) algorithm. The vector quantisation process applied in the VQ-VAE makes the feature learning phase at least 30 times faster than a typical convolutional autoencoder (e.g., Chapter 5). The agglomerative HC (Johnson, 1967) algorithm has no need for any presumption on the distribution of the obtained clusters. Additionally, to sensibly explore galaxy morphologies we included three strategies: (1) to consider a preliminary cluster result in the VQ-VAE when extracting features from images; (2) to allow different distance thresholds used to define clusters in each branch in the HC process; (3) to use the feature of galaxy orientation, which can potentially be a problem in unsupervised machine learning applications, to decide the optimal number of clusters (Section 6.3). The strategies applied to our unsupervised machine result in 27 classification clusters. The clusters are separated based on galaxy shape and structure presented by structural measurements such as the Sérsic index, asymmetry, Gini coefficient, M20. Additionally, we confirm that regardless of the galaxy morphology distribution in the dataset, our unsupervised machine captures consistent features. This characteristic makes our unsupervised methods very useful for large astronomical surveys.

Our method provides 27 preliminary classes for further visual assessment. This unsupervised method significantly accelerates the classification process. Moreover, our unsupervised machine reaches an accuracy of ~ 0.87 for binary classification of early-type (ETGs) and late-type galaxies (LTGs) when we categorise galaxies in an imbalanced dataset, which includes 23% ETGs and 77% LTGS, into only two clusters.

To explore the machine-defined classes, we examined the galaxy properties, and compared them with Hubble types. First, the machine-defined classes show a clear separation between barred and non-barred galaxies that indicates a distinctive difference in structures shown between the two visual types. Second, each machine-defined morphology class is distinctive in a variety of stellar properties such as colour, r-band absolute magnitude, and stellar mass of galaxies. Additionally, the machine morphological classes show a clear transition along with the Hubble types on both the colour-magnitude plane and mass-size plane (Fig. 6.14). This suggests that more morphologies with distinctive galaxy properties can be distinguished from the basic Hubble types such as ellipticals, lenticulars, early spirals, late spirals and irregulars.

7.2.1 Defects in the Visual Classification systems

Visual classification systems such as the Hubble sequence are of great importance in categorising galaxies. However, visual classification systems are defined by humans who might provide less precise decision boundaries when separating different galaxy morphologies.

In Chapter 6, with the unsupervised machine, we find an intrinsic difficulty to accurately classify galaxies with transitional features such as lenticulars and early-type spirals. To associate our given clusters by our unsupervised machine

with Hubble types, we find that no clean cluster is dominated by E or eSp. This is because the structures of ellipticals have a great similarity to lenticular galaxies; meanwhile, lenticular galaxies also share similar structure to eSp. We find that only lenticular galaxies with a particular parameter, e.g. Sérsic index peak at ~ 3, can be distinguished from other visual morphology types (Fig. 6.12).

On the other hand, the eSp types have a broad range of visual features that also exist between $S0$ and lSp. This causes a difficulty to classify eSp into either ETGs or LTGs, or to have a clean classification cluster by our machine.

Moreover, we notice that galaxies with a similar structure can be classified into a variety of visual morphology types; in other words, a mix of galaxy structure can exist in one visual morphology type. Therefore, we conclude that there is an intrinsic uncertainty in any visual classification schemes such as the Hubble sequence.

7.2.2 A Novel Galaxy Classification System by Machine?

In the previous section (see details in Chapter 6), we state the conclusion that the visual morphological classification scheme is not precisely defined. A variety of alternative classification systems such as *CAS systems* can be used to provide an objective morphological classification of galaxies. However, with an unsupervised machine learning technique, we can define the decision boundaries between classifications in high-dimensional feature space that considers galaxy structures, light distribution, galaxy shapes, and other potential factors such as colour and velocity.

In this section, we propose to rethink the visual morphological classification scheme we have known for a century, and to approach related studies with machine-defined morphology classes. The classes suggested by machine learning can possibly provide a more 'accurate' definition in galaxy morphology than a visual classification scheme. For example, in Chapter 6, 27 machine-defined classes have distinctive stellar properties from each other. Additionally, they show a clear transition on both the colour-magnitude diagram and the mass-size relation. This suggests that our machine classifications can be used to develop a novel objectively morphological classification scheme. With this machine-defined scheme, we can re-approach studies of galaxy evolution and formation from a different perspective.

7.3 Future Plans

The work in this thesis could be improved, extended, and followed-up in a variety of ways. For example, throughout this thesis, we focus only on the shape and structure of galaxies using both supervised and unsupervised machine learning methods. Additional progress could be made using images in multiple photometric bands. Additionally, the supervised machine learning binary classification work could be extended using a finer set of morphological classes.

In addition to direct classification tasks using supervised and unsupervised machine learning techniques, we are especially interested in future applications of unsupervised techniques to explore galaxy morphology further. In Section 7.2.2, we presented a machine-defined morphological classification system showing good correlation with the stellar properties of galaxies. Exploring other galaxy properties such as environment, metallicity, and star formation rate would be interesting. We also look forward to applying this unsupervised method to other large surveys with better spatial resolution, such as the Dark Energy Survey and the Hyper Suprime-Cam Subaru Strategic Program. Future surveys such as the Large Synoptic Survey Telescope, the Euclid Space Telescope, etc, will provide further datasets to exploit.

Additionally, a correlation of the classification scheme with redshifts will also be interesting to investigate. For example, does the optimal number of classification clusters suggested by the machine change with redshifts? How well can our method classify galaxies at a higher redshift?

We can foresee plenty of studies extending the work in this thesis. Machine learning techniques are developing fast in a variety of astronomical applications; therefore, in addition to galaxy morphology, the methods developed in this thesis can be extended to different astronomical data such as spectroscopy, or other astronomical objects, such as galaxy clusters and lensing systems, and even to different astronomical challenges such as anomaly detection. It will be interesting to explore some of these.

References

Aaronson, M. (1978). The morphological distribution of bright galaxies in the UVK color plane. , 221:L103–L107.

Abadi, M., Agarwal, A., Barham, P., et al. (2015a). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abadi, M., Agarwal, A., Barham, P., et al. (2015b). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. (2009). The Seventh Data Release of the Sloan Digital Sky Survey. , 182(2):543–558.

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. (2018). The Dark Energy Survey: Data Release 1. , 239(2):18.

Abraham, R. G., van den Bergh, S., and Nair, P. (2003). A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. , 588(1):218–229.

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1988). A Learning Algorithm for Boltzmann Machines, page 285–307. Ablex Publishing Corp., USA.

Al-Rfou, R., Alain, G., Almahairi, A., et al. (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints, abs/1605.02688.

Arcelin, B., Doux, C., Aubourg, E., and Roucelle, C. (2020). Deblending galaxies with Variational Autoencoders: a joint multi-band, multi-instrument approach. arXiv e-prints, page arXiv:2005.12039.

Attias, H. (2000). A variational bayesian framework for graphical models. In In Advances in Neural Information Processing Systems 12, pages 209–215. MIT Press.

Avestruz, C., Li, N., Zhu, H., Lightman, M., Collett, T. E., and Luo, W. (2019a). Automated Lensing Learner: Automated Strong Lensing Identification with a Computer Vision Technique. , 877(1):58.

Avestruz, C., Li, N., Zhu, H., Lightman, M., Collett, T. E., and Luo, W. (2019b). Automated Lensing Learner: Automated Strong Lensing Identification with a Computer Vision Technique. , 877:58.

Baillard, A., Bertin, E., de Lapparent, V., et al. (2011). The EFIGI catalogue of 4458 nearby galaxies with detailed morphology. , 532:A74.

Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. (2004). Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. , 600(2):681–694.

Ball, N. M. and Brunner, R. J. (2010). Data Mining and Machine Learning in Astronomy. International Journal of Modern Physics D, 19(7):1049–1106.

Ball, N. M., Brunner, R. J., Myers, A. D., et al. (2007). Robust Machine Learning Applied to Astronomical Data Sets. II. Quantifying Photometric Redshifts for Quasars Using Instance-based Learning. , 663(2):774–780.

Ball, N. M., Brunner, R. J., Myers, A. D., and Tcheng, D. (2006). Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. , 650(1):497–509.

Ball, N. M., Loveday, J., Fukugita, M., et al. (2004). Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. , 348(3):1038–1046.

Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. (2009). Galaxy Zoo: the dependence of morphology and colour on environment*. , 393(4):1324–1352.

Banerji, M., Lahav, O., Lintott, C. J., et al. (2010). Galaxy Zoo: reproducing galaxy morphologies via machine learning. , 406(1):342–353.

Baron, D. (2019). Machine Learning in Astronomy: a practical overview. arXiv e-prints, page arXiv:1904.07248.

Baron, D. and Poznanski, D. (2017). The weirdest SDSS galaxies: results from an outlier detection algorithm. , 465(4):4530–4555.

Bautista, M. Á., Sanakoyeu, A., Sutter, E., and Ommer, B. (2016). Cliquecnn: Deep unsupervised exemplar learning. CoRR, abs/1608.08792.

Bayer, D., Chatterjee, S., Koopmans, L. V. E., et al. (2018). Observational constraints on the sub-galactic matter-power spectrum from galaxy-galaxy strong gravitational lensing. arXiv e-prints, page arXiv:1803.05952.

Bayliss, M. B., Sharon, K., Acharyya, A., et al. (2017). Spatially Resolved Patchy Lyα Emission within the Central Kiloparsec of a Strongly Lensed Quasar Host Galaxy at z = 2.8. , 845:L14.

Beck, M. R., Scarlata, C., Fortson, L. F., et al. (2018). Integrating human and machine intelligence in galaxy morphology classification tasks. , 476(4):5516–5534.

Bellman, R. (1954). The theory of dynamic programming. Bull. Amer. Math. Soc., 60(6):503–515.

Bershady, M. A., Jangren, A., and Conselice, C. J. (2000). Structural and Photometric Classification of Galaxies. I. Calibration Based on a Nearby Galaxy Sample. , 119(6):2645–2663.

Bertin, E. and Arnouts, S. (1996). SExtractor: Software for source extraction. , 117:393–404.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Bom, C. R., Makler, M., Albuquerque, M. P., and Brandt, C. H. (2017). A neural network gravitational arc finder based on the Mediatrix filamentation method. , 597:A135.

Bonjean, V., Aghanim, N., Salomé, P., Beelen, A., Douspis, M., and Soubrié, E. (2019). Star formation rates and stellar masses from machine learning. , 622:A137.

Borji, A. and Dundar, A. (2017). A new look at clustering through the lens of deep convolutional neural networks. CoRR, abs/1706.05048.

Bottrell, C., Hani, M. H., Teimoorinia, H., et al. (2019). Deep learning predictions of galaxy merger stage and the importance of observational realism. , 490(4):5390–5413.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., and Song, A. (2015). Efficient agglomerative hierarchical clustering. Expert Syst. Appl., 42(5):2785–2797.

Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A., and Lemson, G. (2009). Resolving cosmic structure formation with the Millennium-II Simulation. , 398(3):1150–1164.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7):1145 – 1159.

Breen, P. G., Foley, C. N., Boekholt, T., and Portegies Zwart, S. (2020). Newton versus the machine: solving the chaotic three-body problem using deep neural networks. , 494(2):2465–2470.

Breiman, L. (2001). Random forests. Mach. Learn., 45(1):5–32.

Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. (2019). Deblending and classifying astronomical sources with Mask R-CNN deep learning. , 490(3):3952–3965.

Calderon, V. F. and Berlind, A. A. (2019). Prediction of galaxy halo masses in SDSS DR7 via a machine learning approach. , 490(2):2367–2379.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. CoRR, abs/1807.05520.

Carrasco Kind, M. and Brunner, R. J. (2014). SOMz: photometric redshift PDFs with self-organizing maps and random atlas. , 438:3409–3421.

CarrascoKind, M. and Brunner, R. J. (2014). SOMz: photometric redshift PDFs with self-organizing maps and random atlas. , 438(4):3409–3421.

Cavuoti, S., Amaro, V., Brescia, M., Vellucci, C., Tortora, C., and Longo, G. (2017). METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts. , 465:1959–1973.

Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. (2020a). Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. , 493(3):4209–4228.

Cheng, T.-Y., Li, N., Conselice, C. J., Aragón-Salamanca, A., Dye, S., and Metcalf, R. B. (2020b). Identifying strong lenses with unsupervised machine learning using convolutional autoencoder. , 494(3):3750–3765.

Chester, C. and Roberts, M. S. (1964). Properties of Galaxies: color-magnitude diagram. , 69:635.

Chopra, P. and Yadav, S. (2017). Restricted boltzmann machine and softmax regression for fault detection and classification. Complex Intelligent Systems, pages 1–11.

Chou, F.-C. (2014). Galaxy Zoo Challenge: Classify Galaxy Morphologies from Images.

Collett, T. E. (2015). The Population of GalaxyGalaxy Strong Lenses in Forthcoming Optical Imaging Surveys. , 811:20.

Collett, T. E. and Auger, M. W. (2014). Cosmological constraints from the double source plane lens SDSSJ0946+1006. , 443:969–976.

Conselice, C. J. (2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. , 147(1):1–28.

Conselice, C. J. (2006). The fundamental properties of galaxies and a new galaxy classification system. , 373(4):1389–1408.

Conselice, C. J., Bershady, M. A., and Jangren, A. (2000). The Asymmetry of Galaxies: Physical Morphology for Nearby and High-Redshift Galaxies. , 529(2):886–910.

Conselice, C. J., Blackburne, J. A., and Papovich, C. (2005). The Luminosity, Stellar Mass, and Number Density Evolution of Field Galaxies of Known Morphology from z = 0.5 to 3. , 620(2):564–583.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. In Machine Learning, pages 273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1):21–27.

Cunningham, P. and Delany, S. J. (2007). k-nearest neighbour classifiers.

D'Abrusco, R., Fabbiano, G., Djorgovski, G., Donalek, C., Laurino, O., and Longo, G. (2012). CLaSPS: A New Methodology for Knowledge Extraction from Complex Astronomical Data Sets. , 755:92.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1.

de la Calleja, J. and Fuentes, O. (2004). Machine learning and image analysis for morphological galaxy classification. , 349(1):87–93.

de Vaucouleurs, G. (1948). Recherches sur les Nebuleuses Extragalactiques. Annales d'Astrophysique, 11:247.

de Vaucouleurs, G. (1959). Classification and Morphology of External Galaxies. Handbuch der Physik, 53:275.

de Vaucouleurs, G. (1961). Integrated Colors of Bright Galaxies in the u, b, V System. , 5:233.

de Vaucouleurs, G. (1964). Luminosity Classification of Galaxies and Some Applications. , 69:561.

de Vaucouleurs, G., de Vaucouleurs, A., and Corwin, H. G. (1995a). VizieR Online Data Catalog: RC2 Catalogue (de Vaucouleurs+ 1976). VizieR Online Data Catalog, page VII/112.

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Buta, R. J., Paturel, G., and Fouque, P. (1995b). VizieR Online Data Catalog: Third Reference Cat. of Bright Galaxies (RC3) (de Vaucouleurs+ 1991). VizieR Online Data Catalog, page VII/155.

DeepMind (2018). Sonnet, url = https://github.com/deepmind/sonnet.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 39(1):1–38.

DES Collaboration (2005). The Dark Energy Survey. arXiv e-prints, pages astro–ph/0510346.

DES Collaboration, Abbott, T., Abdalla, F. B., et al. (2016). The Dark Energy Survey: more than dark energy - an overview. , 460(2):1270–1299.

Dieleman, S., Schlüter, J., Raffel, C., et al. (2015). Lasagne: First release.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. , 450(2):1441–1459.

D'Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts. , 609:A111.

Dizaji, K. G., Herandi, A., and Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. CoRR, abs/1704.06327.

Dodge, S. and Karam, L. (2016). Understanding how image quality affects deep neural networks. In 2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016, 2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016. Institute of Electrical and Electronics Engineers Inc. 8th International Conference on Quality of Multimedia Experience, QoMEX 2016 ; Conference date: 06-06-2016 Through 08-06-2016.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., and Fischer, J. L. (2018). Improving galaxy morphologies for SDSS with Deep Learning. , 476(3):3661–3676.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M. A., and Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. CoRR, abs/1406.6909.

Dressler, A. (1980). Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies. , 236:351–365.

Drlica-Wagner, A., Sevilla-Noarbe, I., Rykoff, E. S., et al. (2018). Dark Energy Survey Year 1 Results: The Photometric Data Set for Cosmology. , 235(2):33.

Dubath, P., Rimoldini, L., Süveges, M., et al. (2011). Random forest automated supervised classification of Hipparcos periodic variable stars. , 414(3):2602–2617.

Dundar, A., Jin, J., and Culurciello, E. (2015). Convolutional clustering for unsupervised learning. CoRR, abs/1511.06241.

Dye, S., Furlanetto, C., Dunne, L., et al. (2018). Modelling high-resolution ALMA observations of strongly lensed highly star-forming galaxies detected by Herschel. , 476:4383–4394.

Elmegreen, D. M. and Elmegreen, B. G. (1982). Flocculent and grand design spiral structure in field, binary and group galaxies. , 201:1021–1034.

Elmegreen, D. M. and Elmegreen, B. G. (1987). Arm Classifications for Spiral Galaxies. , 314:3.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, pages 226–231. AAAI Press.

Fawagreh, K., Gaber, M. M., and Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering, 2(1):602–609.

Fawcett, T. (2006). An introduction to roc analysis. Pattern Recognition Letters, 27(8):861 – 874. ROC Analysis in Pattern Recognition.

Ferreira, L., Conselice, C. J., Duncan, K., Cheng, T.-Y., Griffiths, A., and Whitney, A. (2020). Galaxy Merger Rates up to z $\sim$ 3 Using a Bayesian Deep Learning Model: A Major-merger Classifier Using IllustrisTNG Simulation Data. , 895(2):115.

Fix, E. and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review / Revue Internationale de Statistique, 57(3):238–247.

Flaugher, B., Diehl, H. T., Honscheid, K., et al. (2015). The Dark Energy Camera. , 150(5):150.

Fritzke, B. (1994). A growing neural gas network learns topologies. In Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94, pages 625–632, Cambridge, MA, USA. MIT Press.

Fritzke, B. (1995). A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, Advances in Neural Information Processing Systems 7, pages 625–632. MIT Press.

Fukugita, M., Nakamura, O., Okamura, S., et al. (2007). A Catalog of Morphologically Classified Galaxies from the Sloan Digital Sky Survey: North Equatorial Region. , 134(2):579–593.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. Biological Cybernetics, 20:121–136.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193–202.

Fukushima, K., Miyake, S., and Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(5):826–834.

Fustes, D., Manteiga, M., Dafonte, C., et al. (2013). An approach to the analysis of SDSS spectroscopic outliers based on self-organizing maps. Designing the outlier analysis software package for the next Gaia survey. , 559:A7.

Gao, D., Zhang, Y.-X., and Zhao, Y.-H. (2008). Support vector machines and kd-tree for separating quasars from large survey data bases. , 386(3):1417–1425.

Gavazzi, R., Marshall, P. J., Treu, T., and Sonnenfeld, A. (2014). RINGFINDER: Automated Detection of Galaxy-scale Gravitational Lenses in Ground-based Multi-filter Imaging Data. , 785:144.

Geach, J. E. (2012). Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys. , 419:2633–2645.

Giles, D. and Walkowicz, L. (2019). Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection. , 484(1):834–849.

Gilman, D., Birrer, S., Treu, T., Keeton, C. R., and Nierenberg, A. (2018). Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. , 481(1):819–834.

Goderya, S. N. and Lolling, S. M. (2002). Morphological Classification of Galaxies using Computer Vision and Artificial Neural Networks: A Computational Scheme. , 279(4):377–387.

Grazian, A., Fontana, A., De Santis, C., Gallozzi, S., Giallongo, E., and Di Pangrazio, F. (2004). The Large Binocular Camera Image Simulator. , 116(822):750–761.

Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). Deep clustering with convolutional autoencoders. In ICONIP.

Hambleton, K. M., Gibson, B. K., Brook, C. B., et al. (2011). Advanced morphological galaxy classification: a comparison of observed and simulated galaxies. , 418(2):801–810.

Hamming, R. W. (1950). Error detecting and error correcting codes. The Bell System Technical Journal, 29(2):147–160.

Hartley, H. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14(2):174–194. doi:10.2307/2527783.

Hartley, P., Flamary, R., Jackson, N., Tagore, A. S., and Metcalf, R. B. (2017). Support vector machine classification of strong gravitational lenses. , 471:3378–3397.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, 2nd Edition. Springer series in statistics. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

He, S., Li, Y., Feng, Y., et al. (2019). Learning to predict the cosmological structure formation. Proceedings of the National Academy of Science, 116(28):13825–13832.

Hernández-Toledo, H. M., Vázquez-Mata, J. A., Martínez-Vázquez, L. A., et al. (2008). A Morphological Re-Evaluation of Galaxies in Common from the Catalog of Isolated Galaxies and the Sloan Digital Sky Survey (DR6). , 136(5):2115–2135.

Hershey, J. R., Chen, Z., Roux, J. L., and Watanabe, S. (2015). Deep clustering: Discriminative embeddings for segmentation and separation. CoRR, abs/1508.04306.

Hewitt, J. N., Turner, E. L., Schneider, D. P., Burke, B. F., and Langston, G. I. (1988). Unusual radio source MG1131+0456 - A possible Einstein ring. , 333:537–540.

Hezaveh, Y. D., Dalal, N., Marrone, D. P., et al. (2016). Detection of Lensing Substructure Using ALMA Observations of the Dusty Galaxy SDP.81. , 823:37.

Hezaveh, Y. D., Levasseur, L. P., and Marshall, P. J. (2017). Fast automated analysis of strong gravitational lenses with convolutional neural networks. , 548:555–557.

Higson, E., Handley, W., Hobson, M., and Lasenby, A. (2019). Bayesian sparse reconstruction: a brute-force approach to astronomical imaging and machine learning. , 483(4):4828–4846.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. Neural Comput., 14(8):1771–1800.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Comput., 9(8):1735–1780.

Hocking, A., Geach, J. E., Sun, Y., and Davey, N. (2018). An automatic taxonomy of galaxy morphology using unsupervised machine learning. , 473:1108–1129.

Holmberg, E. (1958). A photographic photometry of extragalactic nebulae. Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136:1.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Hsu, Y. and Kira, Z. (2015). Neural network-based clustering using pairwise constraints. CoRR, abs/1511.06321.

Hubble, E. P. (1926). Extragalactic nebulae. , 64:321–369.

Hubble, E. P. (1936). Realm of the Nebulae.

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., and Sánchez Almeida, J. (2011). Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification. , 525:A157.

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. (2015). A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning. , 221(1):8.

Huertas-Company, M., Primack, J. R., Dekel, A., et al. (2018). Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range. , 858(2):114.

Huertas-Company, M., Rodriguez-Gomez, V., Nelson, D., et al. (2019). The Hubble Sequence at z $\sim$ 0 in the IllustrisTNG simulation with deep learning. , 489(2):1859–1879.

Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., and Le Fèvre, O. (2008). A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. I. Method description. , 478(3):971–980.

Huertas-Company, M., Tasca, L., Rouan, D., et al. (2009). A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. II. Quantifying morphological k-correction in the COSMOS field at 1 < z < 2: Ks band vs. I band. , 497(3):743–753.

Huppenkothen, D., Heil, L. M., Hogg, D. W., and Mueller, A. (2017). Using machine learning to explore the long-term evolution of GRS 1915+105. , 466(2):2364–2377.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. (2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873(2):111.

Jacobs, C., Collett, T., Glazebrook, K., et al. (2019). An Extended Catalog of Galaxy-Galaxy Strong Gravitational Lenses Discovered in DES Using Convolutional Neural Networks. , 243(1):17.

Jacobs, C., Glazebrook, K., Collett, T., More, A., and McCarthy, C. (2017). Finding strong lenses in CFHTLS using convolutional neural networks. , 471(1):167–181.

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32:241–254.

Joseph, R., Courbin, F., Metcalf, R. B., et al. (2014). A PCA-based automated finder for galaxy-scale strong lenses. , 566:A63.

Kamble, P. M. and Hegadi, R. S. (2015). Handwritten marathi character recognition using r-hog feature. Procedia Computer Science, 45:266 – 274. International Conference on Advanced Computing Technologies and Applications (ICACTA).

Keogh, E. and Mueen, A. (2017). Curse of Dimensionality, pages 314–315. Springer US, Boston, MA.

Kim, E. J. and Brunner, R. J. (2017). Star-galaxy classification using deep convolutional neural networks. , 464(4):4463–4475.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv e-prints, page arXiv:1312.6114.

Kodi Ramanah, D., Charnock, T., Villaescusa-Navarro, F., and Wandelt, B. D. (2020). Super-resolution emulator of cosmological simulations using deep physical models. , 495(4):4227–4236.

Kohonen, T., editor (1997). Self-organizing Maps. Springer-Verlag, Berlin, Heidelberg.

Kovács, A. and Szapudi, I. (2015). Star-galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogues. , 448(2):1305–1313.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

Krone-Martins, A. and Moitinho, A. (2014). UPMASK: unsupervised photometric membership assignment in stellar clusters. , 561:A57.

Kügler, S. D., Polsterer, K., and Hoecker, M. (2015). Determining spectroscopic redshifts by using k nearest neighbor regression. I. Description of method and analysis. , 576:A132.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Ann. Math. Statist., 22(1):79–86.

Kummer, J., Kahlhoefer, F., and Schmidt-Hoberg, K. (2018). Effective description of dark matter self-interactions in small dark matter haloes. , 474:388–399.

Küng, R., Saha, P., Ferreras, I., et al. (2018). Models of gravitational lens candidates from Space Warps CFHTLS. , 474:3700–3713.

Lahav, O., Naim, A., Sodré, L., J., and Storrie-Lombardi, M. C. (1996). Neural computation as a tool for galaxy classification: methods and examples. , 283:207.

Lanusse, F., Ma, Q., Li, N., et al. (2018). CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding. , 473(3):3895–3906.

Laureijs, R., Amiaux, J., Arduini, S., et al. (2011). Euclid Definition Study Report. arXiv e-prints, page arXiv:1110.3193.

Law, D. R., Steidel, C. C., Erb, D. K., et al. (2007). The Physical Nature of Rest-UV Galaxy Morphology during the Peak Epoch of Galaxy Formation. , 656(1):1–26.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

Li, F., Qiao, H., Zhang, B., and Xi, X. (2017). Discriminatively boosted image clustering with fully convolutional auto-encoders. CoRR, abs/1703.07980.

Li, F. F., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence, 28:594–611.

Lintott, C., Schawinski, K., Bamford, S., et al. (2011). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. , 410(1):166–178.

Lintott, C. J., Schawinski, K., Slosar, A., et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. , 389(3):1179–1189.

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., and Winter, M. K. (2016). Photometric Supernova Classification with Machine Learning. , 225:31.

Lotz, J. M., Davis, M., Faber, S. M., et al. (2008). The Evolution of Galaxy Mergers and Morphology at z < 1.2 in the Extended Groth Strip. , 672(1):177–197.

Lotz, J. M., Primack, J., and Madau, P. (2004). A New Nonparametric Approach to Galaxy Morphological Classification. , 128(1):163–182.

Madgwick, D. S. (2003). Correlating galaxy morphologies and spectra in the 2dF Galaxy Redshift Survey. , 338(1):197–207.

Maehoenen, P. H. and Hakala, P. J. (1995). Automated Source Classification Using a Kohonen Network. , 452:L77.

Mandelbaum, R., Rowe, B., Bosch, J., et al. (2014). The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook. , 212(1):5.

Maranzana, F. E. (1963). On the location of supply points to minimize transportation costs. IBM Syst. J., 2(2):129–135.

Margalef-Bentabol, B., Huertas-Company, M., Charnock, T., et al. (2020). Detecting outliers in astronomical images with deep generative networks. arXiv e-prints, page arXiv:2003.08263.

Marshall, P. J., Hogg, D. W., Moustakas, L. A., et al. (2009). Automated Detection of Galaxy-Scale Gravitational Lenses in High-Resolution Imaging Data. , 694:924–942.

Martin, D. C., Fanson, J., Schiminovich, D., et al. (2005). The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. , 619(1):L1–L6.

Martin, G., Kaviraj, S., Hocking, A., Read, S. C., and Geach, J. E. (2019). Galaxy morphological classification in deep-wide surveys via unsupervised machine learning. arXiv e-prints, page arXiv:1909.10537.

Martin, G., Kaviraj, S., Hocking, A., Read, S. C., and Geach, J. E. (2020). Galaxy morphological classification in deep-wide surveys via unsupervised machine learning. , 491(1):1408–1426.

Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I, ICANN'11, pages 52–59, Berlin, Heidelberg. Springer-Verlag.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, Second Edition. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

McLachlan, G. and Krishnan, T. (1997). The EM algorithm and extensions. Wiley, New York.

Meert, A., Vikram, V., and Bernardi, M. (2015). A catalogue of 2D photometric decompositions in the SDSS-DR7 spectroscopic main galaxy sample: preferred models and systematics. , 446:3943–3974.

Mendel, J. T., Simard, L., Palmer, M., Ellison, S. L., and Patton, D. R. (2014). A Catalog of Bulge, Disk, and Total Stellar Mass Estimates for the Sloan Digital Sky Survey. , 210(1):3.

Meneghetti, M., Melchior, P., Grazian, A., et al. (2008). Realistic simulations of gravitational lensing by galaxy clusters: extracting arc parameters from mock DUNE images. , 482:403–418.

Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. (2019a). The strong gravitational lens finding challenge. , 625:A119.

Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. (2019b). The strong gravitational lens finding challenge. A&A, 625:A119.

Morgan, W. W. (1962). Some Characteristics of Galaxies. , 135:1.

Morgan, W. W. and Mayall, N. U. (1957). A Spectral Classification of Galaxies. , 69(409):291.

Mustafa, M., Bard, D., Bhimji, W., Lukić, Z., Al-Rfou, R., and Kratochvil, J. M. (2019). CosmoGAN: creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks. Computational Astrophysics and Cosmology, 6(1):1.

Naim, A., Lahav, O., Sodre, L., J., and Storrie-Lombardi, M. C. (1995). Automated morphological classification of APM galaxies by supervised artificial neural networks. , 275(3):567–590.

Nair, P. B. and Abraham, R. G. (2010). A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey. , 186(2):427–456.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, pages 807–814, USA. Omnipress.

Nouri, D. (2014). nolearn: scikit-learn compatible neural network library.

Ntampaka, M., Trac, H., Sutherland, D. J., Battaglia, N., Póczos, B., and Schneider, J. (2015). A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters. , 803(2):50.

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., and Zumach, W. A. (1992). Automated Star/Galaxy Discrimination With Neural Networks. , 103:318.

Oh, K., Choi, H., Kim, H.-G., Moon, J.-S., and Yi, S. K. (2013). Demographics of Sloan Digital Sky Survey Galaxies along the Hubble Sequence. , 146(6):151.

Orr, M. J. L. and Science, C. F. C. (1996). Introduction to radial basis function networks. Technical report.

Ostrovski, F., McMahon, R. G., Connolly, A. J., et al. (2017). VDES J2325-5229 a z = 2.7 gravitationally lensed quasar discovered using morphology-independent supervised machine learning. , 465:4325–4334.

Paraficz, D., Courbin, F., Tramacere, A., et al. (2016). The PCA Lens-Finder: application to CFHTLS. , 592:A75.

Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. Expert Syst. Appl., 36(2):3336–3341.

Pearson, W. J., Wang, L., Trayford, J. W., Petrillo, C. E., and van der Tak, F. F. S. (2019). Identifying galaxy mergers in observations and simulations with deep learning. , 626:A49.

Pearson, J., Li, N., and Dye, S. (2019). The use of convolutional neural networks for modelling large optically-selected strong galaxy-lens samples. , 488:991–1004.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Peng, C. Y., Ho, L. C., Impey, C. D., and Rix, H.-W. (2010). Detailed Decomposition of Galaxy Images. II. Beyond Axisymmetric Models. , 139(6):2097–2129.

Perraudin, N., Srivastava, A., Lucchi, A., Kacprzak, T., Hofmann, T., and Réfrégier, A. (2019). Cosmological N-body simulations: a challenge for scalable generative models. Computational Astrophysics and Cosmology, 6(1):5.

Petrillo, C. E., Tortora, C., Chatterjee, S., et al. (2017). Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks. , 472(1):1129–1150.

Polsterer, K. L., Gieseke, F., and Kramer, O. (2012). Galaxy Classification without Feature Extraction, volume 461 of Astronomical Society of the Pacific Conference Series, page 561.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1):37–63.

Rana, A., Jain, D., Mahajan, S., Mukherjee, A., and Holanda, R. F. L. (2017). Probing the cosmic distance duality relation using time delay lenses. , 7:010.

Razavi, A., van den Oord, A., and Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. arXiv e-prints, page arXiv:1906.00446.

Reiman, D. M. and Göhre, B. E. (2019). Deblending galaxy superpositions with branched generative adversarial networks. , 485(2):2617–2627.

Rix, H.-W. and Zaritsky, D. (1995). Nonaxisymmetric Structures in the Stellar Disks of Galaxies. , 447:82.

Rodríguez, A. C., Kacprzak, T., Lucchi, A., et al. (2018). Fast cosmic web simulations with generative adversarial networks. Computational Astrophysics and Cosmology, 5(1):4.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, pages 65–386.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. (2018). The Elephant in the Room. arXiv e-prints, page arXiv:1808.03305.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53 – 65.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. , 323(6088):533–536.

Sadeh, I., Abdalla, F. B., and Lahav, O. (2016). ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. , 128(10):104502.

Salakhutdinov, R. and Hinton, G. (2009). Deep boltzmann machines. In van Dyk, D. and Welling, M., editors, Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, volume 5 of Proceedings of Machine Learning Research, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In Proceedings of the 24th International Conference on Machine Learning, ICML '07, page 791–798, New York, NY, USA. Association for Computing Machinery.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM J. Res. Dev., 3(3):210–229.

Samui, S. and Samui Pal, S. (2017). Photo-z with CuBANz: An improved photometric redshift estimator using Clustering aided Back propagation Neural network. , 51:169–177.

Sandage, A. (1961). The Hubble Atlas of Galaxies.

Scarlata, C., Carollo, C. M., Lilly, S., et al. (2007a). COSMOS Morphological Classification with the Zurich Estimator of Structural Types (ZEST) and the Evolution Since z = 1 of the Luminosity Function of Early, Disk, and Irregular Galaxies. , 172(1):406–433.

Scarlata, C., Carollo, C. M., Lilly, S. J., et al. (2007b). The Redshift Evolution of Early-Type Galaxies in COSMOS: Do Massive Early-Type Galaxies Form by Dry Mergers? , 172(1):494–510.

Scholkopf, B. and Smola, A. J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681.

Sérsic, J. L. (1963). Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6:41–43.

Sérsic, J. L. (1968). Atlas de Galaxias Australes.

Shamir, L. (2009). Automatic morphological classification of galaxy images. , 399(3):1367–1372.

Sharda, P., Federrath, C., daÂ Cunha, E., Swinbank, A. M., and Dye, S. (2018). Testing star formation laws in a starburst galaxy at redshift 3 resolved with ALMA. , 477:4380–4390.

Short, R. D. and Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. IEEE Trans. Information Theory, 27:622–626.

Shu, C., Ding, X., and Fang, C. (2011). Histogram of the oriented gradient for face recognition. Tsinghua Science and Technology, 16(2):216–224.

Shu, Y., Bolton, A. S., Mao, S., et al. (2016a). The BOSS Emission-line Lens Survey. IV. Smooth Lens Models for the BELLS GALLERY Sample. , 833:264.

Shu, Y., Bolton, A. S., Moustakas, L. A., et al. (2016b). Kiloparsec Mass/Light Offsets in the Galaxy Pair-Lyα Emitter Lens System SDSS J1011+0143. , 820:43.

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., and McConnachie, A. W. (2011). A Catalog of Bulge+disk Decompositions and Updated Photometry for 1.12 Million Galaxies in the Sloan Digital Sky Survey. , 196(1):11.

Siudek, M., Małek, K., Pollo, A., et al. (2018a). The VIMOS Public Extragalactic Redshift Survey (VIPERS). Unsupervised classification with photometric redshifts: a method to accurately classify large galaxy samples without spectroscopic information. arXiv e-prints.

Siudek, M., Małek, K., Pollo, A., et al. (2018b). The VIMOS Public Extragalactic Redshift Survey (VIPERS). The complexity of galaxy populations at 0.4 z 1.3 revealed with unsupervised machine-learning algorithms. , 617:A70.

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. Journal of Business Research, 70:263 – 286.

Smolensky, P. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory, page 194–281. MIT Press, Cambridge, MA, USA.

Soler, J. D., Beuther, H., Rugel, M., et al. (2019). Histogram of oriented gradients: a technique for the study of molecular cloud formation. , 622:A166.

Sonnenfeld, A., Chan, J. H. H., Shu, Y., et al. (2018). Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). I. Automatic search for galaxy-scale strong lenses. , 70:S29.

Sonnenfeld, A., Treu, T., Marshall, P. J., et al. (2015). The SL2S Galaxy-scale Lens Sample. V. Dark Matter Halos and Stellar IMF of Massive Early-type Galaxies Out to Redshift 0.8. , 800:94.

Springel, V., White, S. D. M., Jenkins, A., et al. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. , 435(7042):629–636.

Sreejith, S., Pereverzyev, Sergiy, J., Kelvin, L. S., et al. (2018). Galaxy And Mass Assembly: automatic morphological classification of galaxies using statistical learning. , 474(4):5232–5258.

Storrie-Lombardi, M. C., Lahav, O., Sodre, L., J., and Storrie-Lombardi, L. J. (1992). Morphological Classification of Galaxies by Artificial Neural Networks. , 259:8P.

Strateva, I., Ivezić, Ž., Knapp, G. R., et al. (2001). Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. , 122(4):1861–1874.

Suyu, S. H., Bonvin, V., Courbin, F., et al. (2017). H0LiCOW - I. H_0 Lenses in COSMOGRAIL's Wellspring: program overview. , 468:2590–2604.

Tarsitano, F., Hartley, W. G., Amara, A., et al. (2018). A catalogue of structural and morphological measurements for DES Y1. , 481(2):2018–2040.

Tuccillo, D., Huertas-Company, M., Decencière, E., and Velasco-Forero, S. (2017). Deep learning for studies of galaxy morphology. In Brescia, M., Djorgovski, S. G., Feigelson, E. D., Longo, G., and Cavuoti, S., editors, Astroinformatics, volume 325 of IAU Symposium, pages 191–196.

Tuccillo, D., Huertas-Company, M., Decencière, E., Velasco-Forero, S., Domínguez Sánchez, H., and Dimauro, P. (2018). Deep learning for galaxy surface brightness profile fitting. , 475(1):894–909.

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236):433–460.

Vafaei Sadr, A., Vos, E. E., Bassett, B. A., Hosenie, Z., Oozeer, N., and Lochner, M. (2019). DEEPSOURCE: point source detection using deep learning. , 484(2):2793–2806.

van den Bergh, S. (1960). A Preliminary Luminosity Clssification of Late-Type Galaxies. , 131:215.

van den Bergh, S. (1976). A new classification system for galaxies. , 206:883–887.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. (2017). Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., et al., editors, Advances in Neural Information Processing Systems 30, pages 6306–6315. Curran Associates, Inc.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg.

Vegetti, S., Koopmans, L. V. E., Auger, M. W., Treu, T., and Bolton, A. S. (2014). Inference of the cold dark matter substructure mass function at z = 0.2 using strong gravitational lenses. , 442:2017–2035.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., 11:3371–3408.

Walmsley, M., Smith, L., Lintott, C., et al. (2020). Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. , 491(2):1554–1574.

Way, M. J. and Klose, C. D. (2012). Can Self-Organizing Maps Accurately Predict Photometric Redshifts? , 124:274.

Weir, N., Fayyad, U. M., and Djorgovski, S. (1995). Automated Star/Galaxy Classification for Digitized Poss-II. , 109:2401.

Werbos, P. and John, P. (1974). Beyond regression : new tools for prediction and analysis in the behavioral sciences /.

Whitmore, B. C. (1984). An objective classification system for spiral galaxies. I. The two dominant dimensions. , 278:61–80.

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. , 435(4):2835–2860.

Wu, G. and Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning. In In ICML 2003 Workshop on Learning from Imbalanced Data Sets, pages 49–56.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pages 478–487. JMLR.org.

Xiong, L., Poczos, B., Connolly, A., and Schneider, J. (2018). Anomaly detection for astronomical data.

York, D. G., Adelman, J., Anderson, John E., J., et al. (2000). The Sloan Digital Sky Survey: Technical Summary. , 120(3):1579–1587.

Zamojski, M. A., Schiminovich, D., Rich, R. M., et al. (2007). Deep GALEX Imaging of the COSMOS HST Field: A First Look at the Morphology of z ~0.7 Star-forming Galaxies. , 172(1):468–493.

Zanaty, E. (2012). Support vector machines (svms) versus multilayer perception (mlp) in data classification. Egyptian Informatics Journal, 13(3):177 – 183.

Zhang, Y. and Zhao, Y. (2015). Astronomy in the Big Data Era. Data Science Journal, 14:11.