NATIONAL CENTRE FOR NUCLEAR RESEARCH

DOCTORAL THESIS

Search for galaxy mergers in big sky surveys

Author:
Luis E. SUELVES

Supervisor:
Agnieszka POLLO
Auxiliary supervisor:
William J.PEARSON

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Astrophysics division (BP4)

June 16, 2024

Declaration of Authorship

I, Luis E. SUELVES, declare that this thesis titled, "Search for galaxy mergers in big sky surveys" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the National Centre for Nuclear Research.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at the National Centre for Nuclear Research or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"“Why do we fight, Kal, why do we keep going?”
“I don’t know, I’ve forgotten.”
“It’s so we can be with each other.”
“They all die, Tien. Everyone dies.”
“So they do, don’t they?”
“That means it doesn’t matter, none of it matters.”
“See, that’s the wrong way of looking at it. Since we all go to the same place in the end, the moments we spent with each other are the only things that do matter. The times we helped each other.” ”

Brandon Sanderson, Rhythm of War

NATIONAL CENTRE FOR NUCLEAR RESEARCH

Abstract

Search for galaxy mergers in big sky surveys

Luis E. SUELVES

During their lifecycle and evolution, galaxies can approach each other and collide, becoming some of the most impressive views in the sky. These are galaxy mergers, as their interaction can result in two or more galaxies merging into one. These galaxies show a high variability of morphological distortions due to the tidal forces arising during the process. Their aspect depends on the relative masses, the stage within the process, the line-of-sight projection in the sky, the brightness, distance, and size of the merging members, and the properties of the observation and instruments.

This thesis covers the studies I carried out to classify galaxy mergers in the big datasets obtained from wide sky observations. The surveys I worked with are the Sloan Digital Sky Survey (SDSS) and the deeper Subaru/Hyper Suprime-Cam (HSC). Both preceded the upcoming Large Survey of Space and Time (LSST), which will be carried by the Vera Rubin Observatory. LSST will capture in one night the same amount of data that SDSS collected over nearly a decade. The steep increase of data size from survey to survey has made it imperative to find automatized techniques to treat the data. For that, I worked on dimensionality reduction methods, on the image calibration, and on Machine Learning (ML) techniques such as Neural Networks (NN) or dimensionality reduction methods.

In this thesis, I describe the discovery and development of a new potential methodology to identify galaxy mergers in large surveys. This method is based on the effect that the surrounding of galaxies has in the sky background analysis of the astronomical image. It was discovered when testing the performance of a NN using only photometric information. We trained it on a class-balanced dataset of mergers and non-mergers built out of SDSS galaxies, classified visually by online volunteers in the Galaxy Zoo Data Release 1 (GZ DR1). Testing multiple combinations of the photometric parameters as NN inputs led us to find how the SDSS DR6 sky background error was capable of tracing galaxy mergers with a training-set accuracy of 92.64 $\pm$ 0.15 % test-set accuracy of 92.36 $\pm$ 0.21 %. Moreover, studying the sky error plane formed by the g and r bands revealed that a decision boundary line is enough to achieve an accuracy of 91.59%. The interpretation of this result is that the sky background error is tracing low signal-to-noise features around the observed galaxies.

The success of the sky error boundary led us to test its extension to a wider set of galaxies from SDSS DR6. This was in fact the whole GZ DR1 set. We studied the presence of merging galaxies of various types in different regions of the diagram. It was also found that non-merging galaxies with nearby stars and non-interacting galaxies could appear in the regions populated by mergers. In order to avoid this contaminated non-mergers, we built a decision tree that discarded galaxies with nearby stars or galaxies too far to be potentially colliding. This provided a 67.07 % of dirty non-mergers that were successfully discarded, and a 72.44 % percent of mergers that were correctly retained. Thus, further tailoring of our approach with focus on reducing the contamination would make the boundary capable of finding SDSS sources.

Finally, we tested how to implement the sky error method in the deeper HSC images. We used images on the North Ecliptic Pole (NEP), available through the AKARI-NEP

collaboration. In this field, I joined the ML-based search for mergers in the HSC images of the North Ecliptic Pole (NEP) led by my auxiliary supervisor dr. William J. Pearson. We confirmed visually the merger candidates identified by the model and published the first catalogue of mergers in the NEP. The Galaxy Zoo: Cosmic Dawn! program provided a follow-up set of morphological classifications, from which I selected a sample for the HSC sky error extension. Within an aperture around each source, I estimated the background and obtained the distribution of low Signal-to-noise pixels. Because the sky error in SDSS is computed during the sky background measurement, I also created my own data reduction, making sure the Low Surface-Brightness (LSB) features from the tidally striped material around the mergers are not lost. I compared the resulting images with and without subtraction of the dithering-based background model, and concluded it does not eliminate the LSB features in the image. I calculated multiple parameters on the LSB pixels, observed the parameter space through dimensionality reduction methods, and found that they have the potential of creating a new area where mergers can be found.

Overall, in this thesis I demonstrate that galaxy mergers can be identified in large sky surveys by the effect that the low signal-to-noise stripped material has around them in the images.

NATIONAL CENTRE FOR NUCLEAR RESEARCH

Streszczenie

Search for galaxy mergers in big sky surveys

Luis E. SUELVES

Podczas swojego cyklu życia i ewolucji galaktyki mogą zbliżać się do siebie i zderzać, tworząc jedne z najbardziej imponujących widoków na niebie. Interakcja galaktyk może skutkować połączeniem się dwóch lub więcej galaktyk w jedną. Galaktyki w trakcie zderzeń wykazują dużą rozmaitość zniekształceń morfologicznych spowodowanych siłami pływowymi powstającymi podczas tego procesu. Wygląd układu galaktyk w trakcie zderzenia zależy od ich stosunku mas, etapu procesu, orientacji względem linii widzenia, jasności, odległości i rozmiaru łączących się galaktyk, ale także właściwości teleskopów i kamer.

Niniejsza rozprawa obejmuje badania, które przeprowadziłem w celu wykrywania i klasyfikowania zderzeń galaktyk w dużych zbiorach danych uzyskanych z obserwacji dużych obszarów nieba. Przeglądy, z którymi pracowałem, to Sloan Digital Sky Survey (SDSS) i głębszy Subaru/Hyper Suprime-Cam (HSC). Oba poprzedzają nadchodzący Large Survey of Space and Time (LSST), który będzie prowadzony przez Vera Rubin Observatory i który w ciągu jednej nocy będzie zbierał taką samą ilość danych, jaką SDSS zebrał przez prawie dekadę. Gwałtowny wzrost rozmiaru danych z dekady na dekadę sprawa, że konieczne stało się rozwinięcie zautomatyzowanych technik przetwarzania danych. W tym celu pracowałem nad metodami kalibracji obrazu i technikami uczenia maszynowego (ML), takimi jak sieci neuronowe (NN) lub metody redukcji wymiarowości.

W rozprawie opisuję odkrycie i rozwój nowej potencjalnej metodologii identyfikacji zderzeń galaktyk w dużych przeglądach. Metoda ta opiera się na wpływie, jaki otoczenie zderzających się galaktyk ma na tło nieba na zdjęciu astronomicznym. Efekt związany z tłem został odkryty podczas testowania wydajności sieci neuronowej wykorzystującej wyłącznie informacje fotometryczne. Trenowaliśmy sieć na zrównoważonym klasowo zbiorze danych galaktyk łączących się i niełączących się, stworzonym z galaktyk SDSS, sklasyfikowanych wizualnie przez ochotników w Galaxy Zoo Data Release 1 (GZ DR1). Testowanie wielu kombinacji parametrów fotometrycznych jako danych wejściowych NN doprowadziło nas do odkrycia, w jaki sposób błąd tła nieba SDSS DR6 był w stanie śledzić fuzje galaktyk z dokładnością dla 92,64 $\pm$ 0,15 % zestawu treningowego i z dokładnością 92,36 $\pm$ 0,21 % dla zestawu testowego. Co więcej, badanie płaszczyzny błędu nieba utworzonej przez pasma g i r ujawniło, że graniczna linia decyzyjna jest wystarczająca do osiągnięcia dokładności 91,59%. Interpretacja tego wyniku jest taka, że błąd tła nieba jest wrażliwy na źródła o niskim stosunku sygnału do szumu wokół obserwowanych galaktyk, w tym wypadku ogony pływowe zderzających się galaktyk.

Sukces metody opartej na błędzie tła nieba skłonił nas do przetestowania jej rozszerzenia na szerszy zbiór galaktyk z SDSS DR6. W rzeczywistości był to cały zbiór GZ DR1. Zbadałem obecność łączących się galaktyk różnych typów w różnych regionach diagramu. Okazało się, że galaktyki nie łączące się z pobliskimi gwiazdami i galaktyki nie oddziałujące ze sobą mogą również pojawiać się w regionach zajmowanych przez galaktyki łączące się. Aby uniknąć tych zanieczyszczeń, zbudowałem drzewo decyzyjne, które odrzuca galaktyki z pobliskimi gwiazdami lub galaktyki zbyt odległe, aby mogły potencjalnie się zderzyć. Zapewniło to 67,07 % brudnych nie-zderzeń, które zostały pomyślnie odrzucone, oraz 72,44 % procent zderzeń, które zostały poprawnie sklasyfikowane. Zatem dalsze dostosowanie

tego podejścia zoptymalizowane w kierunku zmniejszenia zanieczyszczeń sprawiłoby, że granica byłaby w stanie prawidłowo sklasyfikować źródła SDSS.

Na koniec przetestowaliśmy, jak wdrożyć metodę błędu nieba na głębszych obrazach HSC. Wykorzystaliśmy obrazy północnego bieguna ekliptycznego (NEP), dostępne w ramach współpracy AKARI-NEP. W tej dziedzinie dołączyłem do opartego na ML wyszukiwania zderzeń na obrazach HSC bieguna północnej ekliptyki (NEP), prowadzonego przez mojego promotora pomocniczego dr. Williama J. Pearsona. Potwierdziliśmy wizualnie kandydatury na zderzające się galaktyki zidentyfikowane przez model i opublikowaliśmy pierwszy katalog zderzeń w NEP. Program Galaxy Zoo: Cosmic Dawn! dostarczył kolejny zestaw klasyfikacji morfologicznych, z których wybrałem próbkę do rozszerzenia błędu nieba HSC. W obrębie apertury wokół każdego źródła oszacowałem tło i uzyskałem rozkład pikseli o niskim stosunku sygnału do szumu. Ponieważ błąd nieba w SDSS jest obliczany podczas pomiaru tła nieba, skonstruowałem również własną metodę redukcji danych, upewniając się, że cechy niskiej jasności powierzchniowej (LSB) z materiału pasów pływowych wokół zderzających się galaktyk nie zostaną utracone. Porównałem obrazy z różnymi podejściami do problemu odjęcia tla, dzięki czemu doszedłem do wniosku, że metoda stosowana standardowo dla danych HSC zachowuje elementy o niskiej jasności powierzchniowej w stopniu wystarczającym do wykrywania zderzeń galaktyk. Obliczyłem wiele parametrów dla pikseli LSB, zbadałem przestrzeń parametrów za pomocą metod redukcji wymiarowości i stwierdziłem, że mają one potencjał utworzenia nowego uproszczonego obszaru, w którym można znaleźć zderzenia galaktyk.

Podsumowując, w rozprawie wykazałem, że zderzenia galaktyk mogą być identyfikowane w dużych przeglądach nieba dzięki efektowi, jaki wywiera na tło nieba wokół nich materiał o niskim stosunku sygnału do szumu.

Acknowledgements

I feel really proud to present here my PhD thesis. It covers three years and a half of research, and of course it has been an endeavour that would have never been possible without many people that have followed along.

First of all goes my thesis supervisor prof. Agnieszka Pollo. She trusted me with the position on the summer of 2020, with the Covid pandemic around and with a project that I was quite excited for. Thanks to her, I have managed to learn how to make science, finding both the fun and the important aspects of my research. Thanks to her patience this thesis have a great shape and flow that I hope it is appreciated by the reader.

Next is my great auxiliary supervisor dr. William J. Pearson. While I did not expect to work with him when I applied, if anything has been done in this thesis, it is because of him. Thanks to the many hours we spent together trying to understand results, mainly managing to make any sense of whatever I try to say when I do not understand something. It made it possible that now I can show the models, analysis, results, and interpretation.

Both of them have been a great support not only in science but at a personal level. I really hope our scientific paths cross again in the future.

At the NCBJ institute I have had a lot of support from all my fellow students, colleagues, and professors. Thanks to Margherita and the three years in which we have shared office, chats, and house – well that more like two and a half I think. Our house and time there, together with Tizi, has been a perfect place for fulfilling all of this. Thanks also to the great crew of students with who we started together, sharing the closed first semester and the many changes over the years. In no specific order, thanks to Maitrayee, Hareesh, Yashwanth, Nora, Victor, Mateusz, and Michał. It has been a lot of learning, stress, and fun! Specially Yashwanth, thanks for the great effort we made together going to the library, getting the papers, and supporting our frustrations. And Hareesh, for being a great companion in understanding many aspect applying machine learning to astronomy together. Also thanks to Maitrayee and all the support she gave me in difficult moments, I really appreacited it. And I cannot leave Monika Demwoska, her support has been crucial in so many ways. From a personal side, understanding my problems, and always helping to feel better, but also for explaining million of times how to make the same papers.

At the astrophysics group, the worldwide known NCBJ BP4, I need to thank my partners in the office for their patience with my grumpiness regarding noise and people chatting or meeting in the office. Besides, it was always nice to chat, complain, and learn together. Thanks, Prasad, Krzysiek, Kishan, and of course again Hareesh and Margherita. I want to thank also to all the great BP4 members that made this journey easier in many moments such as bike rides, lunches, dinners, or conferences. I won't name everybody, but it has been a real pleasure. At the rest of the institute, I also want to thank other students in the Graduate School, the Graduate School organization itself, and the professors and scientists that contributed to it.

Now I need to definitely thank my family. None of this would have been even close to be possible without them. Thanks to my parents for listening to me so many days after work, dealing sometimes with my tension release, or my struggles with talking after a tiring day. Thanks to my sister for how we have both understood each other better and better during these years. Maybe sometimes we jumped at some things, but I appreciate all the support and understanding I received, and I hope I have been able to show it myself. Even in the distance, a big part of all the steps I have walked in Poland have been supported by the three of them.

In the outside world, I have enjoyed the warmth of many friends in Warsaw and all around the world. Those colleagues we met at conferences and summer schools, thank you for the great moments and to contribute to such comfortable scientific community. Some

were great friends in the city with who we hanged out, enjoyed great moments, movies, walks, futbol games in the river, or discovered the many corners of this city. I am going to miss you all a lot, starting all over again in a new place is always tricky, but you all made it quite easy and I am happy for all our time together. For sure we will make as many plans as possible when we are all around. Besides, I want to thank Paulina, because many of the things I managed during my PhD would have never been possible without her support. I also want to include all the good friends that accompanied me in my previous steps in Huesca, at Pedro Cerbuna in Zaragoza, and in Bonn, I would have never reached Warsaw if it was not for them.

Regarding the work presented in these pages, I want to thank the contribution of many professionals and institutions. In particular, the analysis of the background subtraction model on the Subaru/Hyper Suprime-Cam (HSC) images would have never been possible without the help of the HSC Software helpdesk and the SMOKA helpdesk. I thank the HSC software helpdesk team for their helpful advice, specially dr. Hiroki Onozato, who played a crucial role in understanding our goal and in narrowing down the solution. I also want to thank the help of the SMOKA helpdesk for helping in accessing all the necessary frames.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The thesis documented in the following pages focuses on galaxy mergers and their identification in modern astronomical surveys. The underlying theme is the astrophysics of Galaxies, their evolution and how they transform through these merging interactions. Astrophysics as a field itself attempts to understand the mechanisms and processes governing every object of the sky. Galaxies are one of the most essential types among them: they delineate the Universe's large-scale structure, as its main building blocks, and at the same time they are composed by the smaller astrophysical elements we know from our own Milky Way: stars, planets, black holes, gas, etc. The evolution and processes in galaxies depend on the dynamics of their inner components, their whole aspect, and their surrounding environment.

1.1 Galaxies

It can be considered that the concept of galaxies was initially coined after the – nowadays known as – "Great Debate" in 1920, where the scientists Harlow Shapley and Heber D. Curtis held opposing positions about the nature of the spiral nebulae observed at that time (e.g. Hoskin, 1976; Smith and Berendzen, 1982; Trimble, 1995, among many). The centre of the debate was not just the nature of those puzzling objects, but the scale of the Universe itself. Many points were made by the two scientists and the audience, based on the evidence they had at that time, their own research, and their backgrounds. A somewhat definitive conclusion on this mystery around the spiral nebulae was given by the discovery of Cepheid-type variable stars in the M31 Andromeda nebula by Edwin Hubble. These are stars that emit oscillating energy levels in periodic intervals, where the period and the total emitted energy are related to each other by what is known as period-luminosity relation (Leavitt and Pickering, 1912). Due to the nature of light propagation, given our detected energies and the known emission, their actual distance can be calculated by the relation between the two energies, (see Fernie, 1969, for a historic review of this relation). The Hubble (1925)'s catalogue of Cepheids in both M31 and M33 confirmed, through the light curves published in Hubble (1926) and Hubble (1929), respectively, that both objects are located much further than the Magellanic clouds, becoming the conclusive evidence that those two nebulae are "island universes" as our own Milky Way.

Galaxy science took off at that time and has gone a long way since then. These universe islands have been found to be composed by the many of the components that we find in our Milky Way, ranging from gas-rich star forming regions to the central Super Massive

Black Holes (SMBH). Despite they have shown to be much smaller than the separation between them, that does not imply galaxies do not interact with each other but quite the opposite. In this introduction, I will attempt to review the most relevant aspects regarding galaxy components, morphologies, and evolution models that are currently known and that are in the edge of the understanding of the astrophysics community.

1.1.1 Components of galaxies

Present day models indicate that a galaxy is mainly formed by dark matter, gas, dust, and stars, in varying amounts and proportions.

1.1.1.1 Dark Matter

Currently, the most widely accepted cosmological model of our Universe is the ΛCDM model, and one of its most fundamental aspects is that the components of galaxies are bound together due to the gravitational well of their host dark matter halo (White and Rees, 1978). Dark matter cannot be observed directly due to its lack of interaction with light, and while this makes it technically impossible to observe directly, there are many indications of its existence due to its effect on the mass distribution and the dynamics of galaxies. The first hints of its presence in fact appeared in galaxy studies, in the separate discoveries made by Fritz Zwicky and Vera Rubin.

Fritz Zwicky analysed the dynamics of the Coma Cluster by assuming the cluster to be in equilibrium and applying the Virial theorem (Zwicky, 1937). He then calculated the virial mass considering a spherically symmetric cluster and the line-of-sight proper velocities to be the galaxies average velocities. Finally, comparing the obtained mass with the observed luminosity as a factor of solar luminosities provided a mass-to-light ratio of 500, which we currently know to be wrong, but that was impossible to reconcile in the absence of dark matter. In the case of Vera Rubin, an analogous lack-of-light appeared when comparing the rotational curve in our Milky Way and in other spiral galaxies with the stellar distribution. Both distributions did not match in the outskirts of the galaxy: while the light distribution decreases towards the outer radii, the rotational curve remains flat (Sofue and Rubin, 2001). In both studies, the mass attributed to the estimated dynamics was much larger than the luminosity observed, indicating the presence of some mass-producing gravitational field but no light emission.

The cosmological and galactic evolution models quickly acquired this dark matter to explain the galactic large-scale structure. The current models can be summarized by the conclusions from Press and Schechter (1974), White and Rees (1978), and White and Frenk (1991): the galaxy distribution has evolved by the hierarchical and self-similar growth directed by dark matter collapse, because galaxies originate from gas condensing, cooling, and forming stars within those dark matter haloes. This process can be understood as a sequence of gravitational instabilities of dark matter overdensities: the dark matter clumps collapse into each other and the haloes become larger. The haloes originally formed by the linear evolution of the set of the primordial perturbations in our Universe – the nature and origin of which is out of the scope of this thesis – that collapse non-linearly when an overdensity threshold is reached. Then, the primordial gas started falling into these dark matter halos and the galaxies started to form. Quoting from White and Rees (1978): "Luminous galaxies up to a certain limiting size can form when this gas cools and becomes sufficiently concentrated in the centre of the [dark matter halo's] potential well to be self-gravitating and liable to fragmentation".

Nonetheless, dark matter in galaxies is not only a theoretical concept, and many further detections that can only be reconciled by its existence have been made over the decades.

Strong gravitational lensing by single galaxies has also shown that the lens cannot be modelled alone by the light emitted, requiring a dark matter halo (e.g. Dye and Warren, 2005; Nightingale et al., 2023). It has also shown its gravitational influence through the weak lensing effect of the large-scale structure from cosmic shear surveys (Heymans et al., 2021; Abbott et al., 2022; Dark Energy Survey and Kilo-Degree Survey Collaboration et al., 2023). The rotation curves of spirals keep showing examples of dark matter in galaxies (Dolgov, 2000; Sofue and Rubin, 2001). Dark matter is in fact necessary to explain the dynamics of the family of galaxies interacting with the Milky Way in our Local Group. X-ray emission due to high energy electrons have been found in the mass distribution of galaxy clusters. This has been attributed to the hydro-static equilibrium of the electrons within the gravitational well: as the field in a cluster is very strong, the electrons become a hot plasma that generates X-rays by bremsstrahlung and heavy element lines (e.g. Fox and Loeb, 1997; Nagai et al., 2007; Schindler and Diaferio, 2008; Reiprich et al., 2013).

Last but not least, the existence of dark matter is also supported by cosmological surveys. The main probe of this type comes from the Baryonic Acoustic Oscillations (BAOs), observed both in the Cosmic Microwave Background (CMB) and in galaxy surveys. To understand the physics behind BAOs, it is necessary to understand the Universe according to ΛCDM before the epoch of recombination, when the CMB was emitted. The Universe was fundamentally composed by a plasma "soup" of protons and electrons interacting electromagnetically in thermal equilibrium through the exchange of photons. In such a system, an over-density – such as the primordial cosmic perturbations – would produce a sound wave that propagates through the medium, undergoing oscillations powered by the opposing forces of gravity and radiative pressure. The other components present at that time were neutrinos, other more massive atomic nucleus generated during the Big Bang Nucleosynthesis – the cosmic period when atomic nuclei formed –, and dark matter. The dark matter had its own and heavier overdensities, but also was dragged to the BAO overdensities by the gravitational pull of the baryons. Due to the non-stopping expansion, the plasma cooled quasi statically to a temperature smaller than the ground energy level of a Hydrogen atom. This allowed the photons to decouple and to stop scattering between charged particles, which were recombining into neutral atoms. The Cosmos in this stage "is known" as the surface of last scattering, and produced the light that has become the light that A. Penzias and R. Wilson found in 1965 (Penzias and Wilson, 1965) and our telescopes study today. The complex superposition of sound waves is imprinted in the temperature fluctuations of the CMB, together with many other types of anisotropies – which are also out of the scope of this thesis. The angular scale of the BAO depends on the sound speed in the plasma, but the amplitude of its waves is correlated with the relative amount of baryonic matter and dark matter. This is because of the attraction between baryonic overdensities and the dark matter they dragged. The success in applying the ΛCDM model to the angular scale of the CMB anisotropies is one of the main arguments supporting the Cold Dark Matter (CDM) paradigm, being (Planck Collaboration et al., 2020) one of the latest results.

Now, while dark matter was dragged by the baryonic overdensities that propagated as BAOs, it also maintained its own overdensities. After recombination, both dark matter and baryons would coalesce over time into overdensities derived from the BAO's wave expansion. Nonetheless, the majority of mass fell into the more massive dark matter overdensities. This led to a statistical distribution of galaxies within dark matter haloes with a characteristic scale provided by the propagation length of the sound waves at the epoch of recombination. Galaxy Surveys are able to recover this scale by the correlation between the distance of galaxies, being (Eisenstein et al., 2005) one of the – *if not THE* – first.

The properties and structures of the Dark Matter haloes around galaxies and group of galaxies have been studied thoroughly over the years. The main target for proving these properties has been the dynamics of the components of the galaxies towards their

outskirts, and their distribution, as a proxy of the halo of Dark Matter. The combination of observations and simulations gave way to empirical halo profiles (Merritt et al., 2006). The most generally accepted ones are the Navarro-Frank-White (NFW) profile (Navarro et al., 1997), which combines in one function two different radial profiles for the inner (r^{-3}) and outer (r^{-4}) parts; or the Einasto profile, consisting of a three-dimensional generalized power law r^{-n}, being n a parameter to fit. One prominent success of the Einasto profile is in the galaxy of Andromeda (Einasto, 1969) was followed up in Tempel et al. (2007), using the optical and infrared observations (Tamm et al., 2007).

1.1.1.2 Baryonic Matter

Current observations of the baryonic components of galaxies come from a combination of the study of our Milky Way, the galaxies in the Local Group, and many other galaxies that our telescopes can observe with enough high resolution to discern internal regions. We now know fairly well the multiple life paths of individual stars and have good models to understand the behaviour of binary stars and stellar clusters. The processes in which clumps of gas become capable of forming stars have also been modelled with considerable confidence: gas cools down radiatively until it is able to collapse into stars through Jeans instabilities (Jeans, 1902). This produces populations of stars which can be modelled through stellar population synthesis (Salpeter, 1955; Conroy, 2013), allowing to build a theoretical spectrum of a galaxy. The combination of light from a galaxy emitted along the whole electromagnetic spectra is called Spectral Energy Distribution (SED). Different components of the galaxies undergo different processes that emit in different wavelengths. Examples of science using SED fitting can be found in (Małek et al., 2010; Nanni et al., 2020; Riccio et al., 2021; Hamed et al., 2023b; Pistis et al., 2024).

The combination of the different stages of stars and the surrounding gas leads to the complexity of Phases of the Interstellar Medium (ISM). It is crucial to relate the ISM gas cycles to understand how they shape galaxy evolution and, more importantly, the processes in galaxy mergers. In the cold gas regions that collapse into Jeans instabilities producing new stars, the gas that is not fuelling the newborn stars gets heated and dispersed by the high energy light and cosmic rays they generate. Therefore, an originally cold gas cloud becomes over time a gas-depleted cluster of stars, covered by progressively colder gas shells.

Dispersing gas is however not the last contribution of stars to the ISM, as during the later stages of their life processes they reintroduce gas through planetary nebulae and supernovae, heating it again (McKee and Ostriker, 1977). This gas shows a larger abundance of heavy elements than the original media. This increases the metallicity of the ISM, which measures the relative amount of elements heavier than hydrogen and helium. It also heats the surroundings and, in the case of supernovae, distorts the media due to the explosion's high speeds, energy, and large distance reach. A new cycle where the gas cools down into forming new stars is possible if the amount of gas that manages to cool down is large enough.

The gas can be observed by its emission in varying wavelengths and specific narrow and broad-band emissions. These lines are emission lines from some of the elements present in the gas, and their width depends on the velocity dispersion of the emission regions. Moreover, the gas in a galaxy does not change only due to the internal processes and gas expulsion, but also follows a cycle where the gas comes back to the galaxy by mergers – see Sect. 1.2 – and inflows from the surrounding medium, known as circumgalactic medium (CGM; e.g. Tumlinson et al., 2017).

Dust grains also can form from the gas, and although not very abundant, dust has a crucial effect on the light emission. They are mainly small solid particles, their size is

between ∼100nm and ∼1μm, although measurements are uncertain (Ysard et al., 2019). The gas can form dust grains when it cools down, and it fulfils the appropriate conditions for any of the multiple dust formation mechanisms (e.g. Gail and Hoppe, 2010).

Dust has two main influences on the galaxy's light emission. It can absorb and scatter the light of nearby stars, producing the extinction or attenuation of the light emission. Moreover, the light it absorbs is re-emitted in longer wavelengths as infrared emission. It produces a very complex net effect where high energy light emission in the Ultraviolet is reduced and the far infrared one increases (e.g Calzetti et al., 2000; Hamed et al., 2023a; Małek et al., 2024)

In all this review, we have not considered yet all the other known and more exotic forms of baryonic matter. Pulsars and Black Holes, the compact objects resulting after SNe explosions. Planets, which from 1992 we know to exist outside the Solar System around pulsars (Wolszczan and Frail, 1992) and since 1995 around other stars (Mayor and Queloz, 1995). Cosmic Rays are particles produced in energetic processes, ranging from those in the stellar coronae to SNe, that can ionize or swipe clumps of gas. Many galaxies also have a Supermassive Black Hole (SMBH) in their centres. These SMBH have been found (Balick and Brown, 1974; Ekers et al., 1975; Lo et al., 1975) and observed (Event Horizon Telescope Collaboration et al., 2022) in the Milky Way, in Andromeda, and many other galaxies.

These SMBHs can power the most energetic phenomena observable in the sky by the accretion of gas. The first observation of such a process was done by Schmidt (1963). Originally, the source was dubbed Quasi Stellar Object (QSO), because the optical telescopes found it to be a point source like a star, although they had an unusual emission in radio wavelengths and showed a high redshift on the hydrogen emission lines that was only consistent with a distance very far outside the Milky Way. Nowadays, we know the process that powers QSOs is an accretion disk rotating around the SMBHs, which is known as Active Galactic Nuclei (AGN). AGNs are another key element in galaxy life and evolution. They have a strong effect on the cycle of gas along the Universe's large-scale structure because their power can eject gas from the galaxies to their surrounding media (e.g. Feldmann et al., 2016).

1.1.2 Types of galaxies: galaxy classification

1.1.2.1 Morphology

Since the first studies of the spiral nebulae in the 1920s and 1930s, it became quickly apparent that galaxies show a high variability. The first famous classification is Hubble tuning fork, shown in Fig. 1.1, which is still a reference for astronomers today. It covered the more frequent galaxy types, in a sequence that was already hinting an evolutionary path: the radially symmetric galaxies (E0), that led to the more ellipsoidal shapes of elliptical galaxies (E3 – E7), which then separated into two paths of spirals galaxies, with (Sa, Sb, Sc) and without (SBa, SBb, SBc) a bar, and with the intermediate stage of lenticular galaxies (S0). This method was nonetheless excluding the irregular galaxies, whose abundance was increasing at that time.

Current understanding of galactic evolution corrects this evolutionary line. The elliptical galaxies dubbed as early-type galaxies are actually old galaxies with little star formation, and therefore are a late life stage. The late-type spirals are the opposite, they show more gas, more star formation, and are comparatively younger in their stellar populations. Nonetheless, the names early-type and late-type still accompany us today.

Elliptical, lenticular, and spiral galaxies are indeed the most abundant galaxies in the local universe, around a 60% of spirals, a 22% of lenticulars, and a 13% of ellipticals were found in the first catalogues (de Vaucouleurs, 1963). While there is obvious dispersion

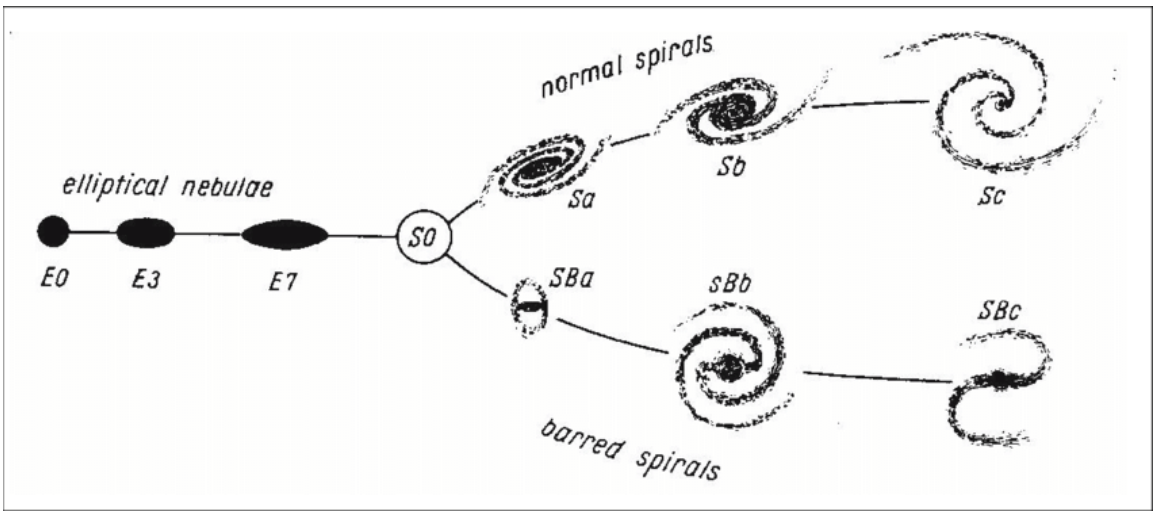
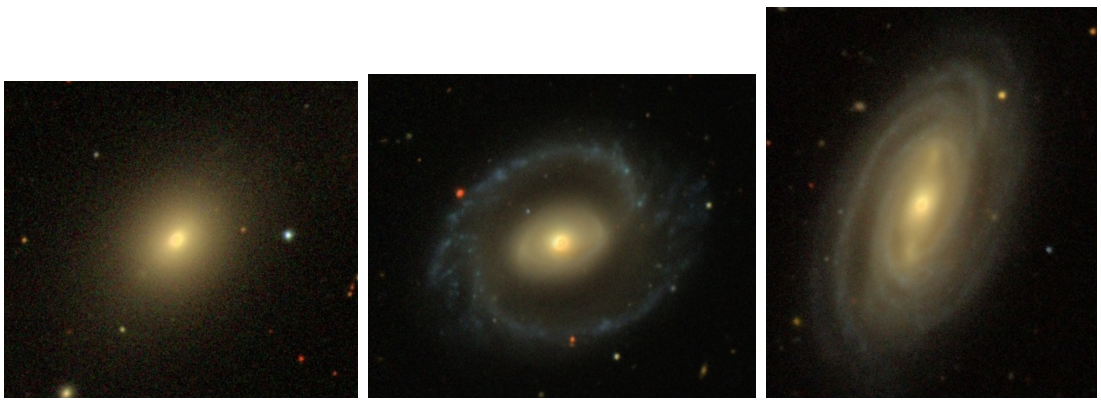


Figure 1.1: Hubble's original diagram.



(a) Elliptical Galaxy. (b) Lenticular Galaxy. (c) Spiral Galaxy.

Figure 1.2: Credit: Aladin SDSS DR9.

in the properties of members of the three types, they do have clear common ones. An example of an isolated elliptical galaxy is shown in Fig. 1.2a. The shape is spheroidal, with no clear substructure, except for the brightness profile that increases towards the centre. The colour is predominantly red because the majority of stars it hosts are red old stars, the less massive and more long-lived ones from the remaining stellar populations. The gas in the ISM has been depleted, the galaxy has been quenched, so that there is no clear presence of gas clumps and star forming regions. In fact, the reason why galaxy evolution leads to this stage is a key question for our modelling and is still science in progress (e.g. Corcho-Caballero et al., 2023; Park et al., 2023; Lorenzon et al., 2024).

Lenticular galaxies are like in Fig. 1.2b. They are somewhere in between elliptical and spirals because, although they have a bulge and a flat disk, the bulge-to-disk size ratio is larger than for spirals and they don't show spiral arms.

Spiral galaxies are like Fig. 1.2c. Their aspect is by far less uniform than ellipticals: they have three main components whose relative size – and even presence – changes for every source: the central core, known as bulge, where older and redder stars are present and that resembles a nuclear elliptical galaxy; the disk, which is flat and axisymmetric at first approximation, with the famous arms spiralling along the disk generally in pairs, and that is in general full of gas, with a very dynamic ISM; and finally the bar, not present in all spirals, that crosses the bulge across its centre, and that seems to drive gas inwards from the disk towards the bulge and to increase star formation along the bar (Knapen et al., 1995; Martel et al., 2013).

There have been multiple further attempts to improve this classification. One of them was created by de Vaucouleurs in the 60s, where he added an orthogonal axis of variability to the evolutionary line (Figure 1.3). This orthogonal cross-section included deviations of more general shapes, such as how long the arms spiral around the galaxy, the relative size between the central bulge and the disc, etc. One other attempt was the Yerkes classification system (Morgan, 1958), that combined the morphological information, mainly the degree of central concentration and extension, and the spectral information.

While the community kept attempting to categorize the morphological zoo of galaxies and understand their life processes and evolution, a fourth broad type of galaxies had posed an additional puzzle since the beginning – if we count the other three as elliptical, spiral, and lenticular. Those were the irregular and peculiar galaxies, and Hubble's and de Vaucouleurs' classification did not have a clear place for them. In the case of irregular galaxies, the main characteristic is that they lack of any symmetry and the structure observed in the other galaxies. In the case of peculiar galaxies, their shapes were clearly, to a greater or lesser extent, highly distorted versions of the three main types. The first big catalogue of irregular and peculiar galaxies was Arp's Atlas of Peculiar Galaxies (Arp, 1966). These types of galaxies we now know are produced by the interaction between galaxies, and depend on the relative stage along the whole interaction process (Toomre and Toomre, 1972).

As the reader might imagine, the complexity in galaxy types has been increasing over the decades. Some of the most abundant ones are dwarf galaxies, which are generally smaller galaxies around the larger elliptical, spiral, or lenticular ones. Their shapes tend to be either spheroidal $\sim$43%, spiral $\sim$45%, or irregular $\sim$13% (Lazar et al., 2024). Clear examples of local irregular dwarfs are the Magellanic clouds or Andromeda's satellite M32.

Other types of galaxies are those which host an AGN, showing a complex range of galaxy "manifestations". The galaxy components themselves are generally too dim in comparison with the nucleus' high luminosity to be observed fur the nearby universe. AGNs show quite strong emission in other bands than the optical. Their emission in radio waves, or narrow and broad emission lines, determine the type of AGN. Overall, the AGN type depends on two things: the presence of a high-speed radio jet, launched along the axis of the accretion disk; and the relative orientation of the system. The orientation is important

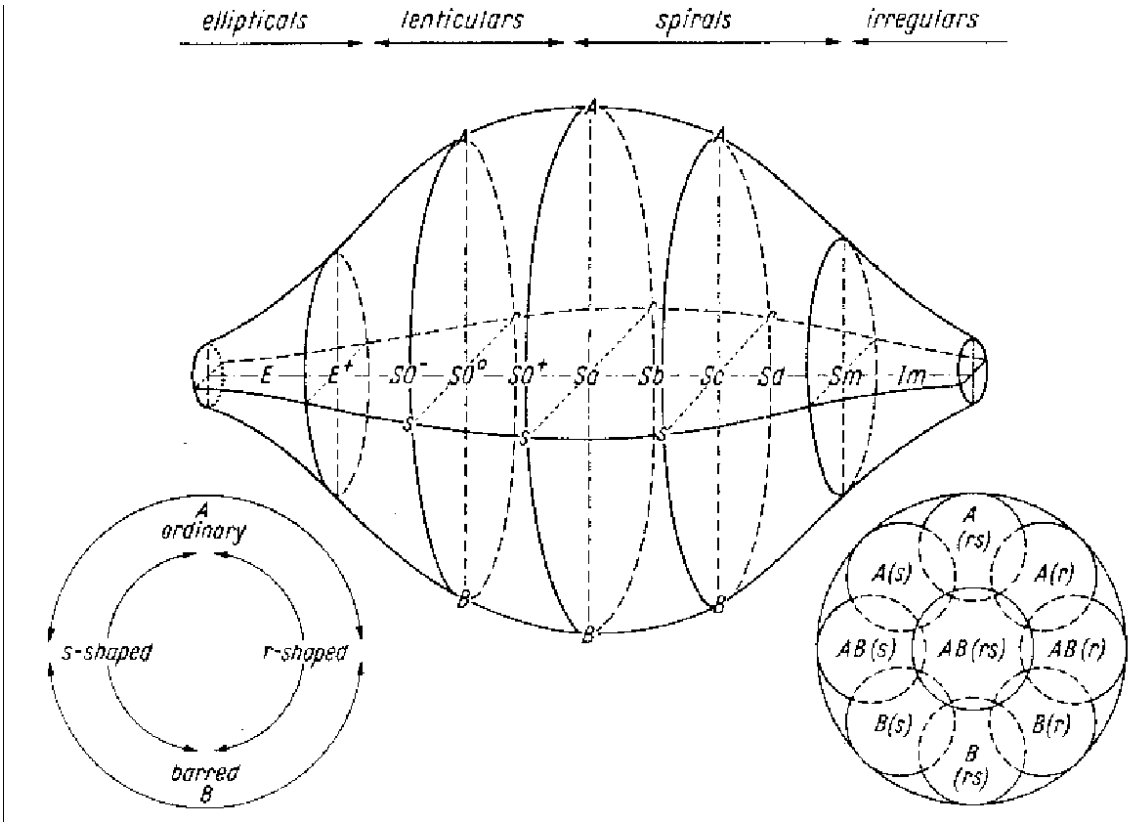


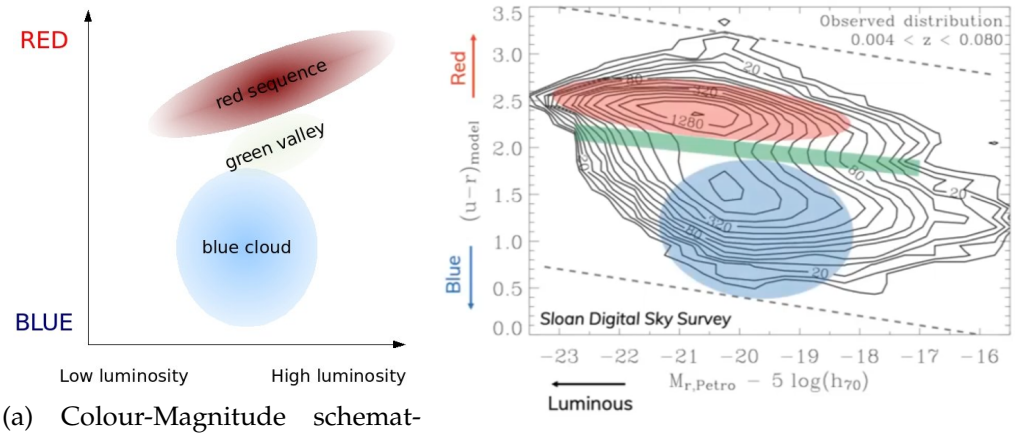
Figure 1.3: de Vaucouleurs' revision (de Vaucouleurs, 1959)

because AGNs have a complex axisymmetric structure, and the emission features depend on the sections of the structure that are visible.

To complicate a bit more the zoo of galaxies, there are the low-surface brightness (LSB) galaxies. They do not fit very well into the current galaxy evolution models, but they are anticipated to become one of the most frequent classes when the large galaxy surveys become sensitive enough to find them (McGaugh, 1996; Bothun et al., 1997; Martin et al., 2019). This is because, as the name indicates, they have a very dim brightness profile, making a high depth in minimum brightness necessary to find them. We have seen many of them already in recently built catalogues such as Thuruthipilly et al. (2024). Other type of galaxies in the LSB regime are the Ultra Diffuse Galaxies (UDG), which have been observed in the Local Group (e.g. McConnachie, 2012), and more might still be found (Newton et al., 2023).

1.1.2.2 Photometry

The astronomical images contain not only the morphological information of the galaxies but also their flux imprint. These are the photometric measurements, they take care of the measurement of the flux that a source has in an astronomical image. In the case of the majority of telescopes, it is done in a single band. The flux of a source is described by magnitudes, which relate to the flux through a logarithmic relation. The main reason for using magnitudes is historical, as the human eye is sensitive to the brightness of stars in the sky in a logarithmic way. In 1856, Norman R. Pogson defined it as $m = -2.5 \log F$, being F the flux and m the observed (or apparent) magnitude. Because of the logarithm, the subtraction of two magnitudes corresponds to the fraction between two fluxes: $m_1 - m_2 = -2.5 \log F_1 / F_2$. Therefore, if the index 1 and 2 correspond to two different bands – r and g



(a) Colour-Magnitude schematics. Credits: Wikipedia Joshua Schroeder

(b) Addapted from (Baldry et al., 2004) by Prof. Francesco M. Valentino.

Figure 1.4: Colour-Magnitude diagram: schematic and SDSS results.

bands would be a contemporary example, covering roughly the red and green areas of the visible light, respectively – their subtraction represents the slope in the SED along the areas covered by the bands. These are the colours of the galaxy, and provide relevant physical information about the galaxies.

The most basic application of photometric measurements to galaxy properties is the colour-magnitude diagram. This diagram shows distinct populations clumping in two different areas. Figure 1.4a indicates a schematic, while Fig. 1.4b shows an actual plot representing galaxies from the Sloan Digital Sky Survey (SDSS). Note the relative inversion of the x-axis, as luminosity increases towards the left in Fig. 1.4b, and towards the right in Fig. 1.4a. Specifically, the x-axis in 1.4b indicated the absolute mass of the galaxies, as calculated within the Petrosian magnitude (Petrosian, 1976) – see Sect. 2.1.1 in Chapter 2.

The red sequence region is where galaxies like the elliptical ones introduced above are located, and the blue cloud is where spiral galaxies are located. As the colour $u - r$ measures the flux fraction -log F_u/F_r = +log F_r/F_u, galaxies with a larger colour should have a larger proportion of red light than blue light. This is one of the main indications of the stellar populations of elliptical and spiral galaxies described in Sect. 1.1.2.1. The elliptical galaxies are redder, they have more red old stars and less blue new stars, while the opposite is true for the spirals. The figure also indicates the green valley, a region in between the two types where fewer galaxies can be found. The presence of the green valley illustrates not only the bimodality of the distribution but also the imperfect separation of both classes. Moreover, the red and blue areas not only show late and early-type galaxies respectively, as there also exist blue elliptical (e.g. Lazar et al., 2023) and red spirals.

1.1.3 Galaxy evolution: hierarchical growth of structure

The bimodality of the colour-magnitude diagram and the morphological types imply two major evolutionary stages for galaxies. The Hubble tuning fork was already aware of this (Figure 1.1). If one considers the presence of gas as the main driver of the transition between the two galaxies, a fairly clear picture of the processes that drive galaxy evolution can be obtained.

Quite interrelated with the galactic gas cycle is the dark matter halo mass function, which indicates the number density for any given mass of haloes present in the Universe. The mass function resulting from the Press-Schecter formalism (Press and Schechter, 1974) is indicated in Fig. 1.5. At the same time, one can assume a luminosity function of galaxies,

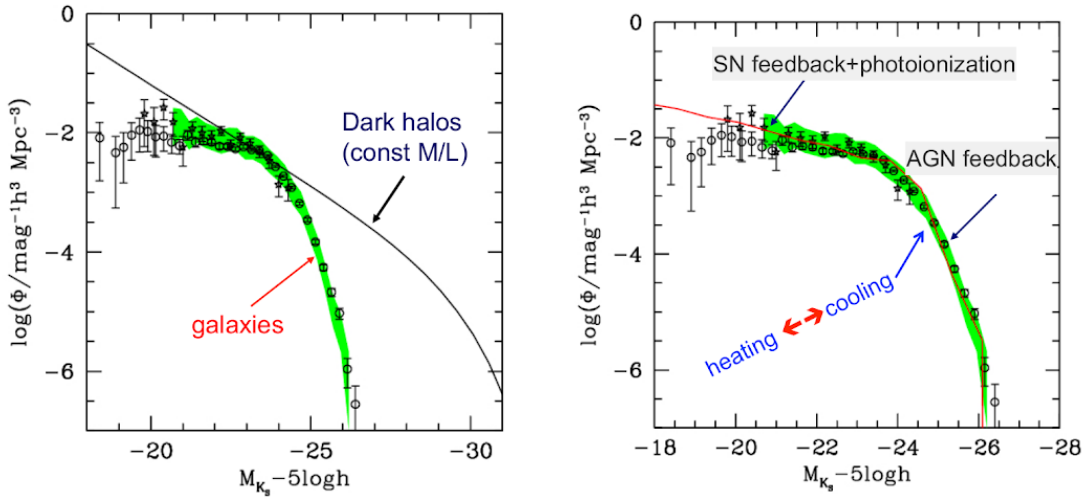


Figure 1.5: Both panels compare a theoretical luminosity function with the luminosity function found in simulations. The theoretical luminosity function has been obtained from the mass function of dark matter haloes following the Press-Schecter formalism and considering a constant M/L. The panel on the left shows how the luminosity function of galaxies in simulations disagrees with the dark matter distribution. The right panel indicates the effect on the luminosity function of multiple physical phenomena observed in real galaxies, that have been applied to the simulations. Courtesy of dr. Wojciech A. Hellwing, in the lecture Introduction to Cosmology.

i.e., the number density of galaxies for any given total luminosity. However, assuming a constant mass-to-light ratio (M/L) does not provide an equivalence between the dark matter mass function and the galactic luminosity function. This is partly due to the complication to estimate the mass of luminous matter in galaxies, as it requires a good quality of information of the galaxy's SED to simply get an initial model. Nonetheless, the masses can be measured with a good enough quality to show clear disagreements. In fact, the effects of baryonic processes in the gas circulation are responsible for this mismatch.

The suites of simulations that include baryonic matter dynamics through hydrodynamics and/or semi-analytic models (see Somerville and Davé, 2015, for a review on modern simulations) have helped enormously understanding why the discrepancy. The two panels in Fig. 1.5 indicate the explanation of these findings. On the left panel, it can be seen clearly that the galaxy luminosity – represented as high absolute magnitude in the K_S band M_{K_S} – adapts to the halo mass function only in one intermediate point, and shows an underpopulation of galaxies for both the lower and larger luminosity tails. The right panel shows how a combination of photoionization by stellar emission and feedback from supernova is able to reduce the amount of galaxies with low stellar mass. This is because both of them reduce the efficiency of gas cooling that allows the stars to be formed. Photoionization is produced by newly formed stars, as their energy is emitted in high energy light and cosmic rays that disperse the surrounding gas in the star forming regions. Supernova have their effect when they emit gas back to the ISM, but can also propel gas outside the galaxy. SNe's net effect reduces star formation in low mass galaxies due to the gas ejection. On the bright end, AGNs are responsible for reducing the stellar mass within the most massive haloes. They do it through the consumption and emission of gas to the inter galactic media produced by the accretion disk.

These insights obtained through simulations are a clear example of the big interrelation of galactic processes. Nonetheless, those above are reduced to stars, gas, and SMBHs, the

surroundings of galaxies and what the interaction with other galaxies could produce. As we know from the Press-Schecter formalism and ΛCDM model, the Universe's distribution is driven by the gravitational assembly of mass. Small and initially overdense regions come to each other and collapse into subsequently larger and larger structures. It is therefore not surprising that one of the other main influences in galaxy evolution come from the actual collapse of galaxies with each other, from galaxy mergers.

The improvement of the simulations is allowing larger numbers of particles, higher resolution, and better understanding of the baryonic physics involved in galaxy evolution. At the same time, they are providing further new insights in the importance of galaxy mergers in the growth of structure. Many simulations have focus on estimating the mass growth produced by mergers (e.g. Rodriguez-Gomez et al., 2016; Angeloudi et al., 2023). Some others have shown the strong influence of the accretion of minor mergers in building massive spirals galaxies (e.g. Jackson et al., 2022), or massive elliptical galaxies (e.g. Rutherford et al., 2024).

Thus, galaxy mergers are a crucial stage in mass assembly of galaxies, although it is not yet known to what extended. Furthermore, they have also shown to have a significant effect in the star formation, morphological transformation, and many other important attributes of galaxies. To continue advancing our understanding of the Universe and the evolution of galaxies, it is necessary to learn more about merging galaxies.

1.2 Galaxy Mergers

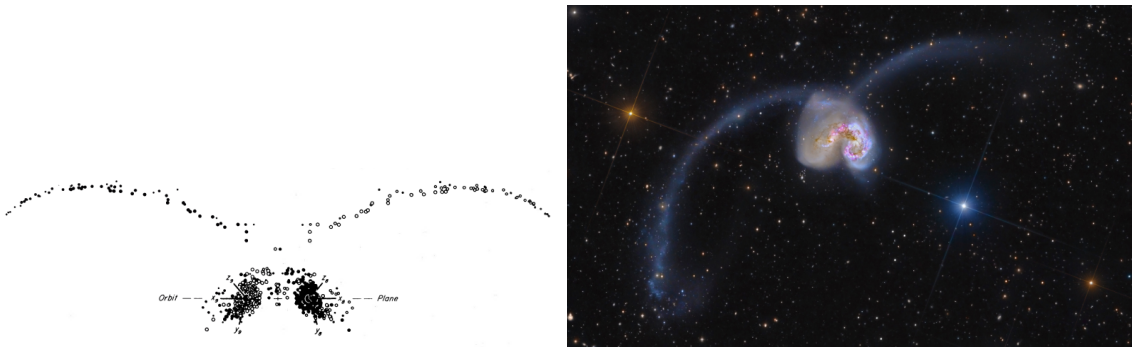
Galaxy mergers are irregular/peculiar galaxies which get their shape because they interact with each other. Merging is a process that culminates into the original parent galaxies becoming a final daughter one. Tidal gravitational forces between the galaxies arise during approaching orbit and highly distort the morphologies, generating various features such as tidal tails, bridges, or shells (Toomre and Toomre, 1972; Quinn, 1984).

Research over the physics behind the beautiful galaxy shapes in Arp's Atlas led to many discussions during the 60s and 70s about their nature. At that time, it was doubted whether gravitational tidal forces could lead to thin bridges or arms, but should only be responsible for much broader shape distortions. In Toomre and Toomre (1972), the Toomre brothers performed a series of simulations of two massive bodies approaching each other surrounded by a disk of ideal non-self-interacting test particles. Using a Runge-Kutta integration method (Pfleiderer, 1963), they combined relative orbits between the two massive bodies, their relative masses, the initial rotation of the disk of the test particle, and the view angle of the eventual tidal features. They showed how the simulations reproduced some types of patterns seen across the night sky, such like the "Antennae" galaxies: Figures 1.6a and 1.6b. The process is essentially gravitational and kinematic, showing how simple and well known laws can lead to highly complex manifestations. These ideal particles would correspond to gas and stars that in reality would be behaving rather like a heterogenous "continuum" and not as the modelled isolated points.

Current detections of mergers indicate that they make up fewer than 10% of the galaxies at low redshifts (Mundy et al., 2017; Duncan et al., 2019), evolving up to 20% in the redshift interval z ϵ [2,3] (Tasca et al., 2014). However, their influence of galaxy evolution at high redshifts is still under debate (Lofthouse et al., 2017).

1.2.1 Effect of Mergers

One of the most commonly accepted effect of mergers is that two spiral galaxies of similar masses become elliptical after the process (Mihos and Hernquist, 1996, Conselice, 2006), and that they trigger star formation along the process (Joseph and Wright, 1985; Patton et al.,



(a) Antennae-like system as simulated in Toomre and Toomre, 1972

(b) Anthennae galaxies, Credit:Hubble ESA/NASA

Figure 1.6: Atennae galaxies: simulated in and in the real sky

2005), although this increase has not always been found (e.g. Pearson et al., 2019a). This star formation seems to be triggered by the movement of gas towards the bulge (Sanders and Mirabel, 1996). These type of mergers are commonly denominated as wet major mergers: wet due to the abundance of gas, major due to a mass ratio between the two merging objects close to 1 – generally from 1 to 1:4. The star formation enhancement has nonetheless not been seeing when two elliptical galaxies merge due to the lack of gas. This merge class is considered as a dry major merger (Hwang et al., 2011). Mergers also seem to be related to the power of the accretion disks in AGNs by inducing the redistribution of gas between the galaxies (e.g. Keel et al., 1985; Scott and Kaviraj, 2014; Ellison et al., 2019). Nonetheless, this is still not fully confirmed and there are works that disagree with this picture (e.g. Kocevski et al., 2012).

Galaxy mergers with a large difference in mass between the two parent galaxies, known as minor mergers – generally those with mass ratios larger than 1:10 –, have been also seen to drive the morphological transition between spiral galaxies and lenticular ones (Bernardi et al., 2011; Eliche-Moral et al., 2018; Maschmann et al., 2020; Tous et al., 2023). We also know that the Milky Way (Helmi, 2020) and many other nearby galaxies show features of the satellites that they have already assimilated. This, supported by the Press-Sechter formalism, has led the community to deduce that the assimilation of satellites is one of the pathways of galaxies to obtain new gas and form new stars.

However, merging is not the final destination of all sources that show signs of interaction. The distorted aspect of a pair of galaxies does not necessary need to imply they are going to merge into one. Those are known as pass by galaxies, being the Cartwheel Galaxy one of the most famous examples. Information about the relative velocity and distance would be necessary to make a prediction. Besides, the process is inherently slow due to the dimensions of galaxies: they can last for $\sim$1 billion years (Kitzbichler and White, 2008; Lotz et al., 2008; Pearson et al., 2024a), and the tidal features can last up to $\sim$3 billion. As a consequence, even with a high image resolution and quality of the data, the merging state can be ambiguous.

The environment of galaxies can also play a role on merging systems, as shown in the two projects I have been part of during my PhD: Pearson et al. (2024b) and Sureshkumar et al. (2024a). Both works reached through different approaches to the conclusions that galaxy interactions are less likely to result in a merger within environments more dense in number of galaxies, where the galaxies have a higher relative velocity. This was in accordance with previous works (e.g. Ostriker, 1980), although it seemed to not be the case for small scales in the simulations of Omori et al. (2023). Thus, how merging depends on the galaxy large-scale distribution is still under debate.

1.2.2 Types of features

During merging, whole galaxies get distorted and some of the material gets stripped to the outskirts of the original systems. The debris generated by the tidal interactions shows different morphologies that depend on the type of collisions. Some of the particular features that can be found are tidal tails, streams, or shells. This variety of distortions is produced by different types of collisions.

Tails Tidal tails are gas-rich (see e.g. Duc et al., 2000) and mostly found in late-type galaxies, i.e., spirals. They host star-forming complexes and are able to form star-clusters or even second-generation dwarf galaxies that get stripped from the final system (Duc and Renaud, 2013). The first attempts to research major mergers were done through numeric simulations of disk-like sources, such as in Toomre and Toomre (1972). It was shown that the tidal arms appear from near-resonance between the orbital speeds of the galactic discs. These tidal tails form from pulled out material of the disks, and are the most prominent characteristic of merging systems. Figure 1.7a shows an example of tidal tails formed from two spiral galaxies interacting.

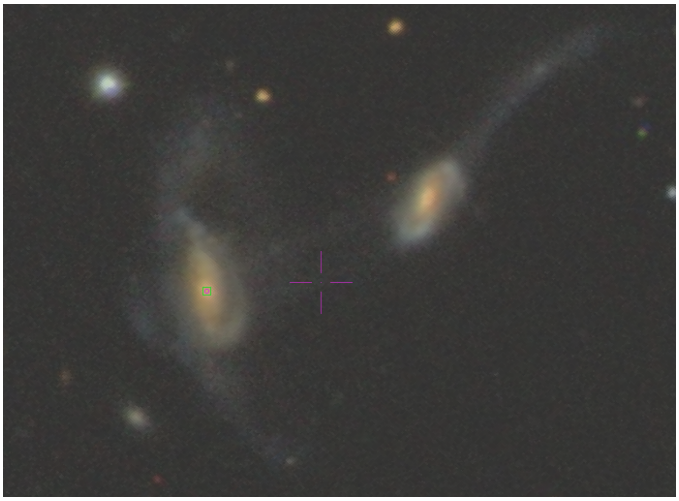
Streams Compared with the tidal tails, the tidal streams are relatively gas poor and show faint stellar absorption lines in spectroscopic studies (Fensch et al., 2020). They are frequently found as stripped material from a low-mass companion orbiting or being consumed by a primary galaxy (Hood et al., 2018). Therefore, they are characteristic of minor mergers, where the mass of the low-mass source is one fourth or less than the high mass one. Figure 1.7c depicts two stellar streams emerging from the sides of an elliptical galaxy. It is worth noting that the southwest streams ends in a shell/arc shape that seems to begin forming.

Shells The models and observations of galactic shells have shown they are relatively gas poor in comparison with tidal tails (Charmandaris et al., 2000). From numerical simulations, it has been inferred that the mechanisms responsible for shell formation are likely to be near-radial infalls of relatively massive minor mergers, with mass ratios above 1:10. Shells are therefore thought to be made up of stars from accreted satellite galaxies, tidally disrupted by the massive main galaxy they merge into. They are relatively frequent around elliptical and lenticular galaxies (Krajnović et al., 2011; Fensch et al., 2020). Shells appear as arcs concentric to the spheroidal shape, as in Fig. 1.7b.

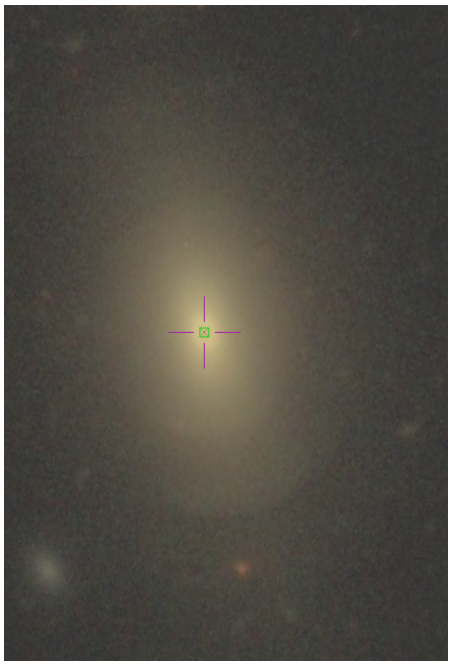
Passing by galaxies, in the minor merger mass ratio range, are also believed to produce shells. They cross them radially close to the centre of the main galaxy, with a low impact parameter, and excite radially the stellar distribution. Also, both radially and non-radial are believed to produce them (Hernquist and Quinn, 1988).

1.3 Merger identification

Galaxy merger identification is the obvious first step for studying the phenomena, but it is not trivial. There are multiple ways of doing it, which can be gathered in four main categories: visual inspection, cross-pair studies, morphological parametrization, and Machine Learning (ML) based classification. All of them have different applicability ranges, depending on if the data is an image or a spectrum, on the resolution, on the photometric bands used, etc.



(a) Galaxy pair interacting with tidal tails, Aladdin SDSS DR9.



(b) Elliptical galaxy with shell, Aladdin SDSS DR9.

(c) Galaxy with stellar stream, Aladdin SDSS DR9.

Figure 1.7: Most frequent tidal features in galaxy mergers. The contrast in the image was enhanced to make the dim features more visible.

1.3.1 Visual inspection

Identification of galaxy mergers through visual inspection is simply the educated or amateur opinion saying if the image of a galaxy shows a merger or not. Depending on the image quality and resolution, the features can appear clearer to the viewer. It is a very time-consuming and subjective method, that has clear human biases. The difficulty presented by ambiguous or not-well-resolved sources can lead to classifications that are inconsistent between observers, or even between repeated iterations by the same observer. The image of the galaxy itself can be ambiguous: it may show distortions that are not due to galaxy-galaxy interactions; even though two galaxies overlap they can be at far distances; or they may show an interaction that is not going to end as a merger, but it is a pass by. Due to the huge time-scales of merging events, there might be cases in which the final state simply cannot be known.

Traditionally, astronomers have characterized the shapes by eye, as was done in the initial galaxy classifications (de Vaucouleurs, 1959) and is currently done for creating catalogues of mergers (de Vaucouleurs, 1963; Arp, 1966; Darg et al., 2010a). It has also been carried out for creating catalogues of merging features (Sola et al., 2022), or to make further performance tests of other types of methods (Pearson et al., 2022).

The tools available through the use of the internet by the general public have allowed the extension of classifications to large numbers of volunteers. The most famous of which is Galaxy Zoo[1] (GZ; Lintott et al., 2008). GZ is a project where volunteers join through the internet and classify the morphologies of galaxies provided. These types of studies are called citizen science. A person using the GZ online interface is presented with the image of a galaxy and a set of questions that belong to a decision tree. These questions try to cover many possible properties of the galaxy, such as the presence of spiral arms, tidal arms, or even artefacts coming from satellites and bright stars. Over the years, GZ has delivered many large catalogues of morphological classifications (Lintott et al., 2011; Willett et al., 2013; Walmsley et al., 2023), and have been taken advantage of by many studies[2]

1.3.2 Cross-Pairs

One drawback of visual inspections that has been mentioned is how two galaxies nearby in the sky might not be physically interacting. A method that can untangle this degeneracy is to estimate how far from us, and therefore between them, those galaxies are located. The cross-pair approach makes use of the photometric (see Conselice et al., 2003; Lotz et al., 2011, but note the method is combined with morphological paramters) or spectroscopic information (Patton et al., 1997; Barton et al., 2000; Le Fèvre et al., 2000; Patton et al., 2002; Lambas et al., 2003; Lin et al., 2004; De Propris et al., 2005; Ellison et al., 2008; Rodrigues et al., 2018; Duncan et al., 2019) to estimate the redshift of the galaxies as a proxy to the distance.

Through the redshift of a source, one can calculate the relative velocity it moves with respect to the observer. Initially, this was understood as the Doppler effect on the light's wavelength: if the galaxy is approaching, then its light is shifted towards higher frequency, what in practical effects would make it more blue. When the source is moving away, then it is shifted to red. The first results of a distance-velocity comparison were done in Hubble (1925): the distance was measured through Cepheid stars and the velocities were based on the redshift. The measurements showed how the galaxy of Andromeda is approaching the Milky Way, although it also showed a linear relation between the distance and the receding velocity of other nearby galaxies. This turned out to be one of the most fundamental aspects

[1] http://www.galaxyzoo.org/
[2] https://www.zooniverse.org/about/publications

of our Universe and the ΛCDM model, that the universe is expanding. It was in fact one of the first predictions of the general relativity (Lemaître, 1933). Nowadays, it is known that it expands linearly at a close distance, following the Hubble Law: $v = H_0 \times d$, where v and d are velocity and distance, and H_0 is the expansion rate known as the Hubble Constant. For further sources, this constant is a parameter $H(t)$ dependent of different components of the Universe: mass, radiations, dark energy, and even the curvature of the cosmological metric itself.

Distances in astrophysics and cosmology are nonetheless not easy to define. Because the Universe is expanding, the separation between two sources changes as the light traverses it. As a consequence, it is necessary to construct lengths based on measurable quantities. The angular diameter distance D_A and the luminosity distance D_L. The D_A of a source is based in the geometrical relation between its original known diameter $2R$, and the observed one δ defines D_A. Differently, the inverse square law characteristic of wave propagation is what defines D_L, relating the emitted luminosity (L) to the observed flux (S):

$$D_A = \frac{2R}{\delta} \quad ; \quad D_L = \sqrt{\frac{L}{4\pi S}} \, . \tag{1.1}$$

which would be equivalent to each other in a Euclidean space. In general, they are related (Etherington, 1933) by

$$D_L = (1+z)^2 D_A = (1+z) D_c \, . \tag{1.2}$$

where D_c is the comoving distance, which corresponds to the distance between two sources is the expansion did not exist. The distance in which a source at D_c is located for a given redshift is known as proper distance $D_p(z)$. All these distances can be calculated as a function of redshift and the cosmological model.

The relative velocity of other galaxies towards ours is not only affected by the cosmological expansion but also by the motion given the gravitational effect of their surroundings. This proper motion has an effect that decreases as distance increases, as the expansion dominates more and more the bulk of the velocity. The proper motion effects are more clear in dense regions like galaxy clusters, as it is sensitive to the cluster dynamics.

The combinations of these two effects leads to allowing the use of the measured motion of galaxies to address if two galaxies are in direction of collision or just together in the sky by chance. The cross-pairs technique mainly uses a range of sky-plane distances and velocities between the merger candidates to determine if they are merging or not. As the accurate measurement of redshift must be carried out through spectroscopic observations, this method requires the spectroscopic resources that are not as economical as photometric ones. Regarding photometric redshift measurements, while the accuracy of those is improving with the combination of galaxy templates and Bayesian methods or ML-based models (e.g. Hildebrandt et al., 2010; Bilicki et al., 2018), it is still necessary to compare them with spec-z values to address their quality.

1.3.3 Morphological Parameters

Given the distortion produced by the tidal interactions, one evident option is to quantify the shape distortion of the images and determine regions of the parameter space occupied mainly by merging sources. Through the years, many parameters have arisen, providing many alternatives. Overall, the morphological parameters are quite dependent on the image resolution, the effects of the background subtraction (see Sect. 1.4.2), and the brightness depth of the images. Morphological parameters are applied by defining regions of the parameter space where galaxies with merging-related features are located.

The initial parameters introduced in the modern literature are those that form the CAS system: Asymmetry, Concentration, and Smoothness. Asymmetry is estimated through the absolute value of the subtraction between the galaxy image and its 180° rotation (Abraham et al., 1996; Conselice et al., 2000). The Concentration is estimated by the ratio between the radii which contain two different amounts of the total light of the galaxy (Kent, 1985; Abraham et al., 1994; Bershady et al., 2000; Conselice, 2003). Generally, it is calculated between the radii at 80% (r_{80}) and at 20% (r_{20}) of the total light. It is worth noting that the extension and total light also have to be defined empirically. Last but not least, the Smoothness comes from applying a smoothing weight to the image and subtracting it from the original one. The residuals are higher in the presence of clumped regions (Takamiya, 1999; Conselice, 2003).

The dependence of the Asymmetry on the sky background and resolution has been addressed in Sazonova et al. (2024). The parameter's dependence on those image properties complicates defining value ranges that determine merging distortions consistently (which occurs also for many of the other parameters introduced above). The sky background effect can be attenuated by changing the absolute value with a root-mean-square formula, because it makes the equation independent of Gaussian noise. Also, the resolution differences can be attenuated by deconvoluting the images with higher image quality.

The other state-of-the-art in merger morphological parameters is the Gini-M_{20} set. The Gini coefficient is defined from the series sum of pixels, each of them weighted by their order from the brightest to the dimmest. It is 0 for a flat distribution and 1 if all light is concentrated in a single pixel (Abraham et al., 2003). M_{20} is calculated from the second-order moment of the brightness distribution for the pixels that contains the 20% of the light, normalized by the second-order moment of the total source. The second moment of bright pixels is sensitive to the spatial distribution of bright features such as bars or spiral arms (Lotz et al., 2004). These two parameters can be combined to define a decision region of the combined plane that characterize galaxy mergers (Snyder et al., 2015; Snyder et al., 2015; Rodriguez-Gomez et al., 2019; Pearson et al., 2022; Sureshkumar et al., 2024b).

It is worth noting that, due to the complex nature of galaxies, an infinite resolution of a galaxy would be affected by the vast amount of galaxy features, making some parametrizations unpractical. This can be understood similarly to the complex nature of the sea coastline (Vulpiani, 2014), that in fact played a crucial role in the development of fractal studies by Lewis Fry Richardson in the 50s. If one wants to measure the length of a coastline section, the minimum distance of the measurement tool has a strong effect on the result. This lower limit determines the range of features that can affect the calculation: the more features, the more the distance estimated will tend to be. Therefore, for a galaxy image with infinite pixels, a morphological parameter defined as a function of all pixels can turn quite unreliable at high resolutions. Nonetheless, the real world resolution range where this could be relevant is achievable only for a small amount of galaxies, where a visual inspection would be more adequate.

It is also worth mentioning that, while we call them morphological parameters, they are actually non-parametric quantities. They are more descriptive statistics, that do not depend on a few parameters to be fit, and instead do not imply any previous assumption or model. They are simply applied to the catalogue of sources, and the empirical results are what indicate their capabilities.

1.3.4 Machine Learning

The progressive improvement of the computing technology over the last decades has led to the introduction a new problem-solving techniques that make a high use of computing resources, the Machine Learning (ML) or Deep Learning (DL) methods. They rely on the

mathematical optimization of many parameters over some dataset, what is known as the training of the model. The main advantage of ML is its automatization, once the parameters are able to solve the desired problem, they become much more efficient than during the model training or than humans performing the same tasks. This is because they respond with machine speed and without any new difficult computation. Nonetheless, there is the caveat that the new data has to be similar enough to the training data for the model to be able to perform well.

The more extremely successful examples of ML models are the newly arising large language models. These models are able to generate reliable written conversations using the access to the whole internet to train models with $\sim 10^9$ parameters. They are revolutionizing the way of working, as quick inputs to this type of models can reduce the time to do a wide variety of tasks.

The state-of-the-art in the application of the large language models to astronomy can be seen in ChatGaia[3], an interface that has access to the Gaia Space Telescope (e.g. Gaia Collaboration et al., 2016, the first data release) data and has been adapted from large language models. The user can ask the ChatGaia about some selection of stars by a simple question, and then ChatGaia understand the question and gets the desired source from the Gaia data repository. Moreover, the user can ask the model to generate a plot with the data, as ChatGaia is able to generate a catalogue file, load it to a plotting code language such as python, and generate the script that runs the plot.

Machine Learning models have also found their way around a wide range of applications in image recognition and classification, ranging from facial recognition in social media to more complex objects in very different situations. The parameters are trained to identify features that characterize different classes, such as the face-on wheels of cars or the vertical trunk of a tree.

Depending on the initial knowledge about the data that one wants to analyse with an ML model, there is a bifurcation in what type of models are more convenient to build. When the data is a set of classes that are already known, this requires a Supervised ML model: the parameters will be trained so that the output classifications try to match those already known from the input. When the initial classifications are not known beforehand or the information is purposely ignored, then an Unsupervised ML model would be at hand. Such models can, for example, cluster images similar in some way by creating a parameter space.

One of the fields that have been taking advantage increasingly of ML classification methods is galaxy classification. Image recognition methods have proven their strength for it. This is the case mainly – but not exclusively – of the convolutional neural networks (CNN) (Dieleman et al., 2015). CNNs have proved capable of learning the aspect of multiple types of galaxies (Domínguez Sánchez et al., 2018; Vega-Ferrero et al., 2021; Thuruthipilly et al., 2022; Walmsley et al., 2022). Galaxy merger identification with CNN has in fact become a fourth standard methodology for merger identification (Ackermann et al., 2018; Bottrell et al., 2019; Nevin et al., 2019; Pearson et al., 2019a,b; Walmsley et al., 2019; Ferreira et al., 2020; Wang et al., 2020; Domínguez Sánchez et al., 2023; Margalef-Bentabol et al., 2024). However, it has shown it has limits: it struggles with misclassifications (Pearson et al., 2022) and with estimating the merging stages and times (Margalef-Bentabol et al., 2024; Pearson et al., 2024a).

[3] https://www.whatplugin.ai/gpts/chatgaia

1.4 Photometric data reduction & Sky background

The images and spectra of the galaxies astronomers study come from modern telescopes that operate in multiple wavelengths, resolutions, image sizes, and either on ground-based observatories or space-satellites. The study of galaxy mergers is done mainly in two regimes: high resolution of individual sources or large sky surveys. While high resolution images try to understand in more detail the information of a reduce number of sources, in large-scale surveys the goal is to cover an area of the sky that provides as many sources as possible with a limiting quality.

The regime of telescope surveys of this thesis is ground-based large-scale surveys in the optical bands. The surveys used for this thesis are the Sloan Digital Sky Survey (SDSS; York et al., 2000), specifically its Sixth Data Release (SDSS DR6; Adelman-McCarthy et al., 2008), and the Hyper Suprime-Cam instrument (HSC; Aihara et al., 2018) on the Subaru Telescope (Iye, 2021). It is worth noting that, while spectroscopic observations have not being carried out in this work, SDSS is in fact a spectroscopic survey, that makes use of the photometric images for selecting the targets for spectrometry.

The optical systems of these telescopes consist of a camera and the telescope itself. These telescopes have a reflector system, where the light first arrives to what is known as the primary mirror, which collects the light into a second mirror that subsequently focuses it into the camera. Figure 1.8 depicts a reflector telescope, specifically the cassegrain model, of similar structure to the one in both SDSS and Subaru/HSC. For SDSS, the structure is the same but the shape of the mirrors changes[4]. The light arrives towards the primary mirror from the top-right corner, where the astronomical sources would be. The primary mirror has the shape of a concave ring, so that the light is pointed towards the secondary mirror located within the ring overture. The secondary mirror then focuses the light to the camera taking the images.

In order to get information in the images that is relevant for physical measurements, the observations must have a reliable flux and sky position. To get them is the main task of the astronomical calibration process, commonly known as data reduction. The main steps are to calibrate the flux detected by the camera, the telescope effect on distributions of light, and the positions in the sky plane.

The predominant type of camera in modern astronomy is the Charge-Coupled Device (CCD) camera. Its pixels are exposed for a given amount of time to the skylight, absorbing photons that progressively excite electrons. The brighter the area of the sky that the pixel is exposed to, the more electrons are excited over time. After the detection, the camera reads sequentially the stored electrons per pixel, and transform the number of electrons into a digital value. This number is generally identified as Analogue to Digital Units (ADU), although in the earlier Data Releases of SDSS it is named as Data Numbers (DN). Thus, the photometric calibration provides the means of converting these ADUs to fluxes. In the following lines, I'll define some subsequent relevant steps in photometric calibration. Because in this thesis I have run the data reduction pipeline of HSC (Bosch et al., 2018) – see Sect. 3.4.2 –, I will cover the general data reduction steps in the same order as they appear in the pipeline.

1.4.1 Data reduction calibration frames

The first trouble that has to be overcome when working with ADU's is that they cannot be negative numbers, and for that CCD camera's automatically apply an offset known as bias to every exposure. In order to correct this bias from the actual observations, one creates the Bias images: quick exposures, generally taken in the shortest time in which the electronic

[4] https://skyserver.sdss.org/dr1/en/sdss/telescope/telescope.asp

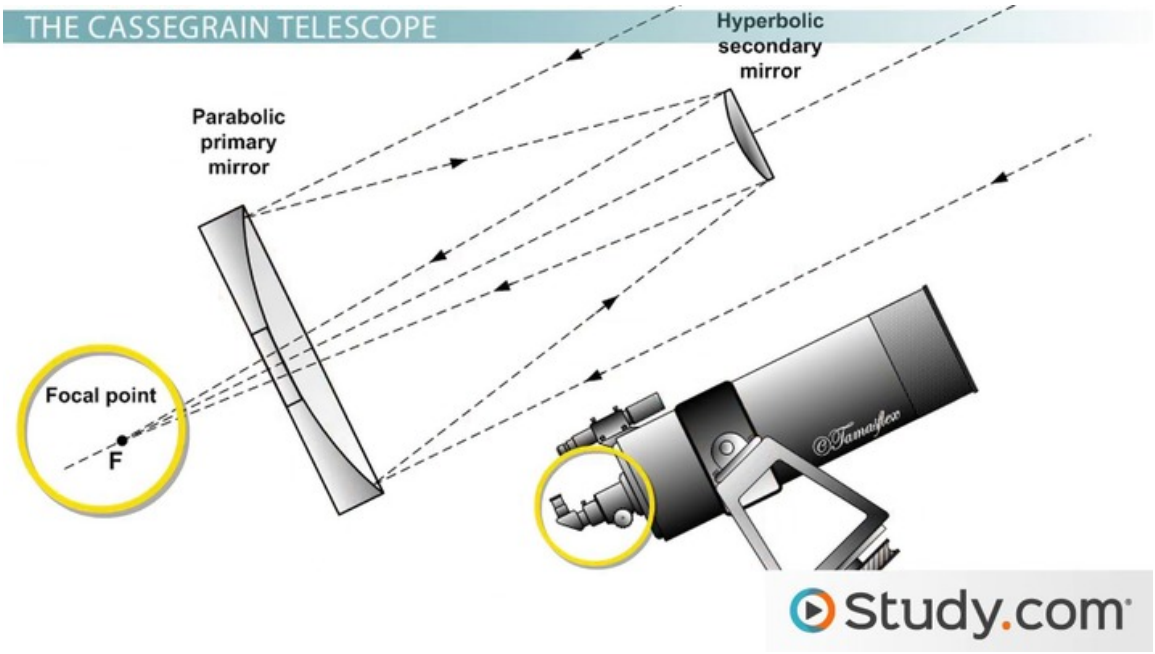


Figure 1.8: Reflector telescope, cassegrain model. The direction of the light is indicated by the arrows on the black dashed lines. They reach first the parabolic primary mirror, that concentrates them onto the hyperbolic secondary mirror. From it, the light rays arrive at the focal point where the camera would be located. Credits: Wikipedia

system is capable of making a picture, where the camera shutter is closed so that no light is entering. They are applied by subtracting the average of the taken bias frames.

Secondly, CCD cameras have an internal noise because of their semiconductor nature. The electrons in the internal lattice can get excited due to the systems' thermal energy. The CCD cameras are therefore under a cooling system to reduce the thermal electrons. Nonetheless, the dark current is still present for low temperature, and in order to correct for the dark noise that arises during the exposure time it is necessary to generate the Dark calibration frames. They are created by performing camera exposures of similar time length as the planned science exposures. This is because the dark current's production of electrons at low temperatures increases linearly with time in first approximation. Then, the Dark frames are averaged and subtracted from the images.

Next is the Flat field exposures, where the goal is to address the slight difference in photoelectric efficiency from pixel to pixel. The procedure is to expose the CCD to a flat field so that all the pixels detect an equal amount of light. They are applied as a factorized weight that corrects for this heterogeneous detection. Creating a perfect flat field is generally quite difficult, but they can be considered good enough for practical purposes. There are two typical options of doing it: one is by illuminating a section of the telescope's covering dome with a homogeneous lamp, what is known as dome flat fields; the other is by exposing the camera to regions of the sky that are illuminated homogeneously in first approximation during twilight or dawn.

1.4.2 Sky Background Subtraction

The next step is the sky background subtraction. The goal of the sky background subtraction is to eliminate light pollution from any sources that are not the astronomical objects themselves. In the visible electromagnetic range, this background can be produced by multiple phenomena. One is the reflexion by the atmosphere of any surrounding element

such as the moon, environmental light around the telescope, natural air glow from molecular or atomic oxygen lines (Broadfoot and Kendall, 1968; Massey and Foltz, 2000), and even from the combined effect of all astronomical sources on the atmosphere (Leinert et al., 1998). Nonetheless, the sky background can come also from internal reflections within the telescope – it is worth noting this does not refer to the fringe patterns that arise in narrow-band photometry, as narrow bands were not used for this work and thus not considered in this introduction – or from bright sources affecting the camera itself, even when they are not observed in the focal plane (e.g. Watkins et al., 2024). There exist other astronomical sources outside the atmosphere that can affect the background. One is the zodiacal light, made by the glow of the solar system's dust distributed across the ecliptic plane, when illuminated in the side of the sky opposite to the Sun. Another is the Galactic cirrus, dusty clouds within the Milky Way that can also affect the low surface brightness regime.

There are multiple methods to calculate the sky background of an image. The general concept consists of characterizing the image ADU level in pixels that can be considered to not be illuminated by any source. This can be done by masking pixels belonging to a source detection, then calculation the background level by a mean or median of the pixel counts. Because the background pixels are generally more abundant than the source pixels, at least outside the regime of surveys focused on LSB targets, even the median itself can be a good enough approximation in many cases. More sophisticated scientific goals require of addressing the spatial dependence of the background by fitting a 2D polynomial instead of using a global statistic.

The method used in SDSS DR6 to calculate the sky level is the clipped median, that consists of calculating the sky level and standard deviation σ, then discarding any pixels brighter than some factor of σ, and remaking recursively the calculation. A clipped median with enough background pixels available converges to the sky background level after several iterations, providing a mean and σ model for a Gaussian background distribution. Because the background can show spatial variability, the SDSS DR6 clipped median was calculated in pixel boxes across the image to generate a background map of the image.

The sky background subtraction is crucial when identifying galaxy mergers, because the tidal features and stripped material that characterize them are in a Signal-to-noise (S/N) regime that can be affected by an over-subtraction. However, the background subtraction can influence many more aspects of the images. For instance, the background modelling in SDSS DR6 showed the problem that large galaxies had their flux, size, and concentration underestimated (Blanton et al., 2005; West, 2005; Hyde and Bernardi, 2009; West et al., 2010). This appeared to come from calculating the background without masking the detected sources, that resulted in an over-estimation and was corrected for the Data Release 8 (DR8) in Blanton et al. (2011).

1.4.3 Source extraction

Another crucial part of the analysis is to detect the sources present in the image. The most basic algorithm, implemented in the most generally used source detection software programs, such as Source Extractor (Bertin and Arnouts, 1996) or `statmorph` (Rodriguez-Gomez et al., 2019), or in the SDSS and HSC pipelines, are the segmentation maps. Segmentation maps are based on finding groups of pixels that arise above the background. They require setting the background level and a measurement of its standard deviation σ, then defining all groups composed of a minimum number of pixels with ADU a factor of σ above the background. Therefore, the segmentation requires knowing the background.

Source detection has further complications. One is the possibility of having two or more actual astronomical sources in the same segmentation map. This is known as blending, and solving it requires of disentangling the multiple detections within the parent source.

Deblending is not an easy problem to solve, and it is in fact one of the main challenges of astronomical imaging. As the new telescopes and cameras manage to observe further and dimmer, the amount of sources that are observed increases, and it becomes more likely that they overlap with each other (Melchior et al., 2021).

1.4.4 Point Spread Function

When the three calibration frames are applied, the effect of the optical system can be addressed through the extracted sources. The combination of internal diffractions is modelled with the Point Spread Function (PSF). A point source observed by the telescope is deformed according to the optical system the same way as any other source, but point sources can be modelled as an initial 2D delta function so that the modelling can be extrapolated to the rest of the sources. Moreover, for ground-based telescopes, point sources suffer from the effect of turbulences in the atmosphere. This produces a Gaussian-like shape to the light distribution that contributes to the instrumental PSF.

In astronomy, a point source is a star, because their angular size is small enough so that it can be considered as a point source entering the telescope. The procedure to define the PSF consists on selecting stars along the image, characterize their shape, and define an empirical PSF. The PSF is not corrected from the images by itself, but rather used as a kernel that can be applied to correct for the shape deformation when calculating fluxes or shapes.

It is important to mention that the brightest objects in the sky, such as well known stars, can spoil the image by saturating the pixels. Saturated pixels have their electrons depleted during the time of exposure, reaching a maximum of detection, and even excite surrounding pixels. Those bright stars therefore cannot be used for PSF estimation and instead become an artefact.

For the case of the HSC pipeline, the PSF estimation, source detections, and sky background are done subsequently multiple times. The sky background is subtracted prior to detecting the sources and making each PSF model. The specific steps on the pipeline, how they are implemented, and how I actually modified them are detailed in Sect. 3.4.2.

1.4.5 Photometric and Astrometric calibrations

Once the frames have been reduced and the sources detected, then the fluxes and the celestial coordinates of the image can be calibrated. These are the photometric and astrometric calibrations. The astrometric calibration searches for position correspondences of a set of sources in the image that can be identified and then referenced to those in already categorized catalogues. The sources used for astrometry are generally stars in a brightness range that does not saturate the camera and whose PSF has been characterized. Some of the well known stars that can be observed with the naked eye could be ideal for calibration, as their position is well determined, but the majority of them indeed saturate the camera's of the modern deep telescopes. The astrometry is calibrated as a ladder, mapping the stellar positions of the main stars from small telescopes, to subsequently dimer stars in deeper surveys. Moreover, because the Sun, the Earth itself, and the stars are in constant motion, a general reference for the observation time has also to be set. This has been generalized by defining an astronomical time epoch. The current convention adopted by the International Astronomical Union (IAU) is the J2000 date, which takes as reference moment the 1st of January of the year 2000 at 12h.

The photometric calibration has the goal of comparing the camera fluxes to a standard system that then can be used to extract relevant physical information. This is calculated by setting the telescope's zero point: the reference source with a magnitude equal to 0 in a given band. The zero point is generally defined by taking some standard source. The

standard source used historically is the star Vega, although many modern surveys use the similar AB system (Oke and Gunn, 1983). Therefore, by setting the zero point of all the sources in the survey to the standard catalogue, one creates the photometric calibration.

1.4.6 Coaddition

The calibration steps that have been discussed above have been referred to individual astronomical exposures. Nonetheless, these can be coadded into one single frame increase the depth. Coaddition is the combination of multiple science exposures through the overlap of images covering the same portions of the sky. The main advantage of coadding images is to reach an image depth equivalent to the combined time of the individual exposures, but using several shorter shots instead of a single long one. This has multiple motivations: to reduce noises that increase with time, to avoid saturations that would arise during too long exposures, to correct for time-dependent background, to avoid transiting objects as satellites or asteroids, and to allow the parallel performance of time-domain science studies.

Coaddition is generated by taking all the images that overlap in the positions of the sky, then making use of the reference coordinates acquired from the astrometric calibration to build a coaddition grid, and finally obtaining the ADUs per pixel by applying some averaging statistics between the pixel values of the individual exposures. For HSC, the coaddition is done by the weighted average of the frames, through a per-pixel weight calculated from the inverse of the mean variance.

1.4.6.1 Dithering

One practical advantage of coaddition comes from applying a method called dithering (Tyson, 1990). This is accomplished by pointing the multiple exposures used for coaddition to different points of the sky, so that the pixels do not observe the same source all the time. By combining systematically shifted frames, it is possible to identify static signatures coming from the telescope and camera, or attenuate the time-dependent background. Due to dithering strategies, the sources in the sky are detected by multiple pixels, reducing the impact of inhomogeneous photoelectric efficiency, or avoiding that pixels with internal problems always observe the same sources. Specifically, for the HSC data reduction that has been employed during the thesis, the dithering makes it possible to identify the static background produced due to the camera's filter response to the sky along the focal plane, as described in Sect. 4.1 of Aihara et al. (2019). The sky background models measured through dithering have been tested as a solution to avoid over-subtraction of LSB structures in multiple works (Duc et al., 2015; Trujillo and Fliri, 2016; Watkins et al., 2024).

1.5 Thesis Contents

This thesis is focused on developing methodologies to find galaxy mergers in large sky surveys. The amount of images, the size of the data, and the number of galaxies observed is going to increase unprecedented in the upcoming years. The Large Survey of Space and Time (LSST; Ivezić et al., 2019), which will be carried by the Vera Rubin Observatory now under construction, will observe as many galaxies in a night as SDSS in a decade. Thus, optimization in the identification of mergers will be crucial to make a proper use of this bast information. Visual inspection will become unfeasible for them, but even cross-pair or ML-based techniques will need to overcome new challenges to keep up. How can we efficiently find galaxy mergers in big sky surveys? In this thesis I will show how I have developed new ways of finding merging galaxies.

I began by seeing if photometry could be enough to find mergers through ML-based models. Photometric measurements are one of the primary outputs of the image analysis, so taken advantage of them for merger classification could be highly beneficial. I built a NN and selected a catalogue for training it to discern between mergers and non-mergers. The galaxies and photometric measurements used came from images of SDSS DR6. How I used this data and discovered that the sky background error as a quite promising measurement to find mergers is described in Chapter 4. Similarly to a morphological parameter, the sky error provided a separation between mergers and non-mergers by a boundary, when representing its value for the training sources in a plot.

Such discovery has the potential of becoming a new tool for merger identification, but there are still two challenges that it would need to overcome before. One would be to show a reduced contamination by non-merging galaxies when applied to a dataset more general and heterogeneous than the training sample. The other extension would be to have it applied robustly to other surveys which would have different image properties and depth, and lack a sky error parameter computed in the same way as for SDSS DR6.

The extension to a larger dataset within SDSS DR6 was carried out by taking all galaxies of GZ DR1, from which the NN training galaxies were obtained. This work is described in Chapter 5. The training set had specific types of mergers, and the relative amount of merging and non-merging sources was not representative of the real sky. I visually inspected sources on the sky error parameter space, noting both the galaxy class and the potential features that may confuse the classification. These contaminations were indeed found, and I addressed them through a decision tree that discards them according to certain conditions.

The extension to deeper sources was done on HSC images, and is described in Chapter 6. As the output photometry in HSC does not have a sky error parameter like SDSS, I designed a technique to characterize the surrounding of galaxies. The goal of this is to reproduce what the sky error parameter does. Besides, I run the HSC data reduction pipeline in order to make sure I analyse the sky without having the sky background modelling modifying it too much.

If the model fulfils these two extensions successfully, then a sky background error methodology could be applied to the next generation of large-scale surveys such as LSST. The text is structured by describing the individual parts of each project together, splitting into the Chapters addressing the Data (2) and the Methodologies (3) of this work. Then, the thesis is structured in three major stages: the NN method that allowed me to unveil the importance of the sky background error, in Chapter 4; the visual inspection and decision tree applied to extend the method to the rest of SDSS DR6-GZ DR1 galaxies, in Chapter 5; and the analysis of mergers in HSC-NEP, first by the visual inspection of ML-based mergers that resulted in the catalogue of Pearson et al. (2022), and finally by the LSB features around HSC galaxies, for the extension to deeper data of the sky error method, in Chapter 6. The thesis concludes with its Summary in Chapter 7.

CHAPTER 2

Data

Different sets of data were employed for creating the galaxy catalogues we studied. The main goal of this chapter is to introduce the telescopes we worked with, and convey the motivation of the data selections. We built catalogues in the Sloan Digital Sky Survey Data Release 6 (SDSS DR6; Adelman-McCarthy et al., 2008) and in the Subaru/Hyper Suprime-Cam (HSC) in the North Ecliptic Pole (NEP). The data catalogues from galaxies imaged in SDSS were visually classified in the Galaxy Zoo Data Release 1 (GZ DR1; Lintott et al., 2011). The catalogues built out of galaxies from HSC were morphologically classified in Pearson et al. (2022) and in Galaxy Zoo: Cosmic Dawn (GZ:CD). We made selections out of the GZ DR1 and GZ:CD sets for testing galaxy merger identification methods in the two depth regimes of SDSS DR6 and HEC-NEP. The properties that characterize the data and the relevant measurements for this thesis are also described.

2.1 Sloan Digital Sky Survey Data Release 6

The Sloan Digital Sky Survey (SDSS; York et al., 2000) is a photometric and spectrometric survey that has aimed to create detailed maps of the sky through a 2.5-metre telescope at the Apache Point Observatory in New Mexico, USA. It began in 2000 and has undergone multiple phases, each expanding its scope and depth, starting from SDSS-I through the current SDSS-V. Each phase provided updates of the instruments and techniques used. SDSS has collected data on millions of objects, providing precise measurements of their positions and fluxes, and it can be considered one of the first large-sky astronomical surveys.

The SDSS galaxies we worked with were observed in the SDSS Data Release 6 (DR6; Adelman-McCarthy et al., 2008), that was within the SDSS-I phase. The telescope's camera, as shown in Fig. 2.1, had its CCDs distribution in six columns, each of them with five rows corresponding to the five SDSS photometric filters: u, g, r, i, and z (Gunn et al., 1998). The SDSS observing strategy consisted of scanning the sky with the 3 arcminute focal place of its camera, drifting the telescope pointing along stripes. From the calibrated images, a selection of spectroscopic targets using the estimated photometry was created. The SDSS-I spectroscopic galaxy targets were selected if they had a Petrosian magnitude – see Sect. 2.1.1 – in the r-band smaller than 17.77 ($r_{\rm petro} > 17.77$), although other sources such as quasars or stars were also observed spectroscopically. Some of the galaxies fulfilling the target criteria could be discarded due to complications such as nearby bright stars or target blending. The position of the targets was registered, and it was used to drill the positions for the SDSS fibres in plates crafted to cover the sky in each exposure. The fibres were inserted one by

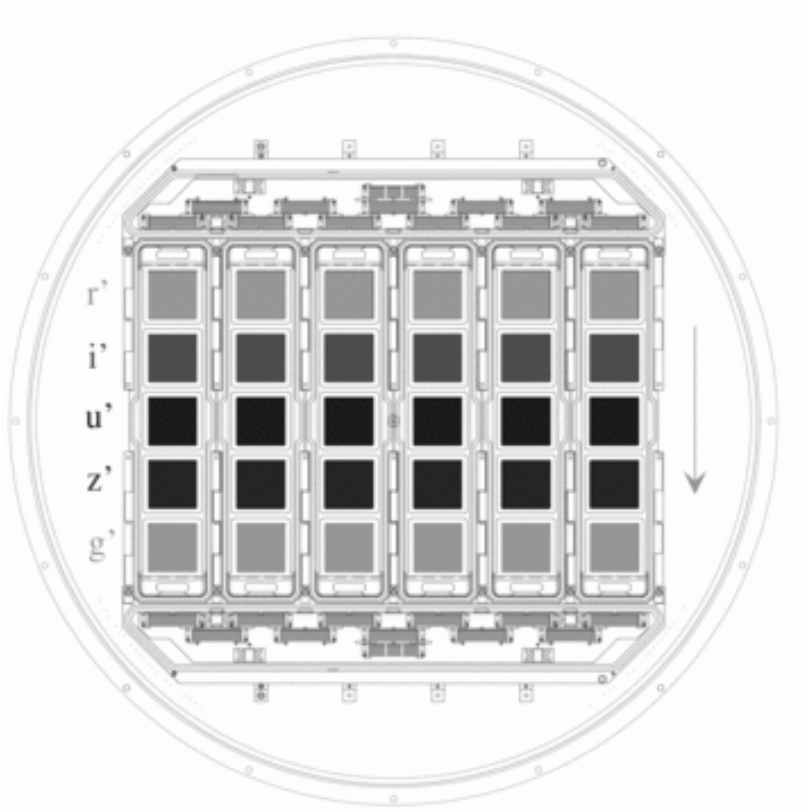


Figure 2.1: Schematic of the SDSS camera in the focal plane shown in Gunn et al. (1998). It depicts the 5×6 CCD cameras distributed in five rows for each photometric band.

one into the holes drilled in the plates, and the telescope took data of the 640 sources per plate simultaneously.

The SDSS DR6 survey reached a magnitude limit of 22.0 in the u, g, and r bands, of 21.3 in i and of 20.5 in the u-band. Besides, the pixel size of the camera was of 0.396 arcseconds, and the median PSF width across the survey was of 1.4 arcseconds. This makes it the most swallow dataset we worked with during this thesis. This is important, because the higher the magnitude limit, the smaller, dimer and more distant galaxies can be observed. Moreover, higher depth also implies more sources and more density in the sky-plane projection, i.e., more blending of sources.

The SDSS DR6 observed 9.5k squared degrees of the sky, detecting almost 240 million sources. Out of them, 800 000 galaxies had spectroscopic observations. The catalogues of galaxies that we built were obtained from these sources. However, we needed for our purposes the morphological classification of those galaxies. For that, we made used of galaxies classified in Galaxy Zoo (GZ; Lintott et al., 2008).

2.1.1 SDSS Photometry

The photometric magnitude of galaxies in SDSS is obtained for its five bands *ugriz*. These magnitudes are calculated by an asinh magnitude function (Lupton et al., 1999) from the measured fluxes. Here we include the definition of the main SDSS magnitudes (Stoughton et al., 2002) used for our experiment, as described in Suelves et al. (2023):

Point-spread-function flux: is calculated by fitting the point spread function (PSF) – see Sect. 1.4.4 – interpolated to the source position. The PSF, with its spatial variation for each image frame and band, is measured by the SDSS's pipeline.

Fibre flux: the flux contained inside the an aperture of the same angular size as the fibres of the SDSS spectrograph. Those fibres cover circular apertures of 3 arcseconds in diameter. In order to simulate better what the fibre sees in reality, the images are convolved with a 2-arcsecond seeing prior to the flux measurement. Appendix A describes how the magnitude and errors are derived from the aperture counts.

Petrosian magnitude: the flux is calculated inside an aperture of radius r_P. This radius is determined by forcing the inner flux to be a fixed factor – the Petrosian ratio R_P – of the mean flux on the circular annulus at the same r_P. The aperture r_P set in the r band is applied for the other four bands, so that the measurement is within a consistent aperture.

Exponential profile fit: the flux is obtained from the fit of the galaxy's brightness distribution with an exponential profile, convolved with the PSF. The 2D exponential profile depends on the radius from the centre r as $I(r)$, the surface brightness at r:

$$I(r) = I_0 \exp\left[-1.68\frac{r}{r_0}\right] , \tag{2.1}$$

where I_0 is the flux at $r = r_0$, the half-light radius.

De Vaucouleurs profile fit: fit with a De Vaucouleurs profile, convolved afterwords with the PSF. Its form is:

$$I(r) = I_0 \exp\left[-7.67\left(\frac{r}{r_0}\right)^{\frac{1}{4}}\right] . \tag{2.2}$$

Model magnitude: standard magnitude employed for SDSS data, computed by a linear combination of the best fits with the exponential and the De Vaucouleurs profiles:

$$F_{\text{model}} = frac_{DeV} F_{deV} + (1 - frac_{DeV}) F_{exp} , \tag{2.3}$$

where $frac_{DeV}$ is the linear combination factor that leads to the best possible fit of the profile combinations.

2.1.1.1 Sky Background Error

The sky background error (`skyErr`) parameter in SDSS DR6 was calculated in the background measurement of the calibrated images. This sky background was estimated in 256×256 centred every 128 pixels across the image Euclidean grid. For each box, the background level was calculated by a clipped median with a clipping factor of $2.32634\times\sigma$, masking only the brightest and saturating stars.

Once the background in each box was estimated, the error was obtained from the interquartile range of the resulting background distribution. The sky level and the sky error were linearly interpolated from the box centres to the centroid of each source. The sky background and error were obtained initially in the units of pixel counts, Data Units (DN), and then transformed to the flux unit `maggie` for the CasJobs repository portal[1]. Maggies are linear flux units that corresponds to 3631 Jy. One maggie is defined to have an AB magnitude of 0, so that they are calibrated according to the zero point of the images [2].

2.1.2 Galaxy Zoo Data Release 1

The Galaxy Zoo Data Release 1 (GZ DR1; Lintott et al., 2011) provided its volunteers with images of galaxies from a set of ∼600 000 out of the 800 000 spectroscopic target galaxies of SDSS DR6. Through its interface, the users were asked questions related to the shape and

[1] https://skyserver.sdss.org/casjobs/

[2] http://cas.sdss.org/dr7/en/help/docs/algorithm.asp?search=mag_model

visual properties, and the gathered answers resulted in a morphological classification. Each GZ DR1 vote was weighted based on the agreement rate of the volunteer that emitted the vote has with the majority opinion of all the objects they viewed. When the online program was completed, the catalogue with the resulting voted classifications were published.

The main GZ DR1 catalogue provided three morphological labels defined from the relative amount of votes. These labels are Elliptical, Spiral, and Uncertain. They were defined after correcting the volunteer votes for the difficulties of discerning spiral galaxies from ellipticals in the faint, small, or far regime. This is because spiral galaxies imaged with low resolution appear similar to elliptical ones. Such low resolution can be because the spiral appears faint in the image and the spiral features are not obvious, or because it is small so that the spiral features are not resolved. This gets enhanced if the galaxy is far: if one could observe the exact same source at redshifts 0.1 and 1, it would become fainter and smaller for the higher redshift.

The correction was performed in Bamford et al. (2009) and reproduced in the final dataset (Lintott et al., 2011). This bias was calculated by building magnitude-size bins, estimating the spiral-to-elliptical ratio for each bin, and assuming this ratio to be constant along the whole survey using the better resolved bin as reference. Thus, every galaxy received a correction factor for the elliptical and spiral votes, i.e., debiased vote fractions. It is important to note that this factor is only related to the limitations of the image themselves, and a different survey would require a different bias estimation. Other biases, such as those coming from the volunteers or the actual morphological dependence on cosmic history, were not included. Each galaxy that had a debiased vote fraction for elliptical or spiral classes above 0.8 were flagged as Elliptical and Spiral galaxies, and those that did not fulfil those two conditions were flagged as Uncertain galaxies.

Regarding merging galaxies, one of the options that were shown to the volunteers asked about the presence of merging signatures. Thus, the galaxies in the catalogue received a merger vote fraction f_m. These votes were not debiased as the votes related to early-versus-late type options explained above. The work in Darg et al. (2010a,b) provided a follow-up catalogue of mergers, where the authors confirmed visually the merging galaxies and selected their merging companions. The mergers they checked were galaxies with $f_m > 0.4$. The released public catalogue[3] is composed of merging galaxy pairs, including some multi-merger cases, with a total of 3 003 merging systems in the redshift range [0.005, 0.1].

One drawback of the GZ DR1 merger votes emitted by the volunteers was that only major merging clear features were explained to them. Thus, it is possible that some less usual merging types such as minor mergers, shells, or coalescing galaxies got low merger votes.

2.1.3 Photometric NN's training dataset

The dataset used for training a Neural Network (NN) is required to contain a representative sample for each different class intended to be identified. We attempted to discern merging and non-merging galaxies using the SDSS DR6 sources that were morphologically determined in GZ DR1.

Our selection of galaxy mergers for training was taken from the Darg et al. (2010a) merger catalogue. To complete the dataset, we needed to include non-merging galaxies. As a preliminary non-merger set, we considered all GZ DR1 galaxies with $f_m < 0.2$ (as in Pearson et al., 2019b). We built the training set to be class balanced, that is, each class that the NN learned to recognize was equally represented in the training data. Using a training set reproducing the real abundances, the NN would become very good at finding non-mergers, but might not learn and confuse some of the less abundant mergers, thereby

[3] https://data.galaxyzoo.org/#section-4

contradicting the purpose of this work. Besides having a class balanced set, we wanted to have a distribution of merging and non-merging galaxies similar in mass, so that they represent comparable populations in every property that is not related to the merging process. For example, if all mergers were significantly more massive than the non-mergers, the NN could be identifying the mass distribution rather than the merging state. While the *r*-band photometric magnitude is a good proxy for galactic mass (e.g. Mahajan et al., 2018), it is a detected flux that depends strongly on the distance. We thus considered the spectrometric redshift (spec-z) as a proxy of the source's distance. We obtained one non-merging nearest neighbour for each merging galaxy based on the model magnitude *r*-mag and the spec-z through a 2D Euclidean distance.

The galaxies had spec-z calculated from the fibre spectroscopy described above, and flux measurements that resulted in the magnitude types described in Sect. 2.1.1. We applied a cut in the spec-z < 0.01 and another cut in the model magnitude *r*-mag < 18.05; the former was because zero non-mergers could be found below that redshift, and the latter because the density of non-mergers decreased with *r*-mag. From the original 3 003 primary mergers, four of them were found to show $f_m < 0.4$, of which three even showed $f_m < 0.2$, which meant that they appeared in both initial samples. Thus, these four galaxies were removed.

The final dataset was reduced to 2 930 mergers after all the restrictions were applied. Therefore, the 2 930 non-mergers were matched and a final full dataset of 5 860 galaxies was obtained. From this, we separated 5 360 class-balanced galaxies for training-validation purposes, and kept the remaining 250 mergers and 250 non-mergers for testing the trained NNs.

2.1.4 Decision Tree dataset

The training dataset defined for training the NN led to useful insights to find mergers using the SDSS DR6 sky background error, described in Chapter 4. In order to extend this sky-background-based merger identification to more galaxies, we made use of the whole GZ DR1-SDSS DR6 catalogue, and split it.

Multiple subsamples of these galaxies were selected for visual inspection. These subsamples were groups of nine galaxies taken from subsets of the GZ DR1 dataset as defined by four properties: by the model magnitude in the *r* band, by GZ merger vote fraction f_m, by GZ morphology flag, and by the location in specific areas of the Fig. 5.1 boundary. We chose of taking nine galaxies in order to get good statistics without compromising too much time in the visual inspection. The first three selections are described below, and the last property is explained in Chapter 5. The nine galaxies per subsample were randomly drawn from each of the four-parameter GZ DR1 subsets.

It is important to insist on clarifying the nomenclature across this dataset: a subset comes from the four-parameter-based selection from the GZ DR1 dataset, a subsample is the nine-galaxy groups drawn from the subsets.

Magnitude: when doing a preliminary inspection of the galaxies along the Fig. 5.1 sky error diagram, it was clear the importance of the magnitude of the galaxy in the visual appreciation of merging features. Three magnitude intervals were defined: the bright interval of magnitude values below 14, the intermediate one from 14 to 16.5, and the dim interval for galaxies of magnitudes above 16.5.

Merger vote fraction f_m: we took advantage of the vote fraction bins employed in Darg et al. (2010a) to determine four possible ranges from the minimum to the maximum votes. The $f_m = 0$ galaxies should be the safest non-merging value, as none volunteer considered them to be interacting. The galaxies with $f_m \in (0,0.2]$, considered as merger by a minor portion of volunteers, might hide some features. The $f_m \in (0.2,0.4]$, interval that in Darg et al. (2010a) was considered as intermediate between mergers and non-mergers, is even

more likely to show mergers that are not major merging pairs. Finally, the higher interval $f_m \in (.4,1.0]$ is where the mergers from Darg et al. (2010a) were taken, and therefore where the majority of merging pairs should appear.

Morphological labels: we considered galaxies showing the three GZ DR1 flags: Elliptical, Spiral, and Uncertain. It is important to note that for Elliptical and Spiral flags f_m is by definition smaller than 0.2, while for Uncertain sources f_m has values along the whole [0,1] interval.

2.2 Subaru/Hyper Suprime-Cam deep field in the North Ecliptic Pole

The HSC is a wide field camera mounted on the Subaru Telescope, with a primary mirror of 8.2 metres, located near the summit of Maunakea, in Hawaii, USA. Opposite to SDSS, it has a filter wheel that allows the multiple optical bands to be rotated. Of these bands, we worked mainly with the g, and r, which are similar to the SDSS ones. It has 116 CCDs, out of which 104 are used for the actual images, 8 for the guiding system and 4 for the focus system[4]. Each of the CCDs has its pixels separated into four vertical channels.

HSC is a wide field camera with a field of view of 1.8 degree in its focal plane. It has been mainly used for creating a large survey of the northern sky. This wide survey is known as the HSC Subaru Strategic Program (HSC-SSP; Aihara et al., 2018). Besides the SSP and it's up to now three data releases, HSC survey also includes deep and ultradeep fields. The deeper the field, the more the coadded exposures that were taken in it.

The North Ecliptic Pole (NEP), observed as one of the HSC deep fields, is the field where Pearson et al. (2022) was carried out. This HSC-NEP deep optical observations (Goto et al., 2017; Oi et al., 2021) were designed to cover the AKARI-NEP infrared wide field survey (Matsuhara et al., 2006; Lee et al., 2009; Kim et al., 2012). The observation was carried out in two rounds. The r-band images were observed in the 1st of July 2014, during a night with high air disturbance. The remaining g, i, z, and y-bands were observed later, between the 8th and 11th of August 2015, under better conditions. The atmospheric turbulences during the r-band observations made its seeing worse than that of the other four bands. The resulting r-band seeing was 1.25 arcseconds and for the rest of the bands, between 0.7 and 0.8 arcseconds. The images used were reduced using the HSC photometric pipeline `hscPipe` version 6.5.3 (Bosch et al., 2018; Aihara et al., 2019).

2.2.1 Galaxy Zoo GAMA-KiDS and Galaxy Zoo: Cosmic Down!

The Galaxy Zoo GAMA-KiDS was carried out on the images of 50 000 images from the Kilo-Degree Survey (KiDS; de Jong et al., 2013a; de Jong et al., 2013b), out of galaxies observed in the Galaxy and Mass Assembly field (GAMA; Driver et al., 2009; Holwerda et al., 2019)[5],[6].

The Galaxy Zoo: Cosmic Down (GZ:CD) program was based on cutouts from the NEP images taken in the 2019 HSC deep field observations of the Hawaii Two-0 (H20) survey (Zalesky, 2021). The classification was performed by GZ volunteers during 2023 and 2024. Similarly to previous GZ programs, the volunteers answered questions regarding merging interactions, although GZ:CD also included major and minor disturbances, in contrast with GZ DR1.

[4] https://hsc.mtk.nao.ac.jp/pipedoc_5_e/hsc_info_e/index.html#hsc-info

[5] https://www.gama-survey.org/

[6] https://indico.in2p3.fr/event/16341/sessions/9954/attachments/49229/62439/Kelvin_20180612-Lyon.pdf

2.2.2 HSC-NEP training dataset

The work in Pearson et al. (2022) classified galaxy mergers by a Deep Learning (DL) model built by my auxiliary supervisor dr. William Pearson, published in Pearson et al. (2022). For training, it used classifications from GAMA-KiDS Galaxy Zoo. The training was done on observations from the HSC-SSP Data Release 2 (DR2; Aihara et al., 2018, 2019). Therefore, the training galaxies were cross-matched from the GAMA-KiDS-based training set in Pearson et al. (2019a) with the HSC-SSP DR2 sources, providing a smaller dataset than that in Pearson et al. (2019a).

To read all the details of how the dataset was created, please refer to the original work (Pearson et al., 2022). The training of the model, its application, and the visual inspection that I contributed to, were all performed on the *r*-band observations, despite the worse seeing quality.

The model was separately applied in two redshift ranges: a low redshift range for $z < 0.15$, and a high redshift interval $0.15 \leq z < 0.30$. The galaxies from the Pearson et al. (2019a) training dataset belonged to the low redshift range, but there is not a catalogue of mergers in the high redshift range. Thus, to train the $0.15 \leq z < 0.30$ galaxies, a data augmentation was applied to the low-redshift galaxies, modifying them to replicate a higher redshift observation. Again, full details on this can be found in Pearson et al. (2022).

2.2.3 HSC sky error extension

We matched the GZ:CD catalogue with the galaxies from the HSC-NEP catalogue created out of the multi-band optical images (Goto et al., 2017; Oi et al., 2021), reduced in Kim et al. (2021). We applied this GZ:CD classification to the galaxies *g*-band HSC-NEP images, although we carried out our own data reduction using the HSC photometric pipeline `hscPipe` version 6.7.

The raw images and calibration frames were downloaded from the SMOKA repository[7]. SMOKA is a web repository developed and maintained by the National Astronomical Observatory of Japan (NAOJ). It includes "public science data obtained at the Subaru Telescope", among others. For HSC, the calibration frames are created monthly, as they are quite stable within long periods of time. For reducing the *g*-band HSC-NEP images taken in July 2015, we made use of the 2015 May bias, dark and dome flat frames. We were suggested to use the May 2015 frames instead of the July ones, as there were some temporal issues for finding the July frames, and the practical difference between them was small.

The basics of the first version of the pipeline, the initial version 4, are described in Bosch et al. (2018). However, a new background estimation method, together with other updates, was introduced for it the `hscPipe` version 6, as described in Aihara et al. (2019). We provide a detailed description of the pipeline in Sect. 3.4.2.

We created a class-balance dataset from the GZ:CD. The galaxy mergers were selected by considering as mergers every galaxy with merging votes above 0.7, following the criteria of the previous GZ release (Walmsley et al., 2022). Analogously to the training dataset for the NN in the first project, we matched each merger with its nearest neighbour, drawn from the galaxies with merging vote equal to zero, in the plane formed by the *r*-band magnitude and the photometric redshift. Because not all the HSC-NEP galaxies have spectroscopic redshift available, we considered the photometric redshift a good initial proxy of the distance. The final catalogue contained exactly 256 mergers and their 256 corresponding non-mergers.

[7] `https://smoka.nao.ac.jp/HSCsearch.jsp`

CHAPTER 3

Methodology

Multiple methodologies were applied along my PhD studies in the work presented in this manuscript. For the Neural Network (NN) applied to our training-set galaxies in the Sloan Digital Sky Surveys (SDSS) – see Sect. 2.1.3 –, we normalized the input parameters in a wide variety of ways, and study them with dimensionality reduction methods. For the visual inspection applied to the Galaxy Zoo Data Release 1 source selection – see Sect. 2.1.4 –, we built a set of questions to classify them. We also built a decision tree, following what was found in the visual inspection, created to automatically select galaxies with surrounding detections that might be contaminating the merger identification. Multiple methods were applied to the Hyper Suprime-Cam North Ecliptic Pole (HSC-NEP) galaxies – see Sects. 2.2.2 and 2.2.3. We contributed to the visual inspection of the classifications generated by the Machine Learning method of Pearson et al. (2022). We also reduced the HSC-NEP g bands images. We finally obtained the pixels within an aperture around the galaxies and measured the background within them. Then, we generated parameters over the histogram of the background pixels, and applied further dimensionality reduction methods to those parameters. The description of the methods applied around the photometric NN, in Sect. 3.1, is described as in Suelves et al. (2023).

3.1 Photometric NN

An NN has combinations of layers, composed of basic mathematical nodes called neurons, connecting the input values to the model's final outputs. Each neuron weights the value of a previous point on the network – which can be all the dataset inputs or all the neurons in the previous layer – by its internal weight $\boldsymbol{g}$ and bias $\boldsymbol{b}$ parameters, performing a linear transformation of the type $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}$ that subsequently goes through a $g(\boldsymbol{f}(\boldsymbol{x}))$ non-linear activation function. $g(\boldsymbol{f}(\boldsymbol{x}))$ simulates a synaptic-like step in which the output is either zero or very small – no connection –, or a larger number that allows the information to 'progress' across the NN structure in some measure. The final output of the NN gives information about the input data such as their classification among a list of classes.

The NNs in this work were trained with known class labels – supervised learning –, so that the neuron weights could be progressively modified and the models learnt to solve the task for which they had been built. Such a process may fall into a situation in which the NN explicitly memorizes the training dataset, a state known as overfitting. Neural network studies are mainly focused not only on how to increase the learning abilities, but also on how to reduce this overfitting and manage the generalization of the trained result.

One of the controversies with NNs is that they tend to be treated as black boxes that somehow solve problems, even though it is not well understood how they manage them and why. This project centres its attention not only on testing our NN models, but also on determining the information the NN finds in the dataset, and for that we made use of some other techniques such as dimensionality reduction, which we also explain in this section. The goal of the NN is not only to classify mergers but also to learn their properties, which can only be achieved by understanding its internal functioning.

3.1.1 Basics of our NN

An NN is essentially defined by the layer-neuron architecture and the properties of the connections. We used dense layers, linking each of one layer's nodes to all those of the previous layer, creating subsequent, fully connected layers from input to output. In the intermediate step between layers, we applied batch normalization (Ioffe and Szegedy, 2015) – to normalize and shift the post-layer value array $\boldsymbol{l}$ to a mean $\bar{\boldsymbol{l}} = 0$ and variance Var($\boldsymbol{l}$) = 1 – in order to make it faster and more stable. Moreover, we applied a dropout rate (Srivastava et al., 2014) for each layer, which consisted in setting some arbitrary percentage of the neurons to zero during each training step. This was done to allow all the neurons to be relevant in some way, forcing them to not rely on a higher influence from others.

The non-linear neuron activation function selected was the commonly used Rectified Linear Unit (ReLU Nair and Hinton, 2010). The NN output is addressed as a two-class classification, also known as binary classification. The final result is given as a softmax probability value for the merger ($P_{\rm mer}$) and non-merger ($P_{\rm nom}$) classes. It fulfils $P_{\rm mer} + P_{\rm nom}$ = 1. The optimization method chosen was the Adam method (Kingma and Ba, 2014), characterized by a dynamic learning rate. The loss function used for optimizing the classifier was the TensorFlow's BinaryCrossentropy class. The layer layout, the dropout rate, and the initial learning rate selected are presented in Sect. 4.1.1.

3.1.2 Training

The model was trained on the pre-selected combination of labelled objects detailed in Sect. 2.1.3. Training datasets are commonly separated into three groups: a training set that is used for the optimization process, a validation set that is studied parallel to the training but without affecting the learning steps, and a test set that is only considered when the training is finished, to check the model's capabilities. During our NN learning, the training updates were done over a batch of 64 galaxies, randomly shuffled for every training epoch. At the same time, the validation set was used as control sample: updates on the neurons were externally saved only when there was an improvement in performance over the validation set. Specifically, we considered that the performance improved when both the validation loss decreased and the validation accuracy increased with respect to the last save. The test set was left unseen by the NN until the end, acting equivalently to applying the NN to a fully new group of objects, except that this new group was drawn from the same distribution as training and validation.

For training the NN, we employed the k-fold cross-validation method (Stone, 1974; Rodriguez et al., 2010). It consists in separating the training dataset by shuffling objects randomly into k equally sized groups. One full training was executed k times so that every train-validation run had $k-1$ groups forming together the training-set and the remaining one as the validation-set, which was switched each time. As a result, we obtained k trained NNs that could show the model instability and at the same time give a more reliable performance test. Moreover, if the learned parameters of the NN happened to be almost the same for all folds, an average of them could be used. For our 5 360 training+validation

galaxies, we decided to split them into five k-folds, partly motivated to avoid extending the training time for too long. This adopted validation method will be referred to in the rest of the text as five-fold cross-validation. The advantage of applying our five-fold validation is that it allows us to compare NN training results with different inputs by the mean and standard deviation of the validation saved-peaks per fold.

3.1.3 Input space normalization

Inputs for NNs are generally normalized to facilitate the optimization process and simplify the numerical accuracy (Yu et al., 2005). For our inputs, we chose the min-max normalization. The original galaxy's feature array X^g was adapted to an [0,1] interval by applying the equation:

$$x^g = \frac{X^g - \mathrm{Min}(X^g)}{\mathrm{Max}(X^g) - \mathrm{Min}(X^g)} \ . \tag{3.1}$$

Different NNs were trained by separately using the different magnitude types described in Sect. 2.1.1. However, the photometric information not only consisted of magnitudes but also of the measurement errors or the ten colour indexes – derived by subtracting magnitudes in each band from one another. We considered multiple combinations of photometric information and labelled them as follows: an NN trained over an input space with only band magnitudes was labelled as B; with only colours was labelled as C; with bands and colours was labelled as BC; and with bands and colours and also with errors was labelled as BCE.

3.1.3.1 Variations of the error normalization

Additionally, for the BCE cases, we considered two more normalization formulas. This was due to the intrinsic relation of the errors σ_B with the magnitudes, and thus normalizing them while ignoring this relation might lead to losing information. A more explicit possibility would be to relate the error normalization σ_b to the measurement-error ratio.

While Eq. 3.1 was kept for bands and colours, two value corrections were used for σ_b, differing whether the post min-max band values b were involved or not. The first case was obtained by the proportion of the original errors Σ_B to the original band measures B, that is, the fractional error:

$$\sigma_b = \frac{\Sigma_B}{B} \ . \tag{3.2}$$

The other case was obtained by considering the pre Σ_B/B and post-normalization σ_b/b ratios to be the same, solving an equality between the two fractions:

$$\sigma_b = b\frac{\Sigma_B}{B} \ . \tag{3.3}$$

Neural networks with their input normalized with Eqs. 3.2 and 3.3 were labelled as BCEp and BCEn, respectively.

3.1.3.2 Min-max normalization of feature space, but not included fully

The multiple photometric parameters we combined to form the NN's input spaces had values varying by up to several orders of magnitude. When we min-max normalized them into the [0,1] interval, this sometimes distorted in some way the relation between them. An illustrating example can be found in how the band magnitudes showed different normalized values between the input spaces of the BCE and the B versions of NNs. For the

BCE case, the parameter with the maximum value was generally the u-band magnitude, and the minimum was the error in some of the five bands. For the B case, while the maximum was the same, the minimum was the magnitude in some other band. Therefore, when applying Eq. 3.1 to the BCE space, the resulting input values for the band magnitudes were close to 1 and for the colours and errors close to 0. However, for the B case one could have a value 1 for the u-band and 0 for the z-band.

To understand the role of specific parameters, it was necessary to isolate them. If one were interested in only studying the band's performance, min-max normalizing them alone would be the immediate option. Nonetheless, the values would by construction differ between the combined BCE and the isolated B input spaces, and making a comparison would not be easy. We attempted to mitigate this by getting the normalized values of a subset within a larger input set but discarding the non-interesting ones. As an example: to understand the role of the magnitude bands in the BCE space, we would perform the same normalization as for the BCE NN, but keeping only the band magnitude values and discarding the colours and errors after applying the min-max formula. We define the nomenclature of this type of input, using the bands' example, as follows: 'bands as if in BCE' or 'bands as if with colours and errors'.

3.1.4 Statistics on the NN results

To quantify the success and performance of the NN, we define the following four classification groups: the mergers classified correctly are regarded as true positives (TPs); true negatives (TNs) are the non-mergers homologous; false negatives (FNs) are mergers mistakenly identified as non-mergers; and false positives (FPs) are non-mergers mistaken as mergers. Moreover, we consider accuracy as the ratio of correctly classified objects with respect to the whole set size:

$$\text{Accuracy} = \frac{\text{TPs} + \text{TNs}}{\text{TPs} + \text{TNs} + \text{FPs} + \text{FNs}} \, . \tag{3.4}$$

For a single validation fold, the denominator would be the 1 720 galaxies that compose it. The mergers' correctly classified rate is given as TPs/(TPs+FNs) and the non-mergers' correctly classified rate is given as TNs/(TNs+FPs).

3.1.5 Dimensionality reduction

Dimensionality reduction techniques transform a high-dimensional set of data into a lower-dimensional representation. With the goal of simplifying a given problem or visualizing the data in a more adequate way, they attempt to maintain as much information of the original data as possible. In our case, we applied them with the purpose of visualizing the galaxies' distribution in 2D, reducing the original input dimensions.

We considered two different ways to do it, one of which was the principal component analysis (PCA; Hotelling, 1933), which is actually not a machine learning model but essentially a matrix diagonalization. The other way was the t-distributed Stochastic Neighbor Embedding (t-SNE; Maaten and Hinton, 2008; Van Der Maaten et al., 2009), which is more oriented to resembling the original distribution.

Principal Component Analysis: a linear method that performs a coordinate transformation of an N-dimensional dataset's feature space – where each dimension corresponds to a data variable – into a new N-dimensional orthonormal coordinate base. This new reference frame is composed of basis vectors called principal components. They arise as a consequence of diagonalizing the covariance matrix of the dataset and ordering the eigenvalues

λ_i. The absolute value of λ_i represents how high the variance is along the correspondent eigenvectors. Through PCA, one can obtain 2D or 3D plots with the projection of the data onto the directions with most feature variations.

t-distributed Stochastic Neighbour Embedding: a machine learning algorithm that reduces an N-dimensional space into a 2D one, while maintaining its statistical distribution as much as possible. This method calculates the relative probability of each pair of high-dimensional objects so that very similar or closeby points have high probabilities and very different ones have low probabilities. A similar probability distribution is initialized in 2D and, through minimizing the Kullback–Leibler divergence between both with respect to the point positions, an embedded map in 2D of the original space is produced.

3.2 Decision Tree

The goal of this project was to identify contaminants around galaxies that might be spoiling the identification methods. These contaminants can be stars, galaxies, or image artefacts next to the target galaxies. They can confuse the classification method, making it to misidentify the target as a merger when it is actually a non-merging galaxy. Two main methodologies were employed for characterizing the target galaxies and the potential contamination: the visual inspection, and a decision tree created to automatically identify some of the features found visually. We applied the cross-pairs method to identify the merging pairs that had spectrometric redshift measurements – see Sect. 1.3.2.

Moreover, the galaxies analysed were selected from the sky background error diagram described in Sect. 4.2.3. We determined the areas of the diagram populated by the mergers and non-mergers from the training sample. This provided reference regions of the diagram where those classes could be found. This was done applying the alpha shape algorithm, which allowed to draw a boundary around the denser clusters of the diagram.

3.2.1 Cross-pairs criteria:

In order to consider the galaxies to be merging by the cross-pairs method, we applied the criteria for relative distance Δr and relative velocity Δv analogous to those in Ellison et al. (2008). We measured the relative distance through the function `kpc_proper_per_arcmin` from the `astropy.cosmology` package. This function calculates the projection in the sky, in units of arcminutes, of the proper size of a kilo parsec (kpc) for a given redshift[1], – see the third paragraph in Sect. 1.3.2, where distances in cosmology were defined. Thus, we obtained Δr given the separation in the sky and the redshift of the target. We measured the velocity (v) directly from the redshift, as $v = z \cdot c$, where c is the speed of light in vacuum.

We considered consistently the same $\Delta v = 500$ km/s across all visual inspection options. However, a slightly larger relative distance than 80 kpc/h was instead chosen for the visual inspection option "One distorted galaxy, with a cross-pair nearby" – see Sect. 3.2.2. This value was $\Delta r = 100$ kpc/h, because we observed tidal distortions in the target galaxy in multiple cases of this classification type in the presence of a companion with Δr between 80 and 10 kpc/h.

3.2.2 Visual Inspection

For visual inspection, multiple samples of nine galaxies were randomly drawn from subsets of GZ DR1, as described in Sects. 2.1.4 and 5.1.1. Each galaxy was visually inspected

[1]Source code of the function: https://github.com/astropy/astropy/blob/main/astropy/cosmology/flrw/base.py

and classified following the flow chart in Fig. 3.1. We dub these galaxies from now on as target galaxies. The visual inspection was performed in practice using Aladin Sky Atlas[2], specifically the Desktop software (Bonnarel et al., 2000). Aladin allows loading images both from the computer or from the Collections its system has available, such as the SDSS DR9 survey images. Each catalogue was loaded from TOPCAT (Taylor, 2005), a software for reading easily `.fits` files. Each catalogue was loaded from TOPCAT into Aladin using the Simple Application Messaging Protocol (SAMP) interoperation system that allows to send datasets between both. The SDSS DR9 colour images available in the Aladin Collections were loaded as an interactive sky and the visual inspection of the catalogue galaxies was done on them.

Here are explained in more detail the options from the flow chart and the reasoning behind them. These options were added as new six columns to the `.fits` file catalogues:

3.2.2.1 Major Merging Pairs

Two Galaxies Clearly interacting: When the target galaxy was clearly interacting with another galaxy that also showed signs of tidal disruption, the target galaxy was considered to be a major merger.

One distorted galaxy, with a cross-pair nearby: In the case of the target galaxy looking distorted, the presence of a galaxy of similar apparent size fulfilling the cross-pair criteria, as described in Sect. 3.2.1, was considered as a sufficient condition for the target to be in a merging interaction.

Cross-Pairs: both galaxies of comparable size: If the target galaxy did not show a clear distortion, but there was a cross-pair galaxy of similar size nearby, the target galaxy was considered to be a major merger.

3.2.2.2 Other Mergers

Cross-Pairs: one galaxy looks like a satellite: If the target galaxy did not show clear distortion, but there was a cross-pair galaxy of clear smaller size nearby, the target galaxy was considered to be a minor merger, and thus to belong to the other mergers cathegory. It is important to note that the small companion rarely had spectrometric redshift available, limiting the application of this criterion to a minority of cases.

Galaxy with satellite infalling: When the target galaxy had signs of distortion due to a smaller galaxy nearby, this was considered to be a minor merger, even if the potential minor pair did not show clear signs of distortion.

Highly distorted coalescing galaxy: This was the case of highly irregular galaxies, and it was considered only for a high enough resolution that allowed to observe the coalescing features. Such features could be double nuclei with distorted surroundings or clear signs of an absorbed satellite galaxy.

3.2.2.3 Non-Mergers

Elliptical or Spiral clearly non-interacting: If the galaxy showed a well define early or late-type morphology, it was clearly a non-merger.

[2] https://aladin.cds.unistra.fr/

Galaxy in a crowded field, without cross-pairs: Some galaxies were bright, showed a large angular size, and appeared in quite crowded fields, blending with many smaller sources. Others were surrounded by galaxies of similar size and even at comparable redshift, forming groups and clusters. Nonetheless, if the relative distance and velocities did not fulfil the cross-pairs condition with any of the nearby galaxies, then the target galaxy was considered a non-merger.

3.2.2.4 Contamination by Stars

A target galaxy was considered contaminated by a star if it was blending with a star of comparable size, or if a star of comparable size was located close to it. The definition of "close" here is subjective and, in practice, it changed slightly from case to case. Overall, "close" meant that the star was located at a distance from the target galaxy by approximately less than twice the radius of said galaxy.

3.2.2.5 Contamination by Visual Pairs

A target galaxy was considered contaminated by a galaxy if it was blending with a source considered to be a galaxy by the SDSS catalogue; if there was clearly a galaxy of comparable size nearby at different redshift; if there was a small galaxy without spectroscopy close to and there was no distortion either in the target galaxy or in the neighbour; or if in a crowded field none of the galaxies fulfilled the cross-pairs conditions. The definition of close here is analogous to Section 3.2.2.4.

3.2.2.6 Contamination by Artefacts

A target galaxy with an image artifact blending or close to it was considered affected by an artifact. These artefacts were mainly diffraction spikes from bright galaxies, satellite trails, or saturated sources.

3.2.3 Decision Tree

Once the results from the visual inspection were gathered for all galaxies in the subsamples, many showed contaminations – see Chapter 5, Sect. 5.1.2. These contaminants could be found automatically without the visual inspection by performing a decision tree.

3.2.3.1 Catalogue of sources surrounding a target galaxy

The first component of the decision tree was a set of SDSS DR6 detections obtained from CasJobs around each of the target galaxies. CasJobs is the name of the server that hosts the SDSS catalogues[3]. Although newer servers are currently in use for the latests SDSS DRs, it still can be accessed for previous iterations such as DR6. Its data can be access by means of Structured Query Language (SQL), written queries that select catalogues according to some properties and provide only a desired set of columns.

To make sure possible cross-pair galaxies were included in these catalogues, we took all the sources within a comoving radius of 100 kpc/h around each source. This comoving radius was calculated applying the `kpc_proper_per_arcmin` function from the `astropy.cosmology` package as described in Sect. 3.2.1.

[3] https://skyserver.sdss.org/casjobs/

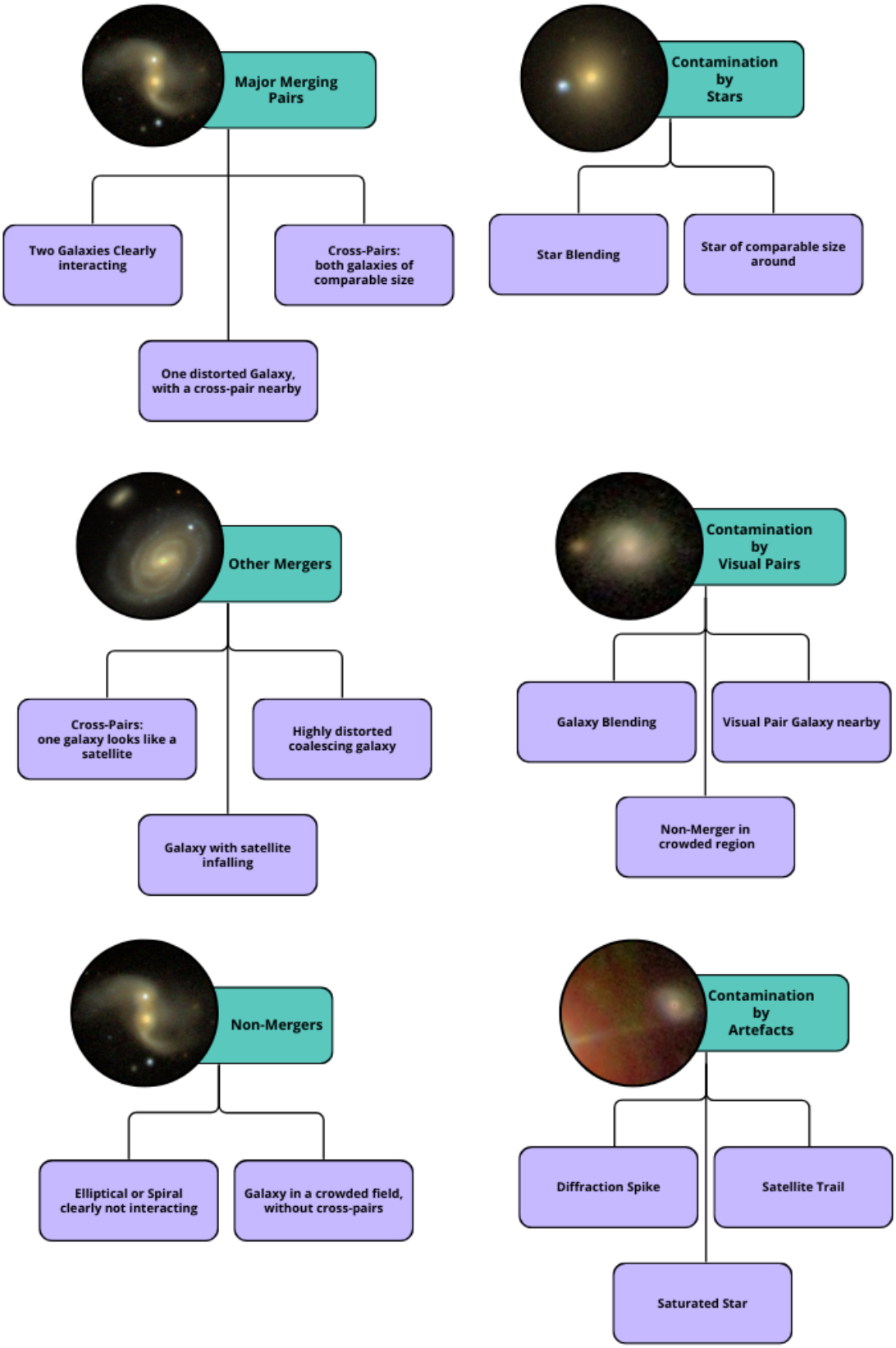


Figure 3.1: Flow chart describing the decision options applied for the visual inspection of galaxies. The left column's three rows describe how galaxies were classified as merging pairs, as other type of mergers, or non-merging galaxies. The right column describes the criteria for defining three different types of contaminations near the galaxies: contamination by stars, contamination by galaxies forming visual pairs, and the contamination by artefacts. This flow chart was created using the online tools provided by the canva webpage (https://www.canva.com/).

3.2.3.2 Flags

All the sources were obtained from a CasJobs query without any flag selection. Nonetheless, the catalogue of surroundings sources included multiple spurious detections. To clean them for the decision tree, some of them were discarded when they presented some specific flags. Moreover, we considered those spurious detections to be affecting too little the surrounding of the target galaxies, and thus to not be interesting for the tree to consider them.

The following list indicates the flags that were considered to imply that a source does not affect the contamination. The flags were chosen by the interpretation of the visual inspection, and a short justification is written for each of them.

- **Flag 2^{46} DEBLEND_NOPEAK** $\rightarrow$ seen in many spurious detections around a bright source
- **2^{47} PSFFLUXINTERP** $\rightarrow$ seen in some other blended spurious sources
- **2^{48} TOOFEWGOOD_DETECTIONS** $\rightarrow$ seen in small sources that blended with galaxies of large angular size

3.2.3.3 Branches of the decision tree

Contamination by Stars: The decision tree considered the target galaxy to be contaminated if there is a star closer than twice the target's Petrosian radius. This also included a star blending within the source itself.

Contamination by Visual Pairs: A galaxy is considered to be a visual pair contaminating the target galaxy if the distance between the centroid of both galaxies is smaller than the sum of the Petrosian radius of also both galaxies. This again included smaller galaxies blending with the source itself.

In order to avoid actual merging pairs to be discarded, the criteria of cross-pairs – Sect. 3.2.1 – was also tested for all galaxies surrounding the target. Therefore, if a cross-pair source is around the galaxy, the contamination criterion was overrun by the cross-pair merger condition.

3.2.3.4 Statistics on the Decision Tree results

Table 3.1 gathers the relevant terms to understand the output of the decision tree. First, it shows the References to the possible contamination results: the dirty and clean visual labels indicate contamination results from the visual inspection, while the clean and dirty from the decision tree (`dt`) are the models outputs. Then, it indicates the Classifications, similar to the formalism in Section 3.1.4: the True Positives (TPs) are visually contaminated sources found to be contaminated in `dt`, while the True Negatives (TNs) are the same but for clean galaxies. False Positives (FPs) are visually clean galaxies that are found dirty in the `dt`, and conversely for the False Negative (FN) results. Finally, it shows the relevant Statistics out of `dt`. Those are the Accuracy, defined as all True results divided by the size of the dataset; the Recall, defined as the rate of dirty sources correctly identified divided by the total number of dirty source in the dataset, i.e, $TPs/(TPs + FNs)$; and the Specificity, defined similarly but for clean sources, i.e., $TNs/(TNs + FPs)$.

3.2.4 Alphashape

The main goal of the visual inspection of the GZ DR1 galaxies was to address the strength of the decision boundary, described in Sect. 4.2.3, for finding mergers by extending the

Table 3.1: Table with the relevant definitions to address the outcome from the decision tree. The References provide the classification types for the decision tree with respect to the visually inspected results. The Classifications describe the output classes, and the Satistics are those that are relevant for quantifying the performance.

Term	Definition
References:	
dirty visually:	any contamination visual flag = True
clean visually	all contamination visual flags = False
dirty from `dt`	a star or galaxy has been found around, and there is no cross-pair
clean from `dt`	neither stars nor galaxies found around, or there is a cross-pair
Classifications:	
TPs	dirty visually and from the `dt`
TNs	clean visually and from the `dt`
FPs	clean visually, but found dirty by the `dt`
FNs	dirty visually, but found clean by the `dt`
Statistics:	
Accuracy:	$(TPs + TNs)/all$
Recall:	$TPs/(TPs + FNs)$
Specificity:	$TNs/(TNs + FPs)$

training dataset to the whole GZ DR1 galaxies. The first step was to constrain the location of galaxy mergers in the diagram. We applied the Alpha shape algorithm (Edelsbrunner et al., 1983) in order to define the region where the bulk of training mergers can be found. Alpha shape is a technique that can be used to define the contour around a group of points in a 2D or 3D graph. Simply speaking, the alpha shape algorithm generates circles with a radius $=1/\alpha$, hence the name. This α is an input of the model, which places circles so that they include points of the distribution on their perimeter, but not inside. When the circles have been defined, then the connection between points are transformed from circular shapes into straight lines. In practice, this builds the outer contour of the system.

For high enough radius $1/\alpha$, the algorithm stops considering outliers as part of the contour. This is because the radius of the circles become small enough that it can fit within points. The circles generated for such points do not reach other points, and as a result, they get disconnected from others. Therefore, only the circles in denser areas of the data contribute the outer edge. The boundary traces points that are closer to each other, the circles do not fit completely between them, and they have to include more than one point on their perimeter. In different words, the higher the α, the more detailed and dense becomes the outer boundary. An example of the effect of high α can be seen on the left side plot in Figure 5.2, and an example of $\alpha = 0$ on the right panel.

3.3 Visual Inspection of the HSC-NEP merger candidates

The main task I carried out for Pearson et al. (2022) was to inspect visually the galaxies classified as mergers by the ML model implemented. My auxiliary supervisor dr. William Pearson and I visually inspected all the galaxies classified by the NN as mergers. We shared the visual inspection of the mergers, being the majority of them observed by dr. William Pearson. The catalogue from where the mergers were identified is described in Chapter 2, Sect. 2.2.2. The goal of this work was to confirm how many of the ML-based merger classifications were correct. This visual inspection was performed similarly to 3.2.2, but

instead of loading the SDSS DR6 images from the internal collection of Aladin, we loaded the HSC-NEP r-band images.

During the work, it was necessary to adapt how Aladin presents the image in order to discern better the presence or absence of tidal features, of shape asymmetries, or of any other type of distortions in the images. The astronomical images shown by Aladin can have their aspect adjusted through two modifications: the scale of the colour bar and the range of pixel values considered. The main scales used were linear, square root, logarithmic, or following and *asinh* function. All of them are shown as a scale of grey shades from white to black. Generally, the square root and *asinh* were more adequate to show some of the features than the others.

The pixel range was set by defining a minimum and a maximum pixel values. This range fixes the limits between which the colour scale changes, making the pixels with the highest or the lowest values to appear as white or black, respectively. Modifying the limits of observation had two main effects that, combined with the different colour scales, could enhance the low-surface brightness features. Setting an upper limit below the majority of bright pixels would make them all white, reducing the dynamical range to low surface-brightness pixels. The lower limit was sometimes forced to be lower than the value obtained from the software autoset. This autoset would take as minimum and maximum the interval of pixel values among those pixels shown in the interface. Making it smaller than the minium pixel was sometimes useful to reduce the variability of the pixels in the background level. A more homogenous background allowed discerning better the low-surface brightness merging features

It is worth noting that this was not done for the inspection of the SDSS galaxies in Sect. 3.2.2. This was because those galaxies were observed through the colour images in the Aladin collection, and they had the pixel range and scale fixed as a red-green-blue combination of the r, g, and i bands.

3.4 HSC sky error extension

The galaxies in the catalogue Galaxy Zoo: Cosmic Dawn (GZ:CD) were matched with those in the HSC-NEP deep field, as described in Sect. 2.2.3. We then obtained squared cutouts for each galaxy using the g-band images, where we made the analyses of the Low-Surface Brightness (LSB) pixels. These cutouts had as centre the centroid from the GZ:CD data, and their sides had a length equal to 20 times their effective radius (r_{eff}). This side length was taken to be much larger than r_{eff}, in order to make sure that there is enough area for the apertures that later will be applied. The cutouts included both the calibrated image and the variance image resulting from the data reduction pipeline. Finally, a segmentation map was created on the cutouts using Source Extractor – See Sect. 1.4.3 –, in order to make sure that the target galaxy was detected and correctly found in the cutout centre.

Given the cutouts around each galaxy in the catalogue, we defined an aperture around them. The apertures' centres were then defined at the same position as the cutout centroids, and with a radius equal to some factor of r_{eff}. This factor is kept as a free parameter, because different aperture radii will include more or less sky background pixels. Thus, the LSB analysis highly depends on it.

The background level was then calculated by a clipped median – See Sect. 1.4.2. We did this by implementing an algorithm that initially takes all the pixels within the aperture and calculates its median and standard deviation. The pixels within the edge of the aperture were included as full pixels, even if only a fraction of them were inside it by the aperture's geometry. Later, the clipped median would proceed by discarding all pixels with higher values than the median plus a factor of the standard deviation, i.e., pixels above *median*

$+ std \cdot f$, where f is the clipped median factor. This clipping factor f was also left as a parameter of the method. The clipped median step is repeated around ten times, providing a final background model that we use as the histogram of the LSB pixel distribution.

The resulting clipped background histogram was saved and a set of multiple parameters characterizing the histogram shape was calculated.

3.4.1 Parameters calculated on the LSB histograms

The parameters applied to the LSB histogram consisted of basic statistics extracting information on the shape of the histograms. The first two are the **mean** and **median**, providing two different measurements of the background level. Then, the interquartile range (**IQR**) was calculated, together with three similar parameters dubbed as **Fraction 1**, **2**, or **3**. These were defined by the following equations, where the q_x variables are the x quantiles:

$$\text{IQR} = q_{75} - q_{25} \ , \tag{3.5}$$

$$\text{Fraction}\,1 = \frac{q_{75} - q_{50}}{q_{50} - q_{25}} \ , \tag{3.6}$$

$$\text{Fraction}\,2 = \frac{q_{100} - q_{75}}{q_{25} - q_{0}} \ , \tag{3.7}$$

$$\text{Fraction}\,3 = \frac{q_{100} - q_{50}}{q_{50} - q_{0}} \ . \tag{3.8}$$

The last two parameters applied were the **Skewness** and the **Kurtosis**. Both of them measure the asymmetry of the distribution, the skewness through quantifying the third order moment and the kurtosis through the fourth order moment.

The resulting combination of parameters provided a set of data that can be simplified through dimensionality reduction methods. We made used of the Neighbourhood Components Analysis (NCA; Goldberger et al., 2004) to enhance the difference between mergers and non-mergers in the resulting eight-dimensional space.

Neighbourhood Components Analysis: it is a non-parametric model that is able to generate an embedding for classification data, while clustering it through the known labels. It does it by optimizing a linear function between the data and the labels. This linear function is a $d \times D$ matrix, where D is the dimensions of the data and d is the dimension of the embedding. For the results presented in this project D=8, and d=2 provides a 2D space.

3.4.2 HSC Photometric Pipeline `hscPipe` version 6.7

The structure of the HSC Pipeline used to reduce the g-band HSC-NEP deep observations can be followed in Fig. 3.3 from left to right. The raw exposures for each of the individual observations are split into the internal CCD's of the camera. Out of the 116 CCD's, 103 are used explicitly for the observations, although one of them was not functioning. The pipeline builds an infrastructure of directories to apply the bias, dark, and flat frames consistently along each CCD and exposure. It is at this point that the pipeline takes advantage of the dithering between exposures to generate one of the two focal plane background that are explained below. This generates the `Sky*.fits` frames per CCD, which store the dithering-based background image, variance, and mask frames. This `Sky*.fits` frame is not used until the stage previous to the coaddition, described below, but it is generated here because it is at this point when the frames have had the bias, dark, and flat frames applied, and this is required previous to the background modeling.

At this point of the pipeline, after the main calibration frames per CCD have been applied, the source detection is performed. First, in each CCD the Instrument Signature Removal (ISR) is applied, which treats the cosmic rays, bad pixels, or saturated pixels by masking them and interpolating new values from the surrounding pixels. When this is done, the pipeline performs three subsequent source detections. Each of them begins with the estimation and subtraction of the sky background, followed by a segmentation and source detection, and proceed with a PSF estimation. As a result, three background subtractions are done, and the final catalogue and the PSF function are generated so that they can be used for the photometric and astrometric calibration.

The three background subtraction images, together with the resulting variance and mask frames, are stored in `.fits` files for each CCD and exposure. The `.fits` format can store multiple images as Header Data Units (HDUs) that are part of one file. Each HDU includes not only the pixel values themselves, but also a header with relevant information. The images of an example background file `skyCorr**.fits` is shown in 3.4. It includes the image, variance, and mask frames of the three detection backgrounds and of the two background models performed later on the focal plane.

The sky background calculated during the source detection is measured in each CCD exposure by splitting it into boxes of 256×256 pixels, with their centroids separated by 128 pixels across a 2D orthogonal grid. This effectively bins the original 2144×4241 pixels of one CCD into a 17×33 pixel image. For each box, the background is calculated by a clipped median with sigma factor of 3. Finally, the background image is fit into a polynomial, smoothing the background level through the CCD field of view. This results on three HDUs stored by the pipeline: one for the image itself, another for the variance, and a last one for masked pixels. One example of these HDUs can be seen in Fig. 3.4, where the image and the variance frame are the second and first frames from the right in the first row, respectively, and the mask frame is the first one from the left in the second row. The mask frame indicates in fact the pixels not observed by the CCD: the bottom-left arc is the camera edge, and the vertical line corresponds to a dead channel, one of the four vertical channels that form each CCD.

The astrometric and photometric calibration of each CCD is done using the PAN-STARRS1 3pi survey for reference (Schlafly et al., 2012; Tonry et al., 2012; Chambers et al., 2016; Flewelling et al., 2020; Magnier et al., 2020a; Magnier et al., 2020b,c; Waters et al., 2020). The astrometric calibration is used to "warp" all the exposures, which is the process of creating the pixel grid that the coadd algorithm considers to stack all the exposures. Such grid also provides the base reference to combine all the CCD of a single exposure into the focal plane of the camera. Taking advantage of this focal plane image, the pipeline creates a global background for the whole exposure, generated through 1024×1024 boxes. The model's image, variance, and mask frames of an example CCD are shown in the first HDU from the right in the third row and the first two from the left in the fourth row of Fig. 3.4. It is sensitive to large-area background contributions, such as light pollution from the moon, artificial light sources, or atmospheric glow.

Before coadding all the images – see Sect. 1.4.6 –, the three background images created during detection are added back to each CCD. Then, the global background in the focal plane is subtracted. Finally, the sky background generated in the first calibration stage making use of the dithering between frames is also subtracted for each CCD. This background's image, variance, and mask frames are shown for an example CDD as the two HDUs on the right of the fourth column of Fig. 3.4, and on the final row. The dithering image gets rid of static background signatures, those that do not depend on the pointing position of the telescope. Those are known to be strongly affected by inhomogeneities in the transmission of the camera filters, and have been well characterized by the HSC team as shown Fig 3.2.

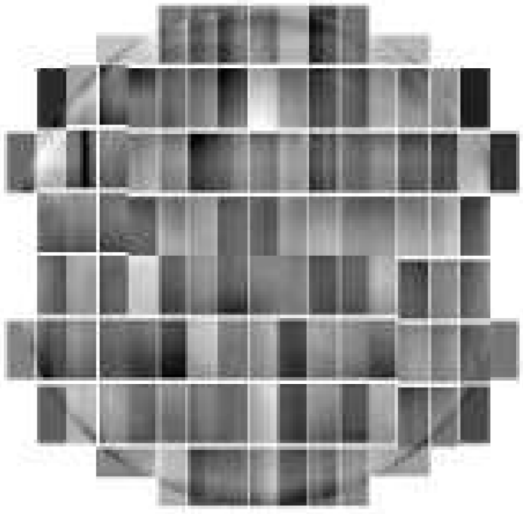


Figure 3.2: It depicts the g-band global sky subtraction, obtained from combining in the focal plane all the `Sky*.fits` files of an example exposure. They provide the static background image generated from the dithering of the images, which is mainly affected by the gain of individual CCDs. The features change from CCD to CCD. Taken from Fig. 4 in (Aihara et al., 2019).

A summary of the sky subtraction can also be read in the `hscPipe` 7 webpage[4].

To complete the data reduction, the Coadd algorithm is performed, and the Calibrated Coadd frames are obtained, as indicated in the two last boxes in Fig. 3.3.

3.4.2.1 Sky background modification

Observing low-surface brightness (LSB) in astronomical images is highly dependent on the sky background subtraction strategy. A background modelling can confuse spurious observations with actual LSB features. We wanted to make sure that this does not happen with the HSC-NEP images, and this was the motivation of generating our own data reduction.

Therefore, we wanted to make sure none of the background subtractions are spoiling the LSB pixels. Out of the five background models, we consider that only the last background, created from dithering, might have a detrimental effect on the low Signal-to-Noise (S/N) surrounding of the galaxies. The three background images created during the detection stage were added back to the image in the coaddition stage. If we assume that this process of subtracting and adding back the fitted background boxes is reversible, meaning that no background feature is lost due to it, then it should not have an impact in the LSB distribution. In the case of the fourth background, the global background calculated in the focal plane using 1024×1024 boxes, the large size of the box implies that it only traces wide background structures and not the smaller LSB structures that we are searching for.

Thus, the only background that we considered to potentially have a negative effect in the low S/N pixels is the last one, which was calculated from the dithering. While this background should only be affected by static noise, it is still sensitive to small scale structures. Consequently, we performed two calibrations for the whole g-band observation in the NEP: the first calibration was run without any modifications to the background, and the second was run avoiding the dithering subtraction. The first calibration is called

[4] https://hsc.mtk.nao.ac.jp/pipedoc/pipedoc_7_e/tips_e/qa_globalsky.html#qa-globalsky

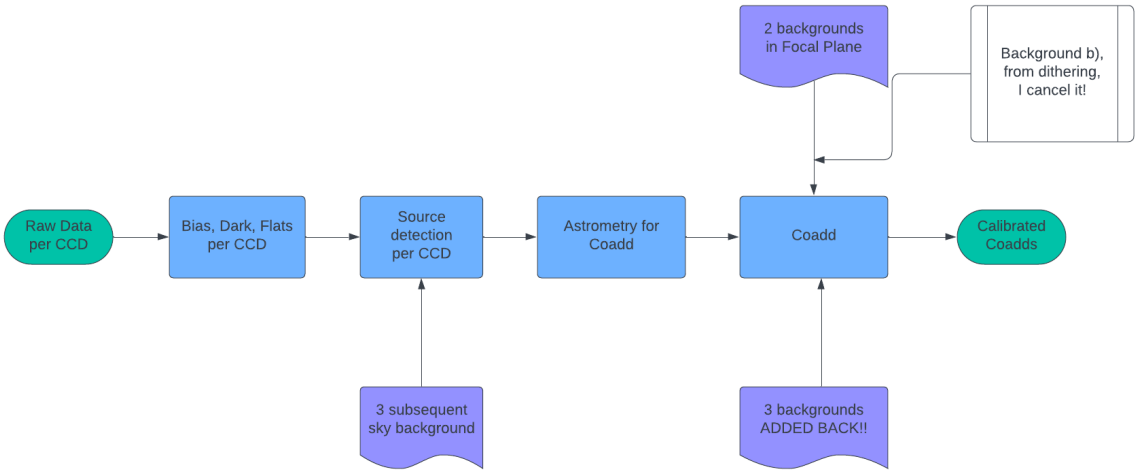


Figure 3.3: Flow chart of the data reduction pipeline HSC Photometric Pipeline `hscPipe` version 6.7 used in *g*-band images of the AKARI HSC-NEP deep field. The process begins on the left with the raw camera exposures, the Raw Data per CCD box. Then, the main data reduction steps are carried subsequently towards the final calibrated coadd exposures, with two instances of sky background subtraction along the way. This flow chart was created using the smartdraw portal (`https://www.smartdraw.com/flowchart/`).

through the text as "Calibration with sky" or as "Normal sky". The second calibration as indicated as "Calibration without sky" or as "My sky".

We managed to avoid the dithering subtraction by setting to 0 the values of the pixels in the three image, variance, and mask frames. The pipeline is run in the exact same way as in the normal case, except that before the `skyCorr*.fits` files are used in the Coadd stage, we run a script that sets its last three HDU to zero. The resulting `skyCorr*.fits` is depicted in Fig.3.5.

Understanding the background subtraction, the steps in the pipeline source code, and how to cancel the subtraction of the dithered sky, were only possible with the help of the HSC Software HelpDesk, specifically of dr. Hiroki Onozato. I would like to dedicate these lines to thank his constant effort and support in responding to my mails.

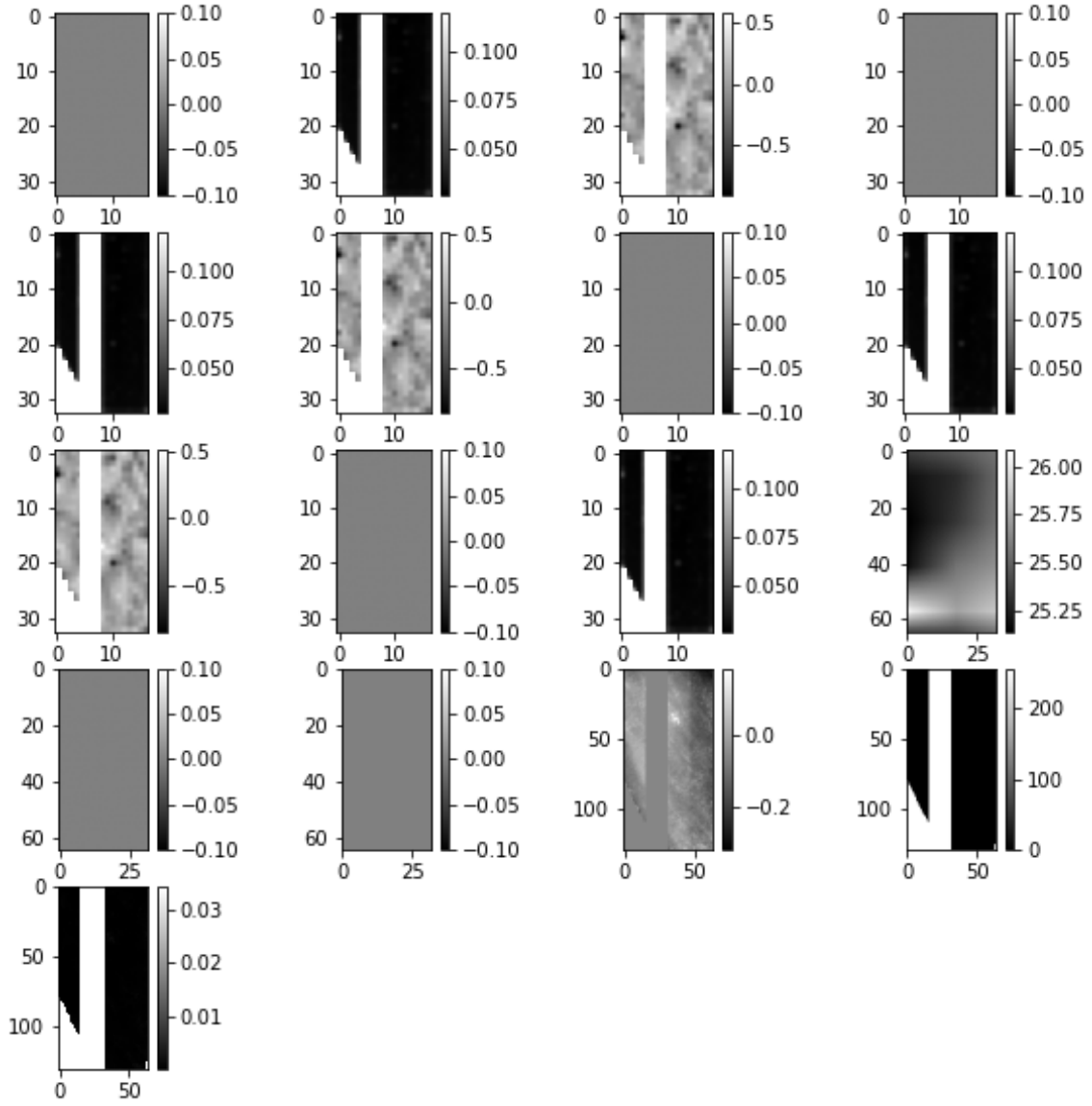


Figure 3.4: All stored HDUs within the files `skyCorr**.fits`, where the `**` in the file's name correspond to the numbers indicating the CCD and exposure for each given file. This file includes all the background information that is used in the coaddition. The three background subtractions done through the source detection stage are stored in groups of three HDUs, corresponding to the background model image, the variance, and the masks, in this order. The three HDUs of the first subtraction are the two panels on the right side of the first row, and the first one in the left of the second row. The other three panels in the second row correspond to the second background, and the panels of the third subtraction are the three starting from the left in the third row. The image themselves have negative sign with respect to the original background, because they are added back previous to including the other backgrounds. The first HDU from the right in the third row and the first two from the left in the fourth row correspond to the global focal plane background. Finally, the last three images, the two on the right of the fourth row and the only image in the fifth, correspond to the background obtained from dithering.

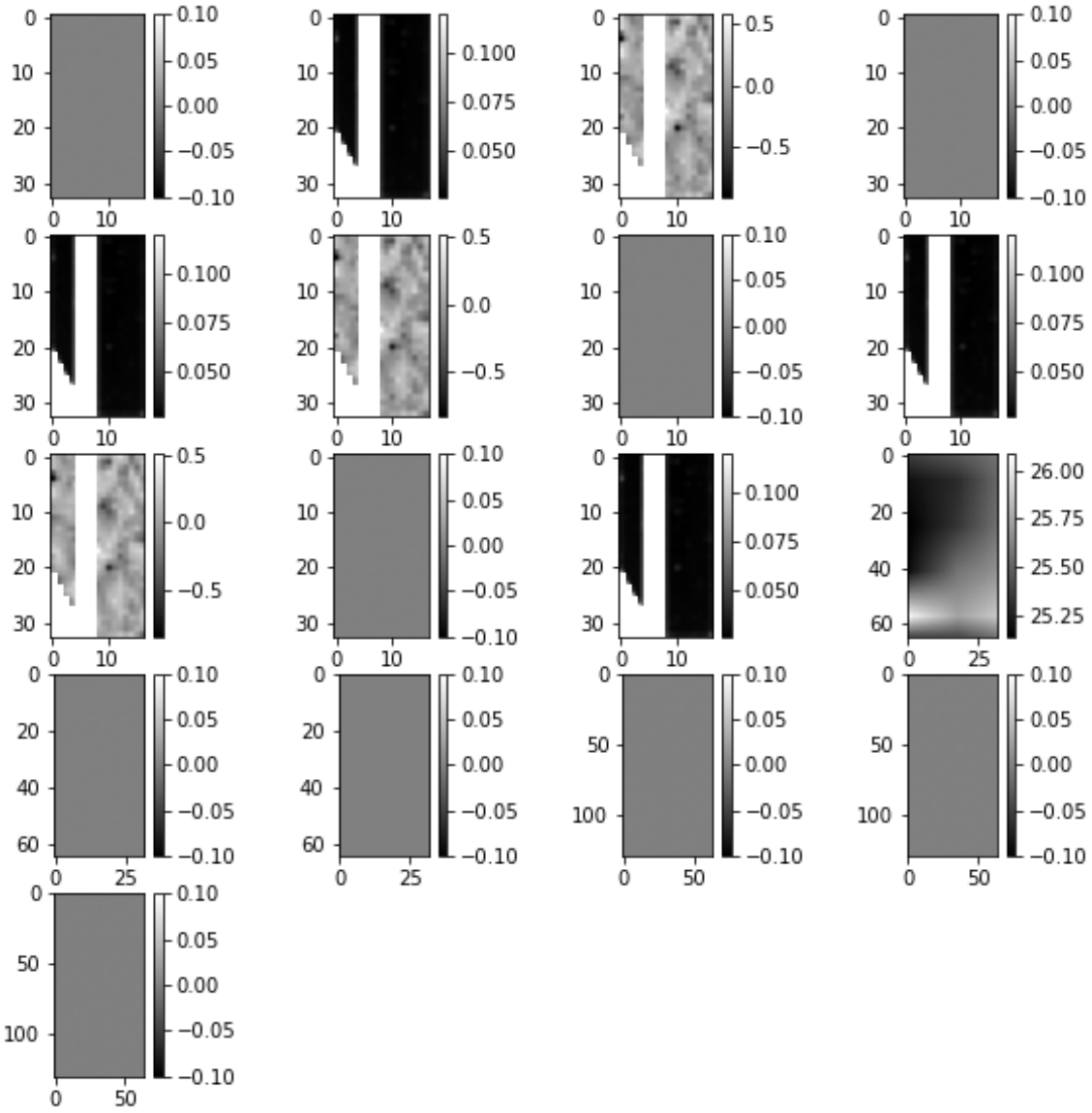


Figure 3.5: The 17 HDUs plotted are the same as in Fig. 3.4, except that the last three have been set to 0 to cancel the last background subtraction, as explained in Sect. 3.4.2.1.

CHAPTER 4

Sky Error

This chapter presents the outcome of the suite of Neural Networks (NNs) that we trained using the dataset described in Sect. 2.1.3. Every NN provided from the photometric input parameters – see Sect. 2.1.1 – the classification between mergers and non-mergers. Using multiple data normalizations and applying data reduction methods to the inputs – see Sect. 3.1 –, we found the potential of the sky background error `skyErr` to classify mergers through a decision diagram. The text in this chapter is based on Suelves et al. (2023), all the work was carried out by myself, I built the catalogue, wrote the codes, prepared all the plots presented in the chapter, etc. The analysis and discussion was developed together with the co-authors, my auxiliary supervisor William J. Pearson, and my supervisor prof. Agnieszka Pollo. During this work, we found that the description of photometric parameters in the SDSS DR6 documentation was not accurate. We made corrections to it, which are presented and justified in the Appendix A.

4.1 Results

4.1.1 Architecture selection

In order to determine our definitive layer layout for the photometry-based NN, we compared the multiple architectures listed in Table 4.1 by their five-fold validation loss. The nomenclature prefix indicates the layer number as nL and the suffix refers to the size when required. Figure 4.1 shows the mean loss and its standard error of the NNs sampled on the BC model magnitude input space. A relatively similar loss was obtained for all versions, except for the cases with too few neurons, 2L_4 and 2L_2. The longest NNs we tried combined up to five layers, but both 5L_b and 5L_s did not improve the performance. The 4L and 3L cases gave loss values as low as those of 2L_64 and 2L_32. Regarding instability, the architecture with the lowest variance was 2L_16. For the definitive NN, we opted for 2L_16 due to its convenient balance: a slightly worse but more stable loss when compared to 4L, 3L, 2L_62, and 2L_32, combined with a shorter computational time. We note that the five-fold galaxy distribution was not fixed but selected randomly for each architecture check.

The other tested NN parameters were the initial learning rate of the Adam optimizer and the dropout rate. The former performed quite badly when significantly diverted from 5×10^{-5}. Regarding the latter, the validation losses of the considered variations are shown in Fig. 4.2. Following the argument for selecting 2L_16, a 0.1 rate would be a more adequate

Table 4.1: Architecture options considered for the NN layout, whose performances are depicted in Fig. 4.1. The Layout column gives the number of neurons per layer, separating each layer with a +.

NN name	Layout
5L_b	128+256+512+256+128
5L_s	64+256+512+256+64
4L	128+256+256+128
3L	32+128+32
2L_64	64+64
2L_32	32+32
2L_16	16+16
2L_8	8+8
2L_4	4+4
2L_2	2+2

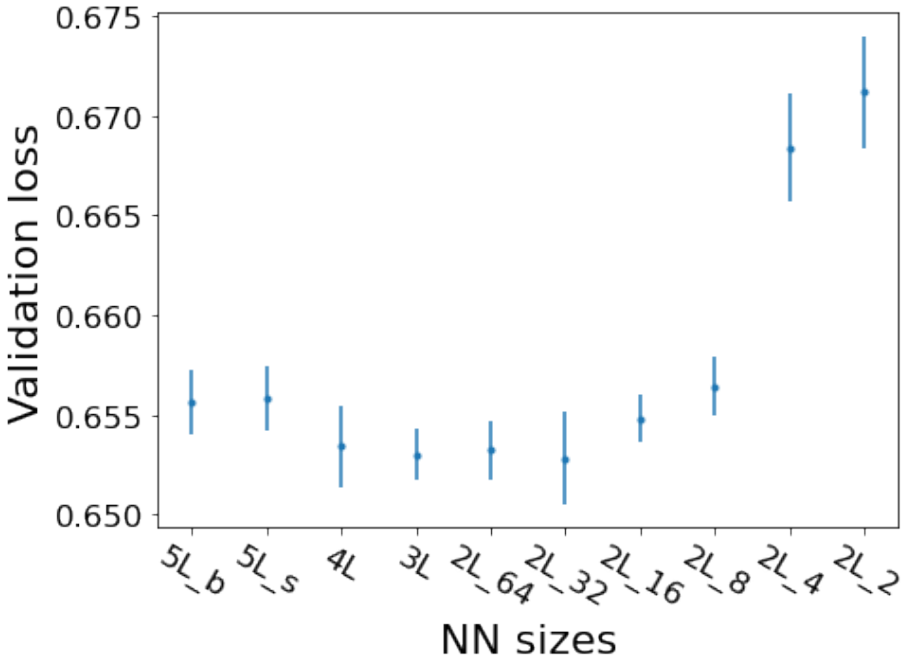


Figure 4.1: Validation loss for each tested NN architecture defined in Table 4.1. It is calculated as the mean and standard error of the loss at the best validation update obtained in each of the five-fold validation cycles.

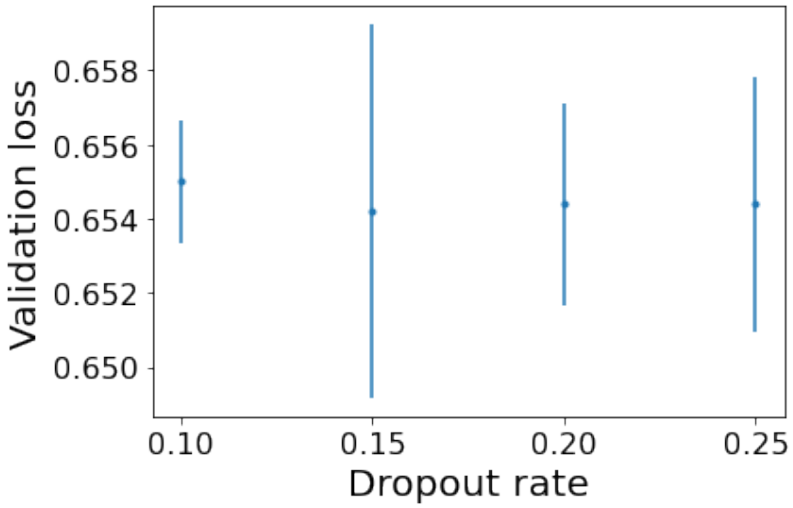


Figure 4.2: Validation loss for each dropout rate chosen. The value is again the mean of the five-fold validation cases and the error bars come from the standard error among the five folds.

Table 4.2: NN training parameters, as defined in Sect. 3.1.4, for the Reference NN input: the bands plus colours model magnitude case. They imply the NN has the potential of figuring out the classification rules.

Accuracy	68.90 $\pm$ 0.72 %
TPs	61.65 $\pm$ 0.77 %
TNs	75.49 $\pm$ 0.92 %

option as it shows a larger loss and smaller standard error. However, we decided to keep the 0.2 rate because the error bar for 0.1 covers only the upper – and worse – part of the interval that 0.2 covers.

4.1.2 Input magnitude variations

We established the BC model magnitude as the initial reference NN for comparisons because it more closely represents the real galaxies' brightness. This input is a 15D space combining the five photometric bands plus the resulting ten colours, normalized in the range [0,1]. It led to an accuracy of 68.90 $\pm$ 0.72%, as shown in Table 4.2. This accuracy implies that we found a stable classifier capable of correctly identifying a substantial amount of objects. Moreover, it is encouraging how around 60% of the mergers were correctly classified, given such a simple input space.

Figure 4.3 shows the resulting accuracy of the NN applied over all six types of magnitudes – as defined in Sect. 2.1.1 – and the input variations – as defined in Sect. 3.1.3. The horizontal orange line is the reference NN accuracy, and the shaded area is its error. Each panel in Fig. 4.3 confirms how the BC magnitude inputs lead to a significant increase in the accuracy with respect to the separated B and C cases. Both bands and colours generate consistent accuracies in all magnitude classes because they essentially contain the same information: one colour index is nothing more than a linear combination of the magnitudes measured in two different bands, in other words, the ratio of fluxes. Including the two of them at the same time seems to facilitate the NNs' performance, and therefore seems to be a way to improve the model. This pattern is found independently of the magnitude type.

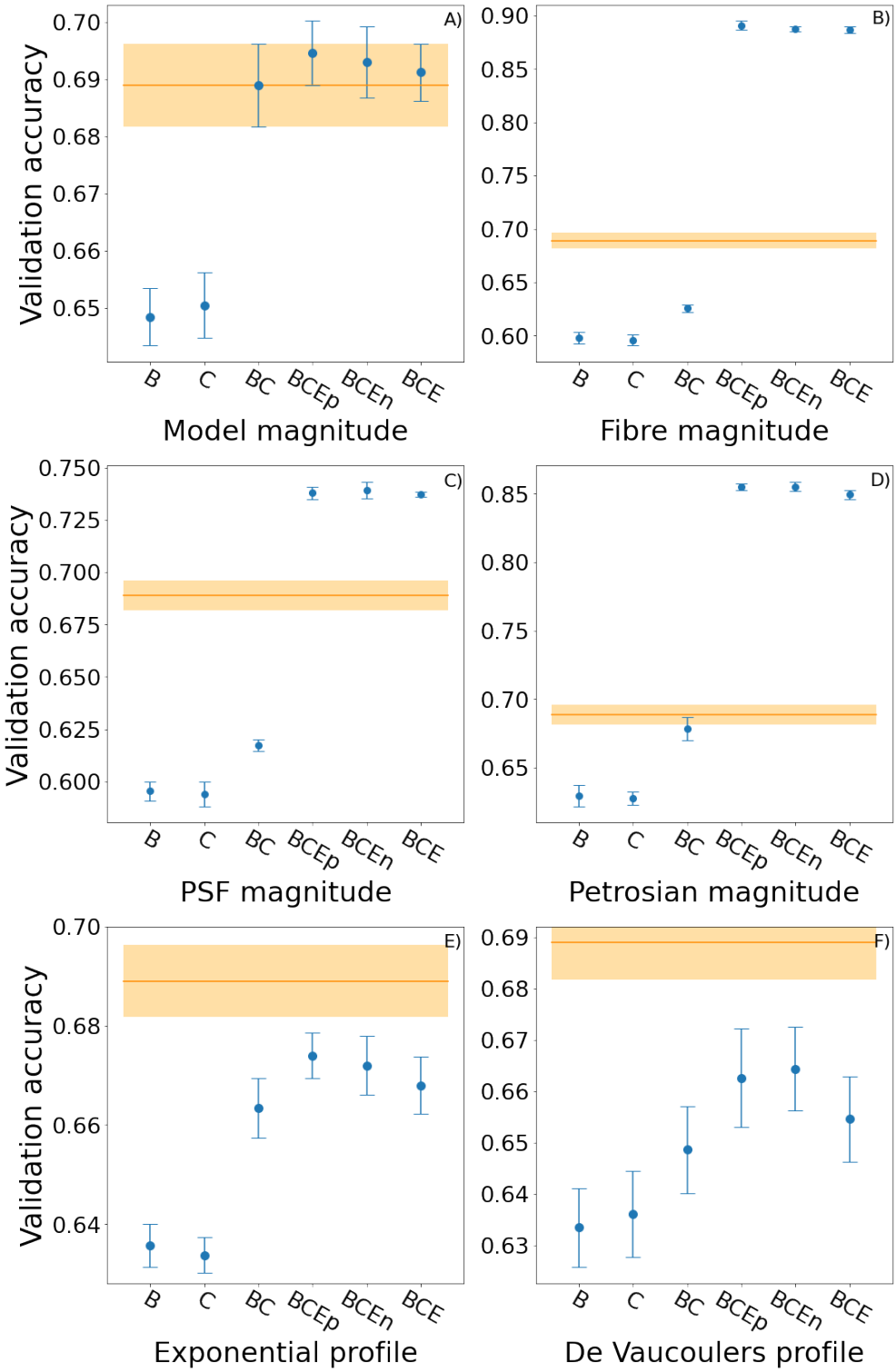


Figure 4.3: Six-panels plot showing the mean validation peak accuracy of the different input variations for each magnitude type. For each magnitude defined in Sect. 2.1.1, we provide several variations: B, which corresponds to the five band values; C, the ten colours obtained from the five bands; BC, a 15-dimensional space combining bands and colours; and BCE, a 20-dimensional space that adds the magnitude errors to the BC cases. All four of these variations follow the min-max normalization defined in Eq. 3.1. Additionally, we show the BCEp and BCEn sets, for which bands and colours were min-max normalized separately to the errors, obtained with Eqs. 3.2 and 3.3 respectively. The distribution of galaxies among the five validation folds is fixed to be the same. Panel A) corresponds to the model magnitude type, B) to the fibre magnitude, C) to the PSF magnitude, D) to the Petrosian magnitude, E) to the exponential magnitude, and F) to the De Vaucouleurs one.

Introducing the errors in the feature space has an influence that highly depends on the magnitude type. As described in Sect. 3.1.3.1, the errors were applied in three ways: including them in the `min-max` normalization with the magnitudes and colours (BCE), calculating them from the fractional error directly (BCEp) – see Eq. 3.2 –, or through the combining the `min-max` normalization with the fractional error (BCEn) – see Eq. 3.3. For the model magnitude, the exponential, and the De Vaucouleurs profiles, the inclusion of errors produces an only slight increase in the accuracy. However, for the PSF, Petrosian, and fibre magnitudes, the accuracy increases highly, reaching almost 90% for the BCEp fibre case.

Regarding the error normalization options, both the BCEp (Eq. 3.2) and BCEn (Eq. 3.3) cases have slightly better accuracy than BCE cases, while being consistent within the standard error. The BCEn cases are generally slightly better than the BCEn ones. This is one indication that suggests that the NN is quite dependent on the normalization method.

4.1.3 Fibre errors mixed with other data

The surprising success from the fibre BCE case deserved a deeper look: not only did it reach a high accuracy, but it also outperformed the reference input – the BC model magnitude. We narrowed down the search of what the main information source was within fibre BCE by combining the bands, colours, and errors of both model and fibre magnitudes in different ways.

Table 4.3 shows the validation mean accuracy for all the relevant combinations, separated into five blocks. The first two blocks indicate that the model B, C, or BC cases combined with model E retain an accuracy of around 67 %, but if the model errors are swapped with fibre errors, then the NN achieves high values, similar to those of the BCE fibre. The third and fourth blocks confirm the same results, but applying fibre B, C, and BC cases. The last block shows the results for model E or fibre E alone. The importance of fibre E is therefore demonstrated not only by how these errors enhance the accuracy when they accompany any magnitude type but also by their performance when used as a 5D input.

4.1.4 Fibre errors' components

According to the NN results presented up to now, one can just use the fibre magnitude error and get a correct classification with an accuracy of $\sim$84%. Such an achievement does not seem intuitively justified solely by the properties of the input data. The next step was then to understand how fibre E values were calculated.

According to the SDSS documentation, the fibre magnitudes are simple aperture magnitudes, meaning we can replicate them with the original astronomical frames and the appropriate calibration data. The whole process of reproducing the fibre magnitudes and errors is detailed in Appendix A. We obtained the errors from the fibre aperture counts in the fpObjc file downloaded from the SDSS repository, sampled in each band on a group of ten galaxies. We justify the updated version of the formula that calculates the error from the initial counts (Eq. A.5).

According to Eq. A.5, the aperture error calculation has four particular inputs per band, which are: the digital unit count in the aperture region `counts`, the error in the CCD camera's dark current `dark variance`, the sky background estimation in the source's centre `sky`, and its error `skyErr`. We retrieved the four inputs per band for each galaxy in our kfold validation sample. We trained several NNs, noted in Table 4.4, considering first all sets of variables and then excluding each one at a time. It arose that the main source of accuracy was `skyErr`, as expressed in the accuracy from fibre E inputs without `skyErr`. Furthermore, running an NN only with `skyErr` showed a very similar accuracy to that of fibre E alone, implying they contain similar information.

Table 4.3: Validation mean accuracy and standard error for all the relevant combinations, separated into five blocks. Each input space undergoes a min-max normalization. The table's first two blocks combine the model B, C, and BC cases with either model E or fibre E, respectively; the third and fourth blocks do the same but for the fibre B, C, or BC cases; and the last block shows what happens for model E or fibre E alone. As in Fig. 4.3, the source distribution among the five validation folds is fixed to be the same.

Input space	Accuracy
model B + model E	67.14 ± 0.31 %
model C + model E	67.24 ± 0.64 %
model BC + model E	69.12 ± 0.50 %
model B + fibre E	77.08 ± 0.73 %
model C + fibre E	88.30 ± 0.45 %
model BC + fieer E	87.40 ± 0.72 %
fibre B + model E	62.56 ± 0.51 %
fibre C + model E	62.40 ± 0.45 %
fibre BC + model E	62.80 ± 0.55 %
fibre B + fibre E	77.14 ± 0.60 %
fibre C + fibre E	88.66 ± 0.37 %
fibre BC + fibre E	87.84 ± 0.42 %
model E	59.06 ± 0.76 %
fibre E	83.76 ± 0.32 %

Moreover, every training in which `skyErr` was accompanied with other features achieved an accuracy better than 90%. The best NN is for `skyErr` and `dark variance` together. Nonetheless, the dark current error does not seem to be related to the galactic properties at all, as it is a property of the CCD camera – see Sect. 1.4.1. This led to the last result, checking the dependence of the NN accuracies on the input normalization.

4.1.5 Normalization dependence

The high accuracy our NN obtained from the `skyErr` and fibre E inputs varied with the companion features. We wanted to see if this behaviour was because of the normalization applied. The first test was to expand the min-max normalization interval from [0,1] to [0,2] for some selected cases. In Table 4.5 the accuracies of the four most relevant inputs are compared between both versions. Only for the fibre E all-input set is there a relatively significant difference, leading us to conclude that the min-max resulting normalized interval is not a critical choice.

The next step was to apply the normalization method introduced in Sect. 3.1.3.2 to different input spaces that included either `skyErr` or fibre E. Table 4.6 shows the accuracy of the NNs we built using the input `skyErr` or fibre E subsets, compared to the complete spaces from which we isolated them. When `SkyErr` is normalized with `dark variance`, or with `counts` and `dark variance`, or with all other error inputs – `counts`, `sky`, and `dark variance` – the accuracy shows little dependence on the companions' presence or absence. For the `skyErr` without normalization, the accuracy is 90.88 %, which is also high. Fibre E is shown to be related to the companion for the fibre BCE and fibre CE cases, but not for the BCEp (fractional errors) or the BCEn case. However, the errors in the BCEp and BCEn cases are explicitly obtained from the bands, so it could be argued that the NN does get that information from the resulting error inputs. The pre-normalized fibre E shows better values than its min-max, but it deviates from the other cases more than `skyErr` does.

Table 4.4: Validation peak mean and error for the NN input spaces using different variables that combine to make up fibre E. The first row considers all inputs shown in Eq. A.5, followed by the results of excluding one variable at a time. The next row shows the case of only `skyErr`, and the final three rows come from combining `skyErr` with every other variable.

Input space	Accuracy
fibre E – all-input-set	91.48 ± 0.31 %
fibre E – without `counts`	91.62 ± 0.25 %
fibre E – without `dark variance`	91.48 ± 0.23 %
fibre E – without `sky`	91.98 ± 0.27 %
fibre E – without `skyErr`	60.62 ± 0.85 %
`skyErr` – only	83.30 ± 0.38 %
`skyErr` + `counts`	91.66 ± 0.34 %
`skyErr` + `dark variance`	92.04 ± 0.26 %
`skyErr` + `sky`	90.90 ± 0.21 %

Table 4.5: Main project NNs but with the min-max normalization set to an [0,2] interval. Little difference can be found from the previous case.

Input Space	Accuracy min-max [0,2]	Accuracy min-max [0,1]
Reference NN	68.82 ± 0.57 %	68.90 ± 0.72 %
Fibre BCE	88.64 ± 0.36 %	88.68 ± 0.31 %
Fibre E – all-input-set	90.82 ± 0.14 %	91.48 ± 0.31 %
`skyErr` – only	83.52 ± 0.38 %	83.30 ± 0.38 %

Table 4.6: Neural network performance for input spaces in which either `skyErr` or fibre E has been normalized with some other parameters that were not included as inputs (see Sect. 3.1.3.2). The last column gives the accuracy for the input space prior to the applied modification. For the `skyErr` and fibre E pre-normalized spaces (rows 4 and 9), the last column compares their respective 5D isolated min-max value.

Input space	Accuracy	Previous accuracy
`skyErr` as if with `dark variance`	92.10 ± 0.28 %	92.04 ± 0.26 %
`skyErr` as if with `counts` & `dark variance`	92.02 ± 0.21 %	91.98 ± 0.27 %
`skyErr` as if in fibre E – all-input-set	91.02 ± 0.32 %	91.48 ± 0.31 %
`skyErr` pre-normalized	90.88 ± 0.25 %	83.30 ± 0.38 %
fibre E as if in BCE	79.72 ± 0.32%	88.68 ± 0.31 %
fibre fractional errors	88.60 ± 0.55 %	89.06 ± 0.42 %
fibre E as if in BCEn	88.40 ± 0.33 %	88.76 ± 0.25 %
fibre E as if in CE	77.66 ± 0.66 %	88.66 ± 0.37 %
fibre E pre-normalized	86.74 ± 0.40 %	83.76 ± 0.32 %
`skyErr` pre-normalized logarithmic scale	92.64 ± 0.15 %	...

It can be interpreted that fibre E does benefit from being together with the magnitudes in specific combinations, but that `skyErr` is sufficient by itself. Nonetheless, the other conclusion is that `skyErr` depends on the normalization to provide the high accuracy observed. Because the best NN result comes from the `skyErr` as if with `dark variance` case, a parameter unrelated to the galaxies, and the resulting accuracy is close to that of the pre-normalized `skyErr` input, we can consider that the role of the `dark variance` is simply to adapt the `skyErr` numerical representation for the NN to better identify the properties of the mergers. Moreover, the last row with the sky error in logarithmic values achieves an even better result, supporting the finding that the `skyErr` is a good merger proxy on its own. The NN for this best case with the saved weights can be found on `GitHub`[1].

4.2 Discussion

The accuracies obtained from the most relevant input spaces are presented in Table 4.7, together with their performance on the test set applying the saved weights. These results demonstrate that the NN has successfully classified galaxy mergers by making use of photometry, and that we have found the sky background error to be the source of the best

[1] https://github.com/LuisEduSuelves/NN16_skyErr-log

Table 4.7: Accuracies for the central NN input spaces of the project. The table also gives the mean accuracy over the test set, calculated using the weights at the peak validation in each fold.

Input space	Accuracy
Reference NN	68.90 ± 0.72 %
Test Reference NN	69.72 ± 0.36 %
Fibre BCE	88.68 ± 0.31 %
Test Fibre BCE	89.60 ± 0.24 %
Fibre E – inputs	91.48 ± 0.31 %
Test Fibre E –inputs	91.20 ± 0.35 %
`skyErr`	83.30 ± 0.38 %
Test `skyErr`	79.56 ± 0.10 %
`skyErr` min-max as if with `dark variance`	92.10 ± 0.28 %
Test `skyErr` min-max as if with `dark variance`	90.92 ± 0.20 %
`skyErr` pre-normalized in logarithmic scale	92.64 ± 0.15 %.
Test `skyErr` pre-normalized logarithmic scale	92.36 ± 0.21 %.

method. Such a calibration parameter potentially points to the importance of differential image analysis, which, to our knowledge, had not been considered as a key method for galaxy merger identification until now. Therefore, this discussion will attempt to justify the advantages of our method. In Sect. 4.2.1, we address its reproducibility and its potential use both in SDSS and in other surveys. Then, in the rest of the section, we study the distribution of galaxies in the five-band `skyErr` space using the min-max normalization as if with `dark variance`, which is the next-to-best accuracy found. For that, we show dimensionality reduction and feature space distributions. To summarize, we infer why the logarithmic sky error should work even better as input space, showing that a simple 2D boundary can be as effective as the NN. Finally, we justify why the sky background error contains information of merging processes.

4.2.1 Reproducibility of the model

The step-by-step measurement of the sky background error[2] should be easy to reproduce in any other optical survey because the measurements are very generic. Nothing the pipeline does should be unusual for another astronomical survey and there is no dependence on the SDSS specifications in any step. It may be that the cut-out box size where the local background is estimated should be adapted to different pixel sizes if necessary. We cannot foresee beforehand if any other SDSS specification might have influence, but it does not seem the case at the present stage. Regarding the training sample, its redshift and mass distribution define the range in which the NN is, in principle, effective. To what extent the NN could be applied to sources outside this data will be addressed in Chapters 5 and 6.

4.2.2 Sky error properties found by the NN. Case `skyErr` as if with `dark variance`

Figures 4.4 and 4.5, show, respectively, the PCA and tSNE methods – Section 3.1.5 – applied to the `skyErr` input space set normalized with `dark variance`. Each data point corresponds

[2] http://classic.sdss.org/dr6/algorithms/sky.html

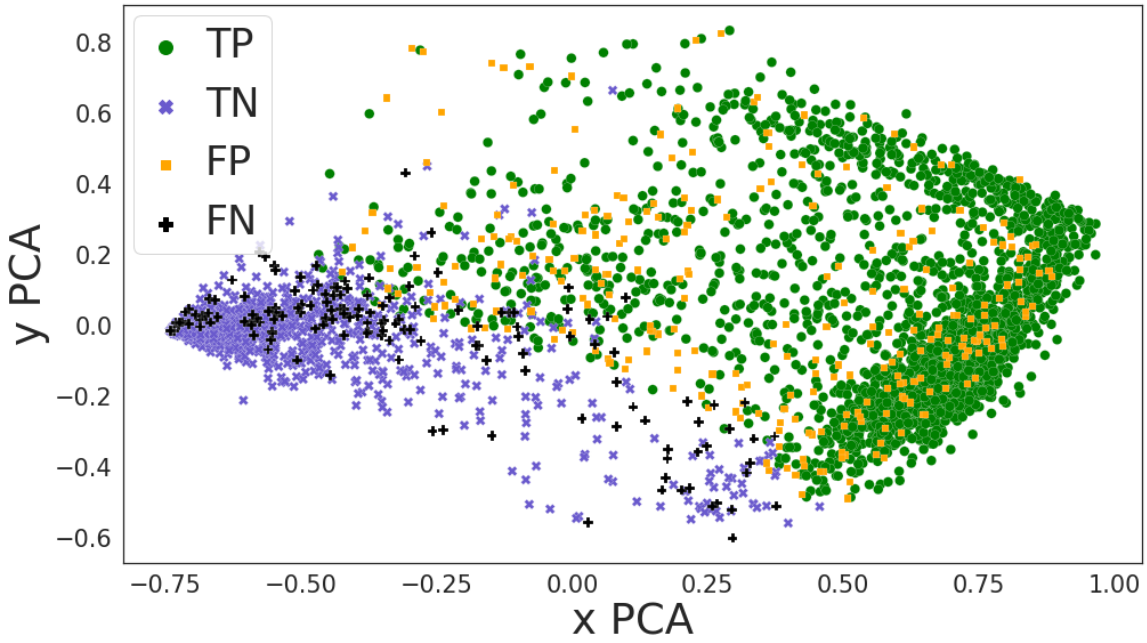


Figure 4.4: 2D embedding using PCA. Classification results from the weights saved at the validation peak of the first of the five folds: TP galaxies are shown by green circles, TNs by blue crosses, FNs by black 'x's, and FPs are in orange. The axes are the first and second principal coordinates. This colour scheme will be repeated for all the plots in the rest of the text.

to a galaxy in the 2D embedding and the colour depends on the classification type (Sect. 3.1.4). Because t-SNE is highly dependent on the initial distribution, we initialized the model based on the data's PCA, a functionality available in the `python sklearn` package that we implemented.

Through the PCA plot – Fig. 4.4 –, the main locations for mergers and non-mergers, where TPs and TNs are denser, can be seen in opposite corners of the rhomboid shape. The TPs are to the right and the TNs are to the left, while in the intermediate area, the plot is less dense and more FNs and FPs arise mixed in between. Some FNs or FPs appear also in the green and blue dense areas, respectively. This is more frequent for FPs, which indicates that the `skyErr` method still does not define an unmistakable distinction between mergers and non-mergers.

The tSNE method – Fig 4.5 – leads to very similar conclusions but from a more defined shape with a more uniform density. The TP versus TN separation is delimited in a clear way. Again, some FPs and FNs are dotting the TP and TN areas. The FNs appear rarely in the blue region; they arise mostly in the TN edges that can be found even intersecting green regions, such as near [0,-10] or [20,30].

Another approach to investigate the data is to create histograms of the inputs. For that, we show the histogram per label in Fig. 4.6 and the histogram per class in Fig. 4.7. Figure 4.6 shows that for mergers, the distribution in the u and z bands has a less steep profile than for non-mergers. The mergers show one peak around 0.2 in z that disappears for the non-mergers. It translates into a defining characteristic, as the distribution is maintained in the TPs in Fig. 4.7. In contrast, the g, r, and i bands present distributions that differ more between labels. For non-mergers, the three of them peak near 0 and decrease in number towards larger normalized errors. For mergers, there is a peak that shifts from central values in g to progressively larger ones in the other two bands. An immediate

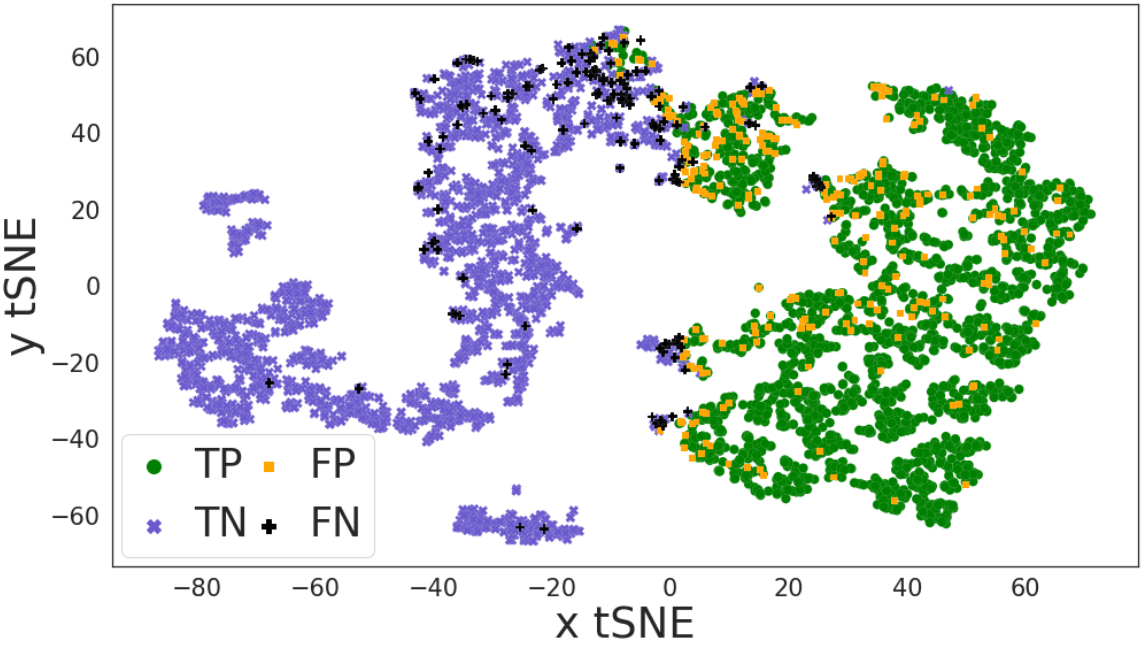


Figure 4.5: 2D embedding using tSNE, with the same classification scheme as for Fig. 4.4. The axes are simply the two tSNE dimensions.

interpretation is that the majority of mergers present an intermediate relative sky error in g, a relatively high error in r, or the highest possible value 1, for a single galaxy after min-max in i. Overall, the TPs distribution in Fig. 4.7 resembles the mergers' one. The TNs simply show a steep slope similar to that in u and z, with the FNs again also flat. The FPs histograms are quite uniform in comparison to the others. It seems that the FPs cover the part of the distribution of non-mergers that goes missing in the TN distributions. The main signs the NN is identifying become evident when comparing TPs and TNs in g, r, and i bands. The mergers have high `skyErr` values but the non-mergers have low ones. The FN and FP profiles indicate that this is neither sufficient nor clearly defined, as the dimensionality reductions were illustrating.

Figures 4.8 and 4.9 show the 2D histograms of the TPs and TNs, and the contour plots of the FNs and FPs of `skyErr` for the u versus z and the g versus r bands, respectively. The TPs and TNs in the first image are clearly separated, although some TNs can be seen in the upper right area, where the TPs are mostly located. The FP and FN contour plots show the area where the confusing galaxies are located. Similar to the patterns appearing in the dimensionality reduction figures, the FPs are mostly around the TP area and the FNs appear both in the intermediate region and near the TNs. For g-r, the location of the TPs is mostly in the upper right corner, near to a value of 1. Analogously to the first image, the TP and TN areas are clearly separated, the FPs are located mostly in the same region as the TPs, and the FNs appear both in the TN region and in the intermediate TP-TN area. Therefore, the properties in the 1D histograms can be seen translated into a 2D representation and the dimensionality reduction patterns are present.

4.2.3 Pre-normalized `skyErr`

To better understand the pre-normalized `skyErr` features, we created individual histograms as in Fig. 4.10, but using logarithmic bins to enable a better visualization. The separation between the distributions depending on the classification type is even more pronounced here than in Fig 4.7. For bands g, r, and i, the TPs and FPs are located in the upper half of

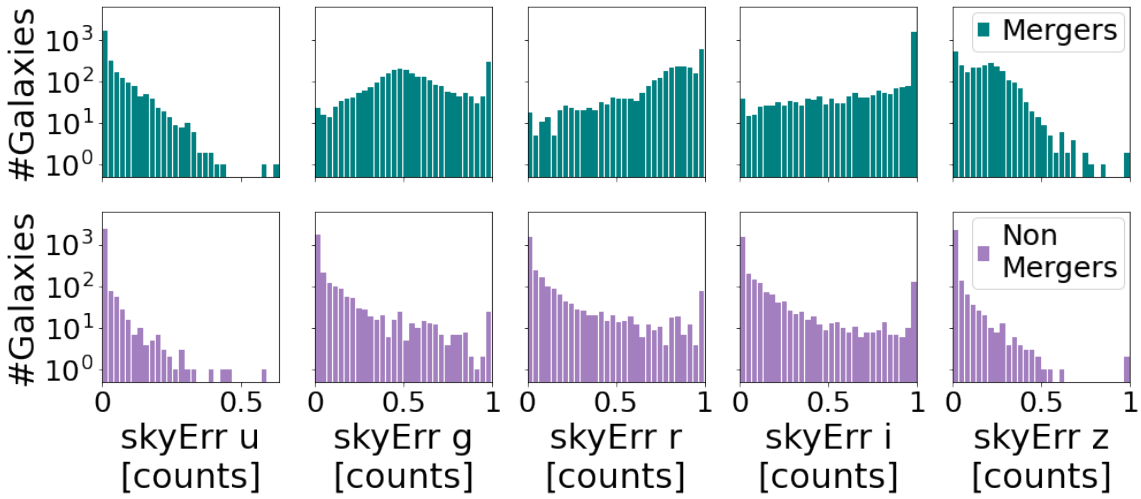


Figure 4.6: Sky background error histogram panels for the five bands. Galaxies labelled as mergers are in blue and non-mergers are in light red. The values were normalized with the dark current variance.

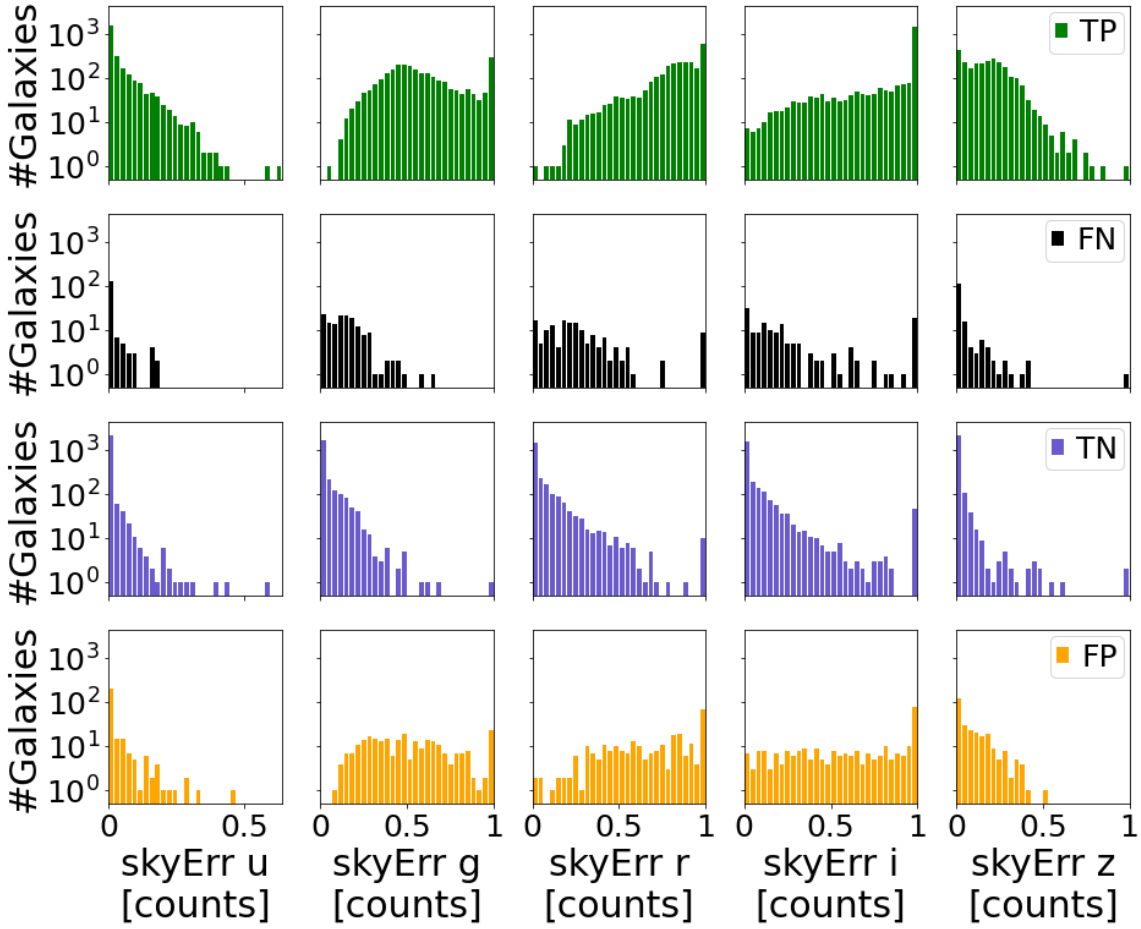


Figure 4.7: Distribution of the same variables as in Fig. 4.6, but split into the four classification types.

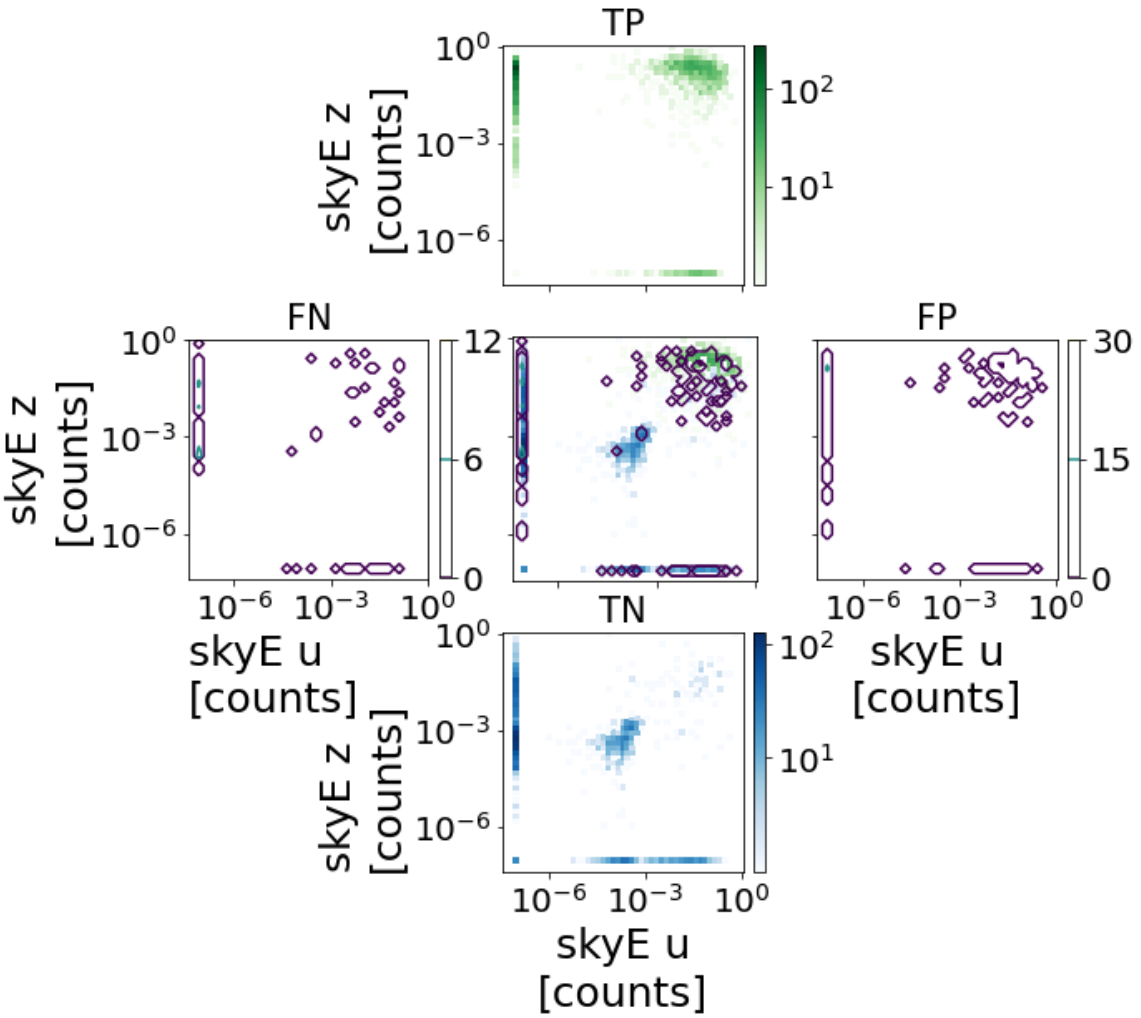


Figure 4.8: Distribution of galaxies in the 2D histograms of the TPs (in green, separated above) and TNs (blue, below), and the contour plots of the FNs (left), FPs (right), and of all galaxies (centre) for `skyErr` in the u-band vs the z-band plane. The 2D histograms show logarithmic colour-bars and the axes are in logarithmic scale. To avoid undefined values for the galaxies with post-normalization features equal to zero, a constant value of 10^{-7} was added to these. Consequently, they appear as vertical and horizontal lines at the bottom and left sides of each panel. This allows us to see what happens with those. It should be noted that some TNs are in 10^{-7} for each band, meaning the pre-normalized `skyErr` in both bands was exactly the same for those galaxies. Those are located in the bottom left corner.

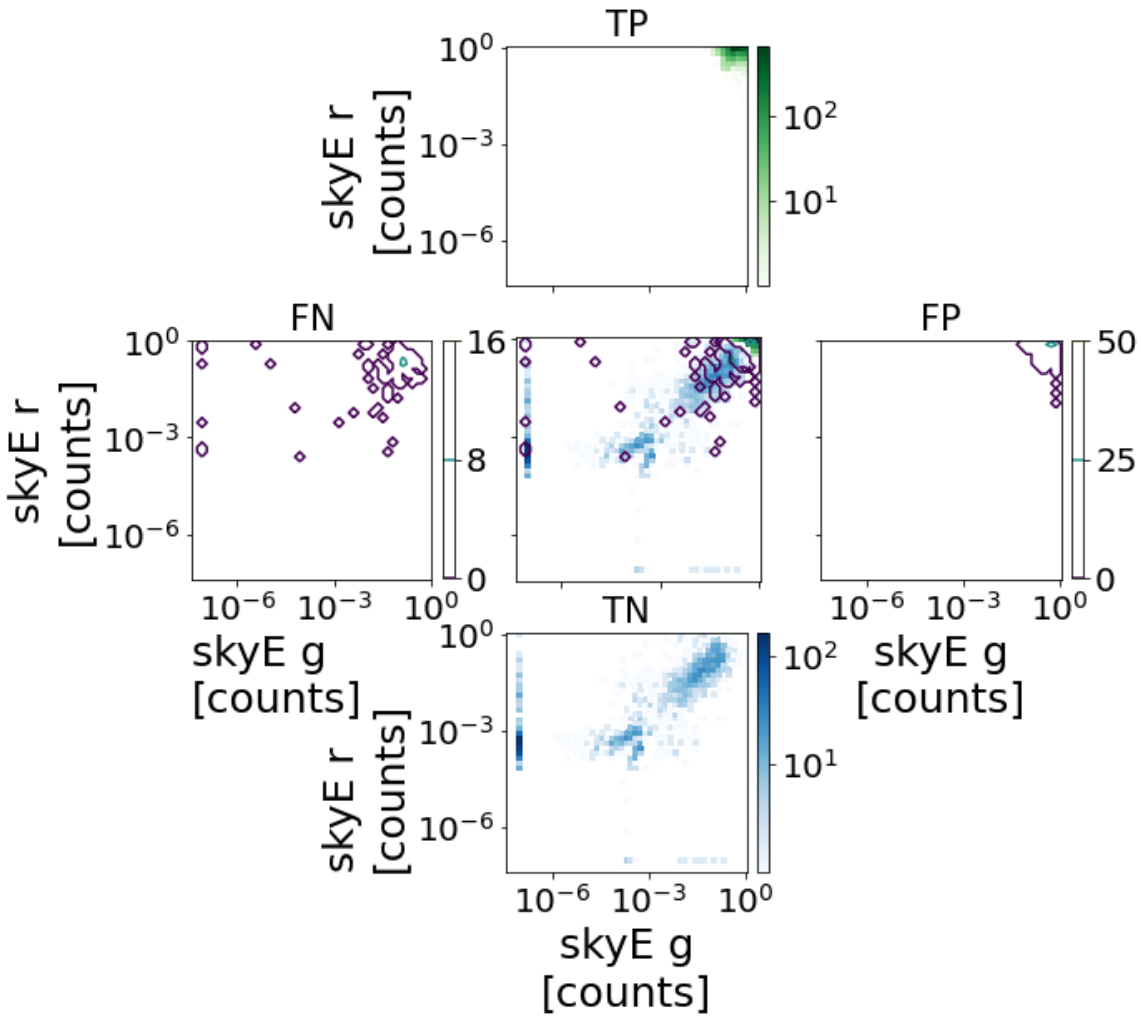


Figure 4.9: Same panels as in Fig. 4.8, but this time for bands *g* and *r*.

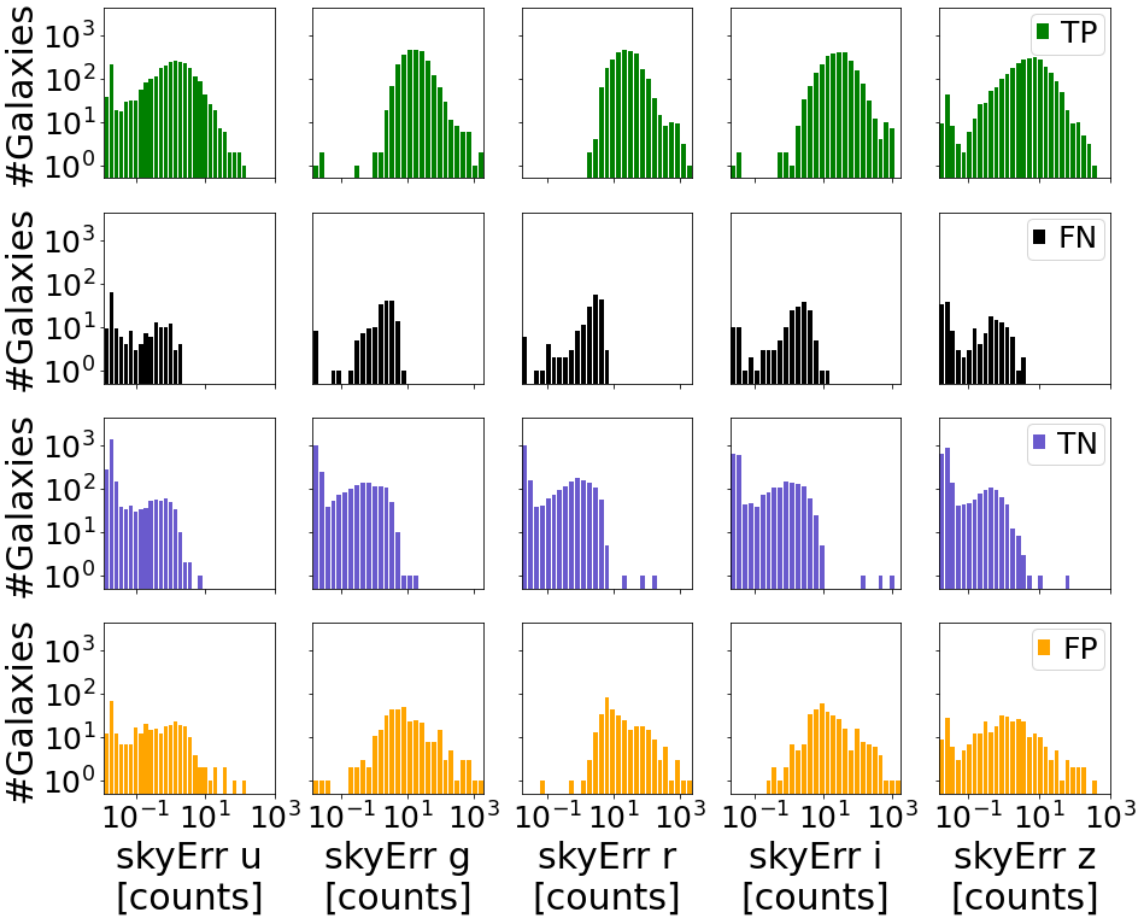


Figure 4.10: Sky background error original histograms in logarithmic bins and bin widths.

the data, and the TNs and FNs are in the lower half. This strong separation is not seen in the u and z bands. These patterns are analogous to those described in the previous section, but are even more explicit. We created one last input space with the logarithm of the `skyErr` five bands to check if the NN is aware of this difference in the data's presentation. This provided an accuracy of 92.64 ± 0.15 %, the best result obtained in this project and shown in the last row in Tables 4.6 and 4.7. The accuracy is 2% better than that corresponding to the linear pre-normalized `skyErr`, confirming the importance of the normalization. Therefore, we can conclude that the normalization of the data plays a crucial role in the success of our model.

Moreover, the shape of the histograms in the three central bands g, r, and i seems to hint at the regions of merging and non-merging galaxies in the `skyErr` space. This is very likely what the NN is identifying. Figure 4.11 illustrates not only the clear separation of the classes in the g-versus-r plane, but also that simply drawing a boundary line is capable of providing an accuracy of 91.59%. The line was built by performing a grid search, first for the intercepts using a fixed slope of -1, which is approximately perpendicular to the distribution of galaxies, and followed by a subsequent grid search for the slope with the obtained intercept fixed. A similar boundary was found for g versus i and r versus i, with accuracies of 91.16% and 90.47%, respectively. Table 4.8 compares the accuracy and the rate of mergers and non-mergers correctly identified using either the NN or the boundary cut. It shows that the boundary is less accurate at identifying the non-mergers than the NN, while it does not lose accuracy for the mergers.

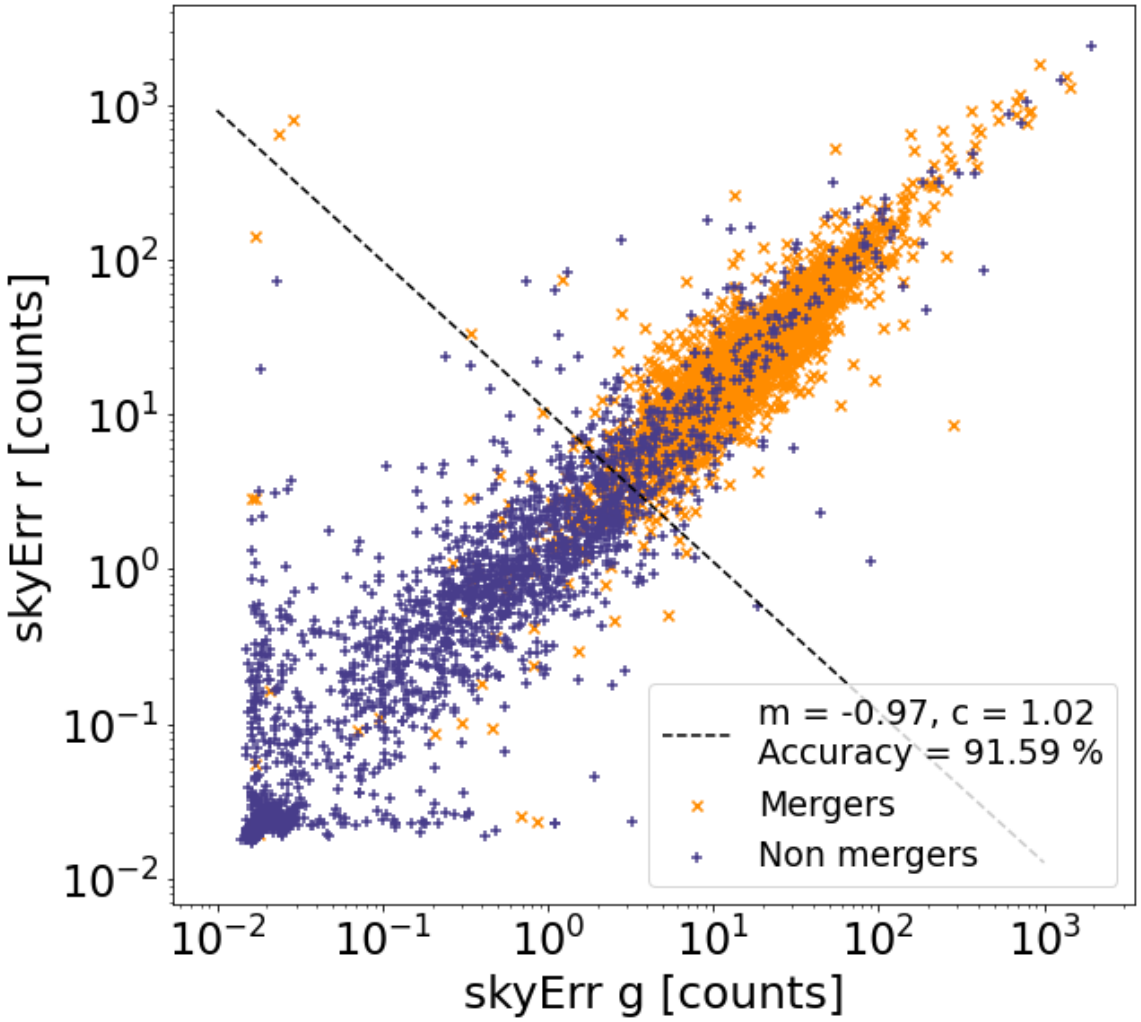


Figure 4.11: Distribution of galaxies in the 2D plane of `skyErr` in the *g* and *r* bands. The mergers are shown by orange crosses and the non-mergers by dark blue plus symbols. The boundary is the dashed black line, with its parameters given in the label, together with the accuracy of the classification using this cut.

Table 4.8: Comparison between the application of the NN or of the boundary cut to the logarithm of the `skyErr`. The rows indicate not only the accuracy but also the rate of TPs and TNs for each method when applied on the full training dataset. The NN results correspond to the saved weights of the first cross-validation fold.

Method	NN	Boundary
Accuracy	92.79 %	91.59 %
TPs rate	95.48 %	95.04 %
TNs rate	90.11 %	88.13 %

4.2.4 Sky error analysis

All these gathered results and visualizations confirm that the NN favours the input data that explicitly represent the sky error features contrasting between mergers and non-mergers. The best example of this separation is observed in Fig. 4.11. It indicates that the error in the central bands characterizes the presence of a merging process.

In physical terms, the sky error performance could have a simple explanation. Mergers produce a chaotic flow of material between the components that in some cases cannot be observed unequivocally in the images because it is not bright enough compared to the galaxy itself. This low brightness could be one of the reasons why image recognition, both by humans and by deep learning methods, can fail or can be inconclusive. Nonetheless, these merging traces could still create a detectable signal that only arises in the sky background around the mergers, and it is so low that only the differential analysis or error estimation is able to discern it. However, this has been obtained from our training dataset, which is limited in the type of galaxy mergers, specifically pre-mergers, and deepness of the images.

4.2.4.1 Deeper surveys

Our training dataset covers a specific deepness region defined by the SDSS imaging, and the galaxy's *r*-band magnitude and spectrometric redshift. As we understand it, the sky error method would require the noise around the mergers to be affected by their low-signal regions. Deeper imaging would transform blurred surroundings into sharp boundaries, impairing the method's accuracy. This makes extending the method to deeper data a profound challenge.

In order to estimate the `skyErr` method's performance on deeper data, we decided to search for galaxies within our training set that were observed in the Stripe 82 of SDSS[3]. This Stripe 82 is an area that was imaged by multiple scans, providing a magnitude that was twice as deep as the single-pass SDSS frames (Annis et al., 2014). The sky background error available for the Stripe 82 galaxies was calculated using the same pipeline as in SDSS DR7, and therefore DR6. We encountered 208 counterparts through an astrometric match. Out of those sources, we could retrieve the sky error values in counts only for 192 of them, divided into 92 mergers and 100 non-mergers.

We applied the identification methods we have built to this deeper set. The classification of the Stripe 82 galaxies obtained by locating the `skyErr` values in the decision boundary provided a 57.81 % accuracy, and by applying the NN, we obtained a 56.98 $\pm$ 0.35 % accuracy.

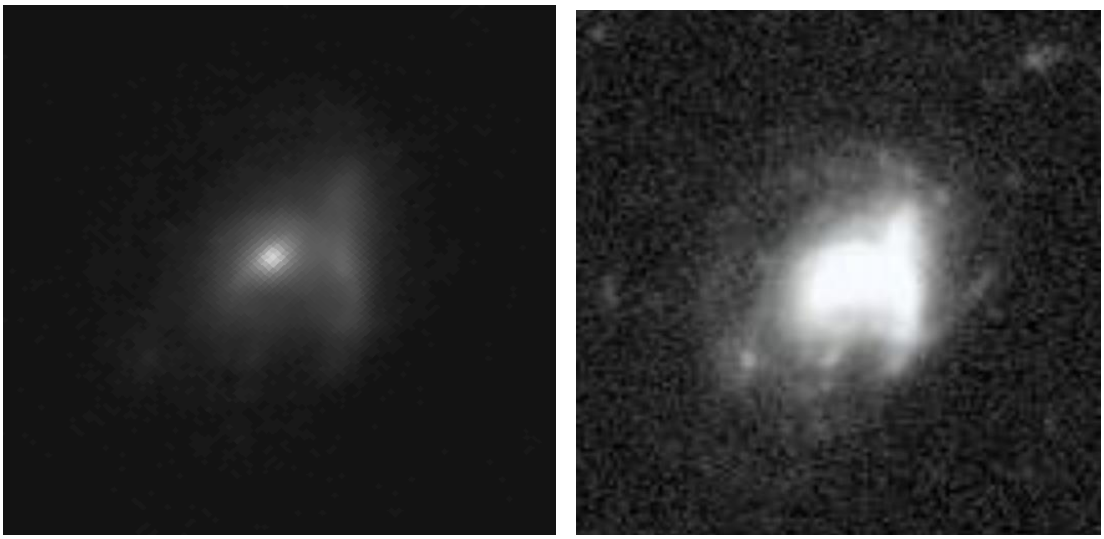
We inspected the differences between the DR6 and Stripe 82 merger observations. Figure 4.12 shows the difference between a merger properly identified in DR6 (Fig. 4.12a) but missed in Stripe 82 (Fig. 4.12b). In this example, the surroundings appear to be more diffuse in deeper data than previously found. This leads us to conclude that the reason why the deepness changes the results is the relative amount of noise and signal in the galaxy's surroundings.

4.2.4.2 Merger remnants and post-mergers:

The Darg et al., 2010a catalogue from which we selected the training mergers consists exclusively of merging pairs. As a consequence, we lack post-merging stages in our sample that can indicate whether the sky error method would also identify them or not.

In order to find merger remnants using our current method, we built a catalogue of galaxies in SDSS DR6 within the *r* magnitude and spec-z intervals of our training set. These

[3] http://cas.sdss.org/stripe82/en/



(a) DR6 (b) Stripe 82

Figure 4.12: Astronomical frame of a galaxy labelled as a merger from our training dataset. Both images correspond to the *r*-band. On the left side, the DR6 frame is shown, and on the right is the deeper Stripe 82 frame.

intervals were [12.24, 18.05] for *r*-mag and [0.01,0.1] for the spec-z, providing up to 286 616 sources. We then carried out two main studies. First, we located them in the `skyErr` decision boundary and visually inspected different regions in the merger's upper half. Second, we made use of the classification in the Galaxy Zoo DECaLS (GZ-D) Campaign 5 (Walmsley et al., 2022). Galaxies with a vote fraction for the classification answer 'major disturbance' above 0.6 were defined as post-mergers (Walmsley et al., 2022). Our goal was to visually inspect the astrometric matches with the SDSS DR6 set of these GZ-D post-mergers. We made a lower cut on the number of votes per galaxy to both reduce the inspection time and to make sure they were extensively visualized, avoiding sources that were picked out by their variable retirement rate (Walmsley et al., 2019). The resulting post-merger catalogue contained 45 galaxies.

Among the SDSS DR6 sources we inspected, we did find at least one clear post-merger that had been correctly identified by both the NN and the decision boundary. Among the GZ-D 45 confirmed galaxy post-mergers, only seven of them were found in the merger region. Using the NN, we obtained the same classification. Those galaxies all showed a surrounding material that mixed with the background. Except for the missed merger remnants, the surrounding seemed to be less diffuse than the other seven galaxies.

4.3 Conclusions

We created an NN and applied it to a class-balanced set of mergers and non-mergers using only photometric information. The dataset was composed of galaxies from SDSS DR6 identified during GZ DR1. The 2 930 mergers from Darg et al., 2010b were combined with the same number of non-mergers in GZ DR1 by a nearest-neighbour match in spec-z and *r*-band magnitude. The NN applied was fully connected: it had two layers with 16 neurons whose activation function is ReLU and it had a dropout rate of 0.2; the learning method was Adam with an initial training rate of 5×10^{-5}; the output was a softmax probability for a two-class classification; and the classifier was the TensorFlow's BinaryCrossentropy class.

First, we used the band magnitudes, colours, and errors of six different SDSS flux measurement methods. Using the model magnitude, we found a reference accuracy of 68.90 $\pm$ 0.72%. Checking the other magnitude types brought up the importance of the fibre magnitude error. This fibre errors provided an 83.76 $\pm$ 0.32 % validation accuracy alone and 88.68 $\pm$ 0.31 % with bands and colours. Further research showed that the components of the fibre magnitude error could achieve an accuracy of 91.48 $\pm$ 0.31 %. We found that the parameter that contributed mainly to this high accuracy is the sky background error. We proved that the sky error is able to show differences between mergers and non-mergers that can be identified in the histogram, PCA, t-SNE, and 2D histogram representations, together with the NN results. Finally, we found that the input space of the logarithm of the pre-normalized five-band sky background error in units of counts is able to reach a validation accuracy of 92.64 $\pm$ 0.15 %. A version of the NN for this last input is published on `GitHub`[4] with the saved weights. Moreover, the NN could be substituted by a decision boundary in the planes between the g, r, and i bands, achieving an accuracy of up to 91.59 % for the g-versus-r plane.

A likely interpretation of this result is that the higher values of the sky background error reflect the traces of merging processes – for example, faint tidal tails – otherwise missed by the sky background measurement due to the dominance of the signal from a galaxy itself. Moreover, we think that the multi-band analysis of the sky background error additionally makes our network and decision boundary sensitive to the colours of this residual flux that originates from the matter surrounding a merging galaxy.

[4] https://github.com/LuisEduSuelves/NN16_skyErr-log

CHAPTER 5

Decision Tree

The simplification into a decision boundary of the NN trained by the five-band `skyErr`, as described in Chapter 4's discussion and depicted in Fig. 4.11, has the potential of facilitating galaxy merger classification in large sky surveys. With the goal of understanding the boundary when a more general dataset is applied, we took the whole catalogue of galaxies observed in Galaxy Zoo Data Release 1 (GZ DR1), as NN's training sources were a subset of it. This chapter presents a proof of concept of the method which we plan to fine tun further before its submission for publication.

5.1 Results

The results came from the visual morphological classifications and contaminations found, plus the subsequent decision tree (`dt`) aiming to reproduce those contaminations. We visually inspected and indicated the classifications and contaminations for GZ DR1 sources distributed in nine-galaxy groups. Along the text, we denominate those nine-galaxy groups as subsamples defined by four labels. This is because, as described in Sect. 2.1.4, they were drawn from groups of galaxies, that we denominate as subsets, that were selected depending on said four subdivisions. Here, we list the labels for the properties that combine to define any given subsample:

- Magnitude label:
  - r1 $\in$ [$\sim$ 12, 14]
  - r2 $\in$ [14, 16.5]
  - r3 $\in$ [16.5, $\sim$ 18j]
- GZ morphology flag:
  - Elliptical
  - Spiral
  - Uncertain
- Merger vote fraction f_m:
  - $f_m = 0$
  - $f_m \in (0, 0.2]$

  - $f_m \in (0.2, 0.4]$ (only for galaxies with Uncertain GZ flag)
  - $f_m \in (0.4, 1]$ (only for galaxies with Uncertain GZ flag)

- Area in the diagram (described below in Sect. 5.1.1:
  - Mergers
  - Non-mergers A
  - Non-mergers B
  - Top
  - Bottom
  - Right
  - Left

5.1.1 Areas of the `skyErr` diagram:

The seven areas in which the GZ DR1 subsets have been distributed were based on the application of the alpha shape method – described in Sect. 3.2.4 – to the denser area of mergers and non-mergers of the diagram.

First, when extending the methodology to the larger data GZ DR1, the `skyErr` in units of `counts`, as used for the NN, is not available through the CasJobs webpage. This is because the sky error in counts was obtained from the uncalibrated catalogues fpObjc defined in Table A.1 at the Appendix A. Instead, the sky background for the full GZ DR1 set can only be obtained in units of `maggies`. As a consequence of changing the `skyErr` units, the grid search had to be redone with this new unit in order to create a new decision boundary analogous to Figure 4.11. The process was similar to how it was described in Sect. 4.2.3. The resulting boundary parameters and accuracy are indicated in the label of Fig. 5.1.

Alpha shape on mergers: In order to constrain the region of mergers in the extension to SDSS DR6 – Fig. 5.1 –, first we selected all the mergers from the training dataset above the decision boundary. Then we run three subsequent alpha shape algorithms through the python `alphashape` package[1]. The first iteration, depicted in the left panel of 5.2, was done with an $\alpha = 28$. It can be seen that multiple isolated outliers are not included. The second iteration, shown in the central panel, was rerun with the same $\alpha = 28$ on the sources within the alpha shape of the previous step. After this, the final alpha shape with $\alpha = 0$ was formed, which does not discard any point, as shown in the right panel of 5.2. This is the leaf-shaped region that we considered for mergers.

Alpha shape on non-mergers: A similar procedure was done for the non-mergers from the training dataset. The reason for running it was to locate where the non-mergers cluster in the diagram. In this case, two subsequent iterations of the alpha shape algorithm with $\alpha = 20$ were run – Fig. 5.3 top left for the first iteration and right for the second. The reason for choosing a smaller α was that non-mergers in the bottom side of the diagram appeared to be less densily clustered than the mergers above. As a result, the non-mergers area included less dense parts of the distribution. While arbitrary, this decision was motivated by covering a larger portion of the non-merger side of the diagram.

After two $\alpha = 20$ iterations, it became clear that there were two main areas of non-mergers in the diagram, and therefore we created two final alpha shapes with $\alpha = 0$, the non-mergers areas A and B shown in the left and right panels of Figure 5.3, respectively.

[1] https://pypi.org/project/alphashape/

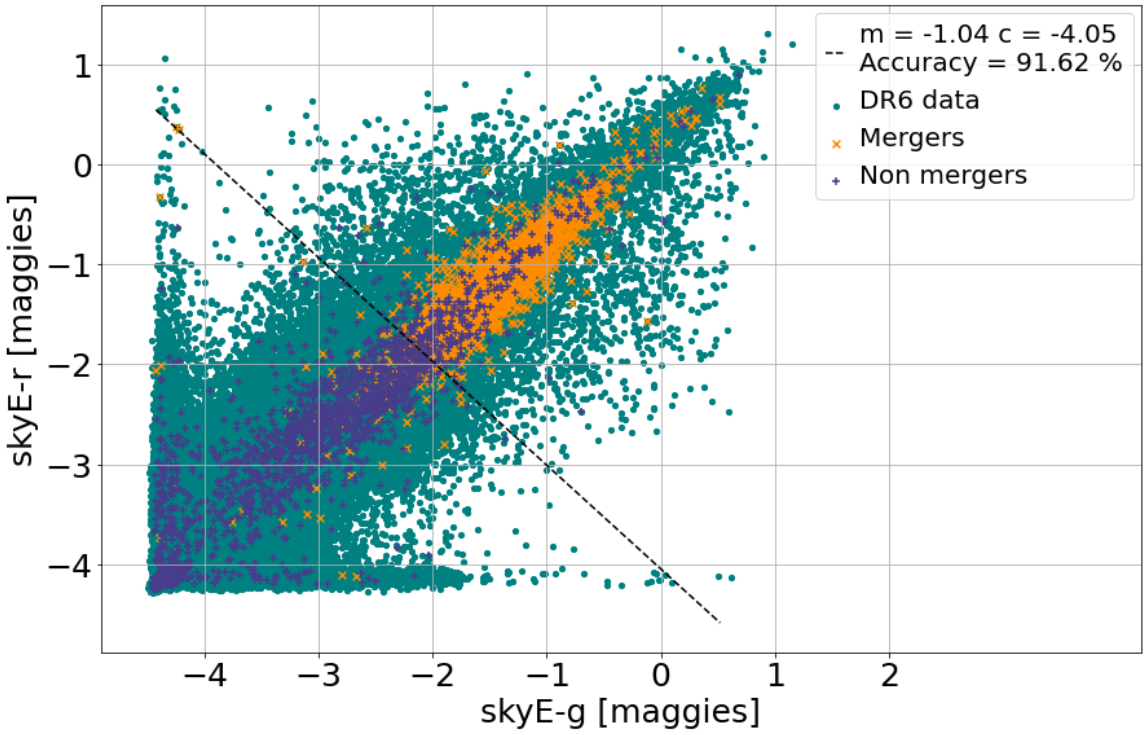


Figure 5.1: Distribution of all GZ DR1 galaxies from SDSS DR6 in the 2D plane of `skyErr` in units of `maggies` for g and r band in the x and y-axis respectively. The mergers and non-mergers are represented as in Figure 4.11, and the galaxies forming the full dataset are the green dots plotted behind. The new boundary dashed black line, with its parameters given in the label, shows the new accuracy obtained on the training data classification.

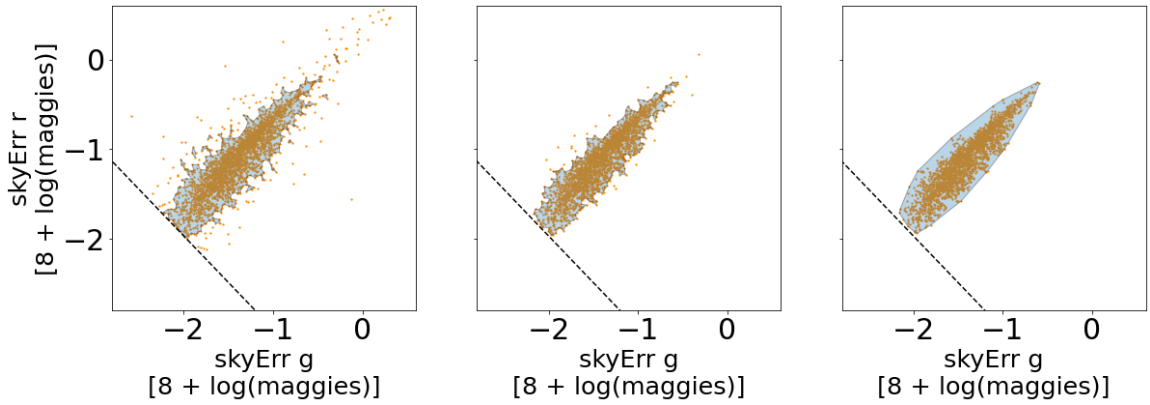


Figure 5.2: Alpha shape stages applied on the training merging galaxies located above the decision boundary of Figure 4.11. The galaxies are plotted as small orange dots. The left panel shows the whole set of mergers and the alpha shape generated with $\alpha = 28$. The central panel shows the second alpha shape iteration, applied on the galaxies located inside the first shape, which are the only ones plotted. It was also run with $\alpha = 28$. The right and final panel shows the mergers within the second shape, and final iteration done with $\alpha = 0$.

When looking at the leaf-shaped merger area and the two non-mergers area, it can be argued that some quite empty portions of the diagram have still not been discarded. While this is in part because of the arbitrary selection of the α parameters, it is important to consider what is the actual distribution of training galaxies covered in the two panels of 5.4. On the left panel all the training mergers are shown, and on the right panel all the training non-mergers, clarifying the defined areas are not as empty as it looks like.

Surroundings of the merger "leaf": The following step was to define the surroundings of the leaf-shaped merging area. The goal of this is to characterize how the galaxy classification changes along the four directions of the diagram. The rectangular box was arbitrarily built to cover 0.5 units in log(`maggies`) from three directions. One goes from the upper-right tip of the leaf in the direction perpendicular to the boundary. The other two directions start from the two corners at the lower-left base of the leaf, and are pointed respectively to the sides in the direction parallel to the boundary. The extension of the rectangle in the bottom-left is limited by the presence of the boundary itself. The rectangle is formed by joining the outer edges of the resultant shape.

This formed four different regions. Figure 5.5 depicts the three areas defined by the alpha shape algorithm and the four areas surrounding the mergers' leaf. The name of these four areas comes from their location respectively to the leaf, when rotation the plot so that the decision boundary appears as horizontal: the yellow triangle would appear in the Top part, the grey and teal side wings would appear in the Left and Right, and in between the leaf and the boundary a Bottom small section appears. The Bottom section is shown zoomed in with more detail at 5.5b.

5.1.2 Visual inspection

Because of the complexity of the data separation, these subsamples are not representative of the underlying distributions. This is exemplified in Table 5.1, that compares a proxy to the underlying distribution with the average of all subsamples within the Top area. While there are similar tendencies between the two distributions, such as a large contamination by stars or a similar amount of mergers, the overall disagreement supports that both distributions are different.

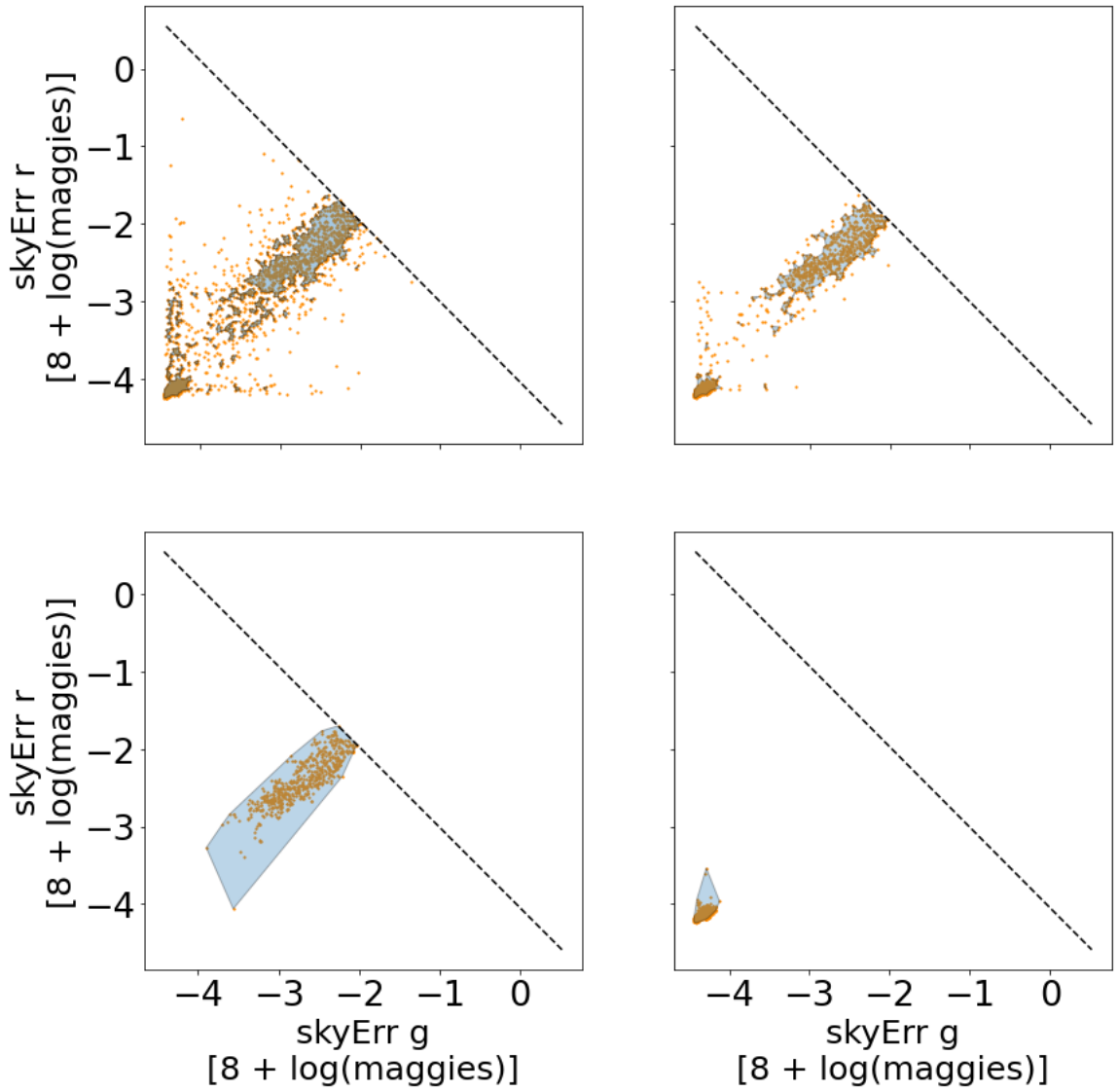


Figure 5.3: Alpha shape stages applied on the training non-merging galaxies located below the decision boundary of Figure 4.11. The galaxies are plotted as small orange dots, as in Fig. 5.2. The top-left panel shows the whole set of non-mergers and the alpha shape generated with $\alpha = 20$. The top-right panel shows the second alpha shape iteration, applied on the galaxies located inside the first shape, which are the only ones plotted. It was also run with $\alpha = 20$. The bottom-left and bottom right panels show the two final shapes, generated separately, where the majority of non-mergers are located.

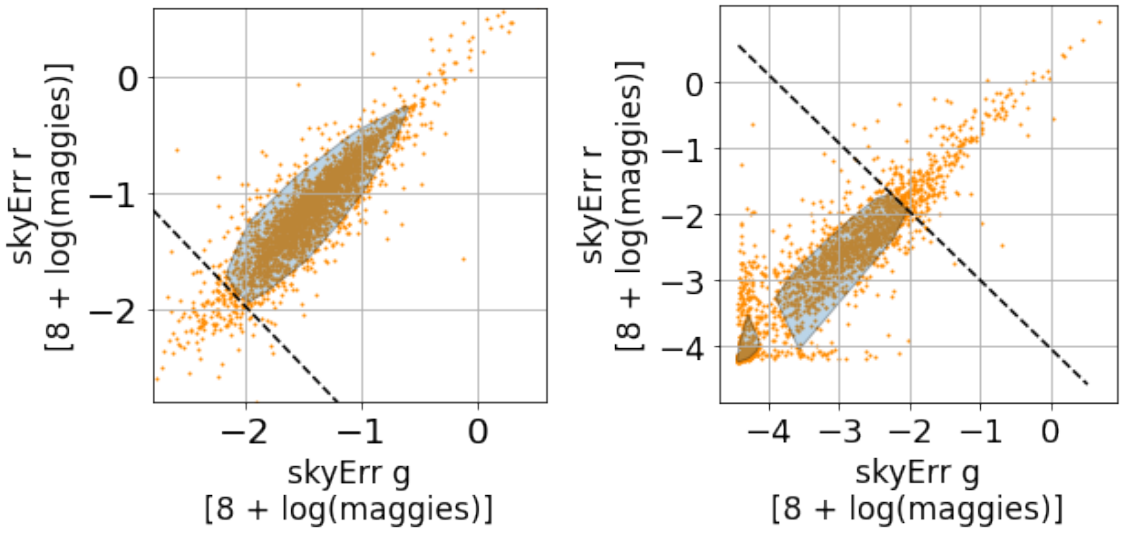


Figure 5.4: Final Merger area – on the left – and Non-Merger A and B areas – on the right –, with the scatter plot of the distributions of training merging and non-merging galaxies, respectively.

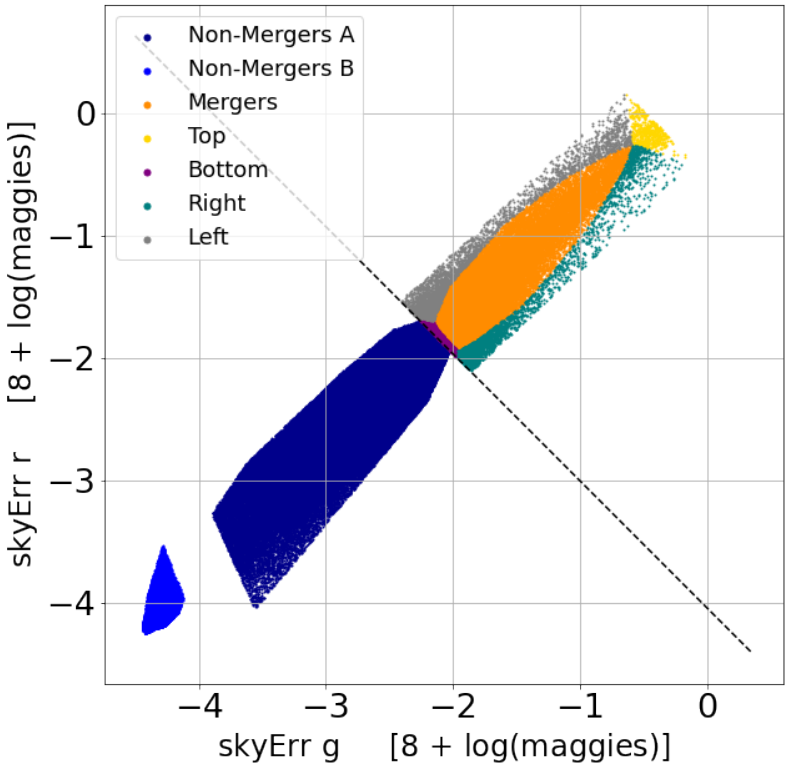
This can be expected because the four-parameter subsets in the Top area have quite heterogenous sizes. For example, the total number of galaxies in the subset [r2, Elliptical, f_m = 0, Top] is much larger than the number of galaxies in the [r1, Elliptical, f_m = 0, Top] subset. The visually inspected subsamples drawn from these two subsets both have nine galaxies. However, the votes of a galaxy in [r1, Elliptical, f_m = 0, Top] are more representative of the subset's distribution than a galaxy in [r2, Elliptical, f_m = 0, Top].

The final votes could be weighted by the relative size of all subsets, as calculated in the last row of Table 5.1. The importance of the weighting for extracting relevant information can be considered to be small due to multiple reasons. First, because the objective of analysing each subsample separately is to understand what type of galaxies can appear depending on the four-parameter selection. Second, because the analysis of the whole dataset has as a goal to find clean and dirty galaxies along the diagram. In other words, the underlying distribution is less important than identifying contaminants. Finally, because there is not a high improvement in recovering the underlying distribution by using the weighted average. Only nine galaxies were used for checking the underlying distribution, as shown by comparing the first and third rows of Table 5.1. A similar tendency was found for the other six areas in the `skyErr` decision diagram.

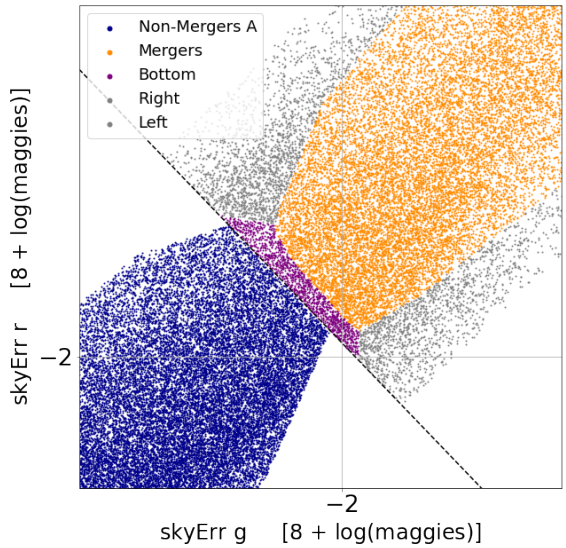
5.1.2.1 Galaxy types

From the visual inspection, we found multiple combinations for subsamples that provided statistically significant samples of mergers. These morphological classifications are shown without taking the contamination into account. The percentages determined in the following lines are calculated with respect to all the galaxies in a described combination of subsets, e.g., a 50% of mergers in Elliptical galaxies with f_m = 0 of the Merger area for all magnitudes means that the percentage has been calculated for the following four-parameter subsamples: [r1, Elliptical, f_m = 0, Top]; [r2, Elliptical, f_m = 0, Top]; and [r3, Elliptical, f_m = 0, Top].

Regarding the major mergers, there is a 63.82% of mergers among all areas and magnitude bins for Uncertain galaxies with $f_m \in (.4, 1]$. This is consistent with the criteria in which the catalogue in Darg et al. (2010a) was built, because they visually confirmed



(a) Complete decision boundary.



(b) Zoom in into the Bottom area.

Figure 5.5: Areas generated in the decision diagram – Figure 4.11 – through the process described in Section 5.1.1. The dots correspond to galaxies in each area, and the dot size is very small to better depict the areas themselves. The colour "pattern" change with each area. The Mergers area has its galaxies in orange, while the two Non-mergers areas A and B are in dark blue and blue, respectively. The Top area appears in yellow, the Left wing in grey, the Right wing in teal, and the Bottom in purple. The second image in below zooms into the separation between areas near the decision boundary itself, with the aim of showing better the shape of the Bottom area.

Table 5.1: This table shows the results of the six visual inspection options for the Top area, calculated in three different ways. The average row was calculated by dividing all the votes for each visual inspection option by the total number of galaxies in the combined subsamples in the Top area. The underlying distribution row was calculated through a proxy obtained by the visual inspection results of nine galaxies drawn from a flat distribution of all galaxies in the Top area. The weighted average row was calculated by the weighted sum of each subsample's votes. This weight was obtained by the relative size of the subset from which the subsamples were drawn, with respect to the total size of the Top area's catalogue.

	maj merg	other merg	non merg	cont blend	cont overlp	cont other
Top: average	12.26	19.35	68.39	87.10	18.06	4.52
Top: underlyig distribution	11.11	33.33	55.56	77.78	33.33	0.0
Top: weighted average	**8.12**	**23.21**	**68.68**	**92.93**	**15.50**	**3.78**

sources in that same vote fraction range. Besides, for the Uncertain galaxies with $f_m \in$ (.2, .4], a 26.00% of major mergers were found.

For other mergers, we found that for Spiral galaxies in the Left, Right, and Bottom areas, in all magnitude and all f_m bins: the percentage of other mergers observed is of 41.18%, 40.38%, and 38.89%, respectively. More specifically, for Spirals in the Left wing with $f_m \in$ (0, 0.2] there is 45.83% among all magnitudes, and similarly, for Uncertain galaxies with $f_M \in$ (0.2, 0.4] among all magnitudes, there is a 45%. Moreover, for Spirals in the Right wing with $f_m \in$ (0, 0.2] among all magnitudes there is 56% of other mergers. Finally, for the Bottom area the amount of other mergers for Spirals with $f_M \in$ (0, 0.2] among all magnitudes is of 51.85 %, and for Uncertain galaxies with $f_M \in$ (0.2, 0.4] among all magnitudes it is a 40.73%.

This not only confirms the potential of the decision diagram for using the sky background error to find galaxy mergers of any types, but also provides a potential recipe to locate multiple type of mergers by studying different regions of the diagram.

5.1.2.2 Clean and dirty sources

Figure 5.6 indicates the location in the decision diagram areas of Fig. 5.5, of galaxies showing the three morphological options from visual inspection. The galaxies found as clean, meaning that none of the three contamination possibilities were found for them, are located in the right column, and the dirty contaminated galaxies are in the left column.

Regarding major mergers, they appear as clean sources along all the diagram in a relatively homogenous way. This is because we have selected galaxies with high f_m in all the areas. Besides, as described above, a 63.82% of those galaxies were indeed major mergers. However, while the clean mergers get distributed along all the diagram, the dirty mergers can be found mainly in three regimes: in the leaf-shaped Merger area; in the Left, Right, and Bottom regions, although with less frequency; and mostly in the Top triangle.

The other identified type of mergers show a distribution similar to the major merging pairs, although with a shift towards the bottom-left corner of the plot in the direction orthogonal to the boundary, as can be seen in the clean other-mergers panel. They also

appear contaminated more often on the Right and Left wings than in the Merger's leaf. Again, the Top area is the dirtiest one.

The key result of these plots is found in the panels of the non-merging galaxies. While they appear cleaner the closer they are found to the bottom-right corner of the plot, they appear dirty in a relatively uniform way across all areas. This is the main motivation for creating a decision tree that is able to identify contaminated non-merging galaxies. If the dirty sources are discarded by the decision tree, then the sky error boundary can be used to find a sample of merging sources with a higher percentage of correctly identified mergers, as shown in the top and central-left panels. The percentage of galaxies that we observed, averaged among the seven areas in the diagram, to show contamination by stars was around 40%, by visual pairs around 30%, and by artefacts around 5%.

Figure 5.7 shows the distribution of sources contaminated by either nearby stars or by visual pairs, separated by the visual morphological classification. As the left columns indicate, the majority of sources with a nearby star contaminating its sky background are located in the Top area. The density of the star-contaminated sources decreases as one moves from the Top area in the bottom-left direction perpendicular to the diagram.

Comparing the contamination of major mergers and of other mergers by stars, by galaxy pairs, or by all contamination as shown in Fig. 5.6, one can conclude that the contaminated mergers have both stars and visual pairs around. Because the amount of artefact-only contamination is quite small – as indicated above, it is an average of ∼5% among all areas –, the dirty panels in Fig. 5.6 have to be populated by galaxies contaminated both by stars and by visual pairs in order to be consistent with Fig. 5.7.

5.1.3 Decision tree

With the goal of finding galaxy mergers in the sky error diagram, it is necessary to avoid those sources that do not have a high sky error because they are mergers, but instead because of the presence of contamination. These so-called clean and dirty sources have been identified in the visual inspection results described above in Section 5.1.2.2. The decision tree thus attempts to find as many of those dirty sources without discarding the clean ones.

Given the relevant statistical concepts described in Sect. 3.2.3.4, Table 5.2 shows the accuracy, recall, and specificity out of `dt` for merging galaxies, and Table 5.3 the same but for non-mergers. Here mergers are sources that were considered as either merging pair or other type of mergers during the visual inspection. The accuracy for mergers is around a 62% and for non-mergers around 64%, making the model better than a random choice, but not very strong. The recall and the specificity for mergers is ∼72% for both. For non-mergers, the recall and specificity are ∼67% and ∼55% respectively

The main goal of the decision tree is to find dirty non-mergers while keeping a high number of clean mergers. Thus, the recall of dirty non-mergers and the specificity of clean mergers are the most important results to assess the quality of the method. In fact, the internal parameters of the decision tree have been fine-tuned to optimize the two statistics. Those hyperparameters are the flags employed in the catalogue of surrounding galaxies, and the distance from the target galaxy for which stars and galaxies are considered to be contaminating either by stars or visual pairs. They are described in see Section 3.2.3. This distance is fully dependent on how the visual inspection was performed, and thus it is a good approximation of the visual criteria applied.

Figure 5.8 shows the density of galaxies along the `skyErr` decision boundary. The two bottom panels indicate the density for clean and dirty galaxies in the right and left respectively. The other four panels indicate how the classification types distribute along

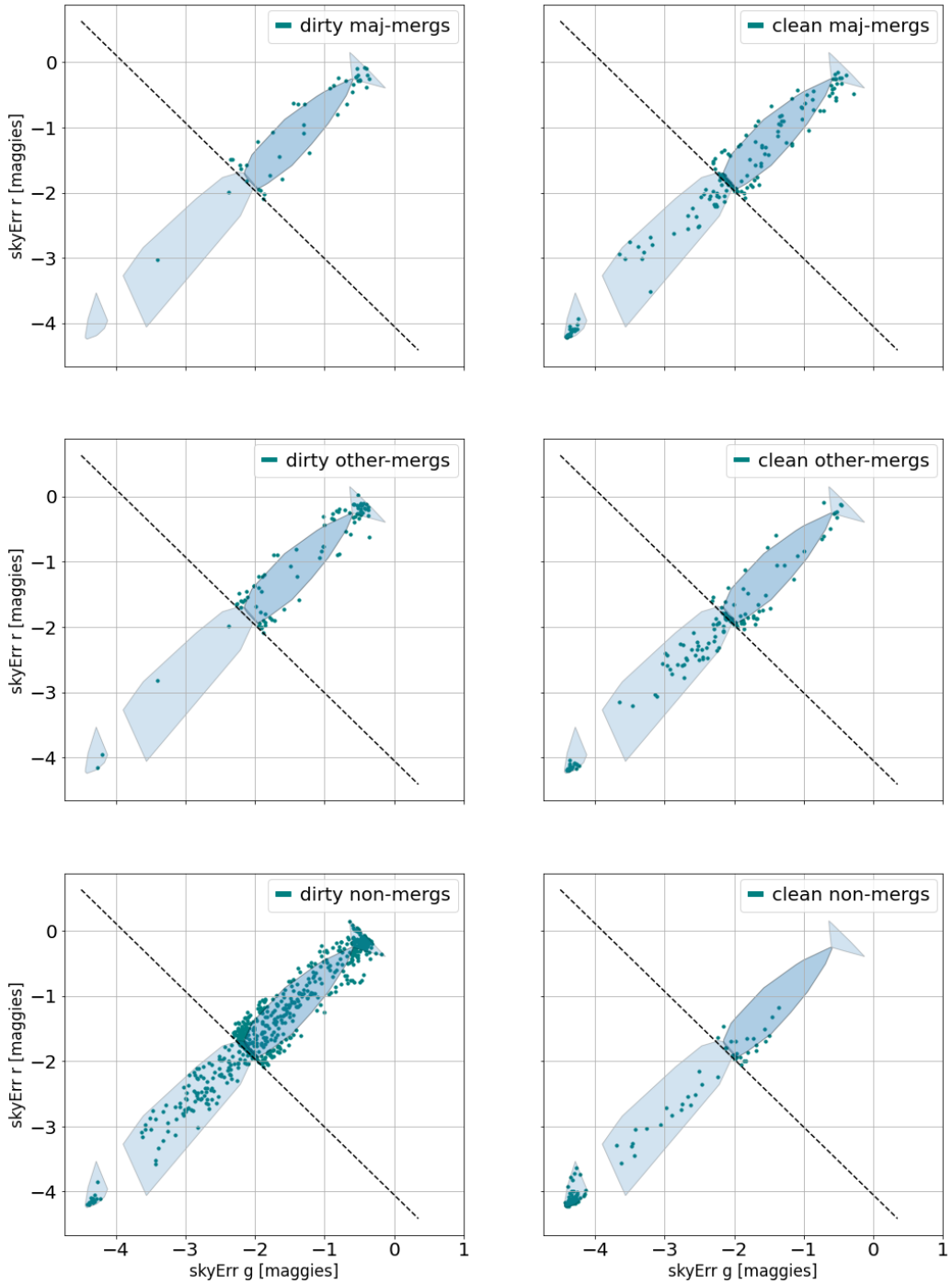


Figure 5.6: Scatter plots indicating the location of the visually inspected galaxies. The dirty and clean galaxies are in the left and right columns respectively. The major mergers, the other types of mergers, and the non-merging galaxies, are in the top, central, and bottom rows. The panels show the decision boundary, the two Non-mergers areas, the Merger leaf and the Top triangle. The Right, Left, and Bottom areas are not shown but can be guessed by the shape of the included regions.

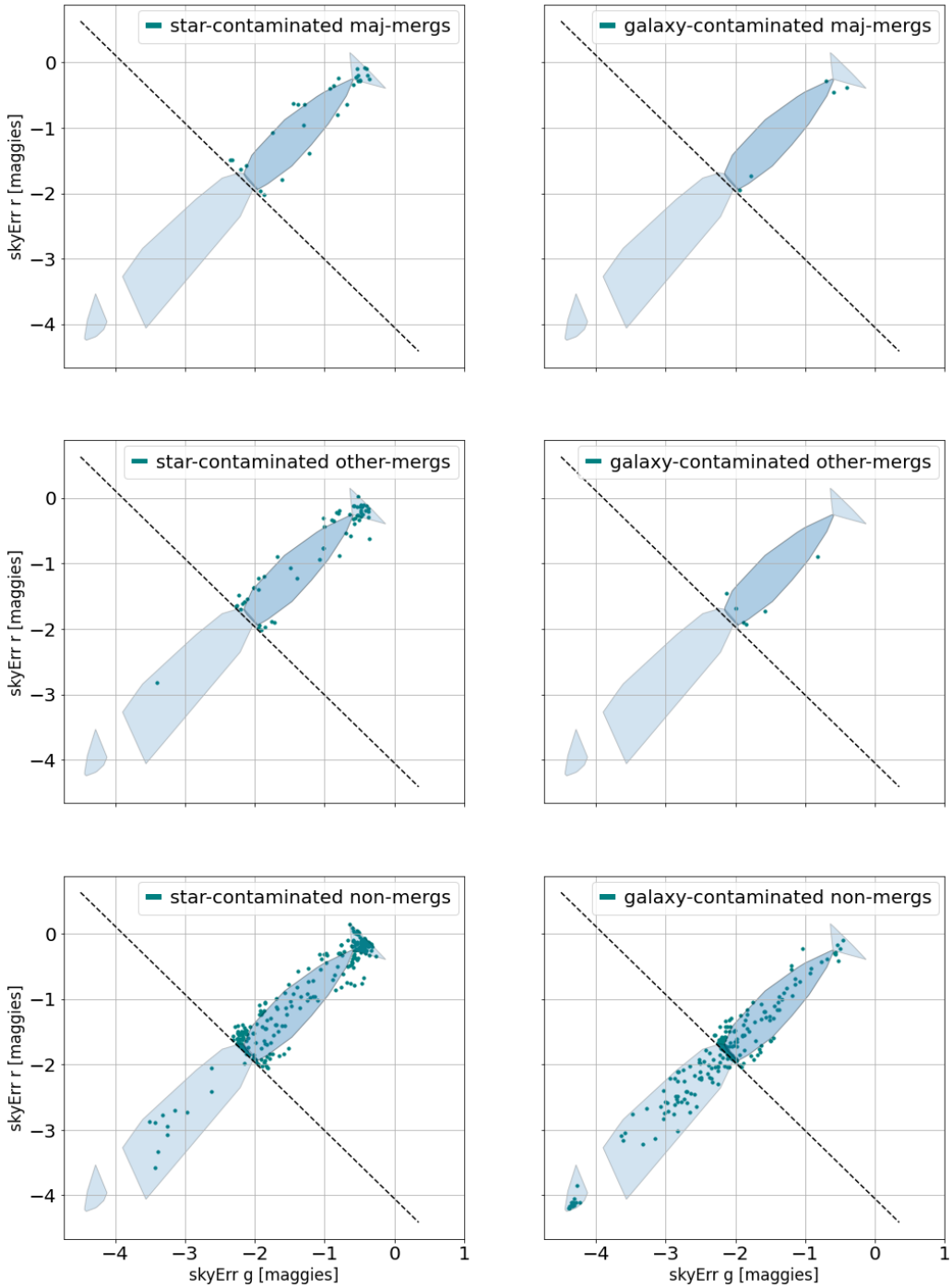


Figure 5.7: Scatter plots indicating the location of the visually inspected galaxies, in this case showing the distribution of sources contaminated exclusively by stars or by galaxies. The sources contaminated by stars are located in the left column, and those contaminated by galaxies acting as visual pairs, on the right column. The rows show the same galaxy morphologies as Fig. 5.6, and the areas of the boundary are also the same as depicted in Fig. 5.6.

Table 5.2: Table showing the accuracy, recall, and specificity from `dt` as described in 3.1 of all galaxies visually classified as mergers. The third column from the left includes the formula used for calculating it, and the last column described the implication of the parameters.

statistic	value	formula	description
Accuracy :	62.72 %	$\frac{TPs+TNs}{all}$	
Recall :	**71.92** %	$\frac{TPs}{TPs+FNs}$	Fraction of dirty mergers correctly identified by `dt` with respect to the total number of mergers considered as dirty by the visual inspection
Specificity :	**72.44** %	$\frac{TNs}{TNs+FPs}$	Fraction of clean mergers correctly identified by `dt` with respect to total number of mergers classified as clean by the visual inspection

Table 5.3: It shows analogous information to Table 5.2, but for galaxies visually identified as non-mergers.

statistic	value	formula	description
Accuracy:	64.16 %	$\frac{TPs+TNs}{all}$	
Recall:	**67.07** %	$\frac{TPs}{TPs+FNs}$	Fraction of dirty non-mergers correctly identified by `dt` with respect to the total number of non-mergers considered as dirty by the visual inspection
Specificity:	55.35 %	$\frac{TNs}{TNs+FPs}$	Fraction of clean non-mergers correctly identified by `dt` with respect to total number of non-mergers classified as clean by visual the inspection

the boundary. These panels do not have a common normalization, with the intention of showing more clearly the distribution within each of them.

While the dirty galaxies are located all along the diagram, the clean sources cluster in the bottom-left corner, where the Non-mergers B area is. Consequently, the majority of TNs, which correspond to clean sources correctly determined to be clean by the decision tree, also cluster there. At the same time, the FPs, clean sources found to be dirty, cluster at the Non-mergers B area and also at the Bottom area. The cluster at the Bottom area can be justified by the `dt` finding minor mergers to be dirty sources. This can be expected because of the combination of two arguments. One is that the visual pair branch finds contamination when the neighbour galaxies are located at a distance to the target that often overlaps with the distance where a minor merging pair starts showing tidal features. The other argument is that the small galaxies don't usually have spectrometric measurement, and therefore they cannot be corrected by the cross-pair branch.

The TPs, dirty sources identified in `dt` as contaminated, show a similar distribution as the dirty galaxies. The FNs, dirty sources identified as clean, cluster in the areas Non-mergers B, Bottom, and Top. This again follows the distribution of dirty galaxies.

5.2 Discussion

5.2.1 Visual Inspection

The main outcomes from the visual inspection were the insights about the subsamples where mergers can be found and the subsamples where contaminated sources cluster.

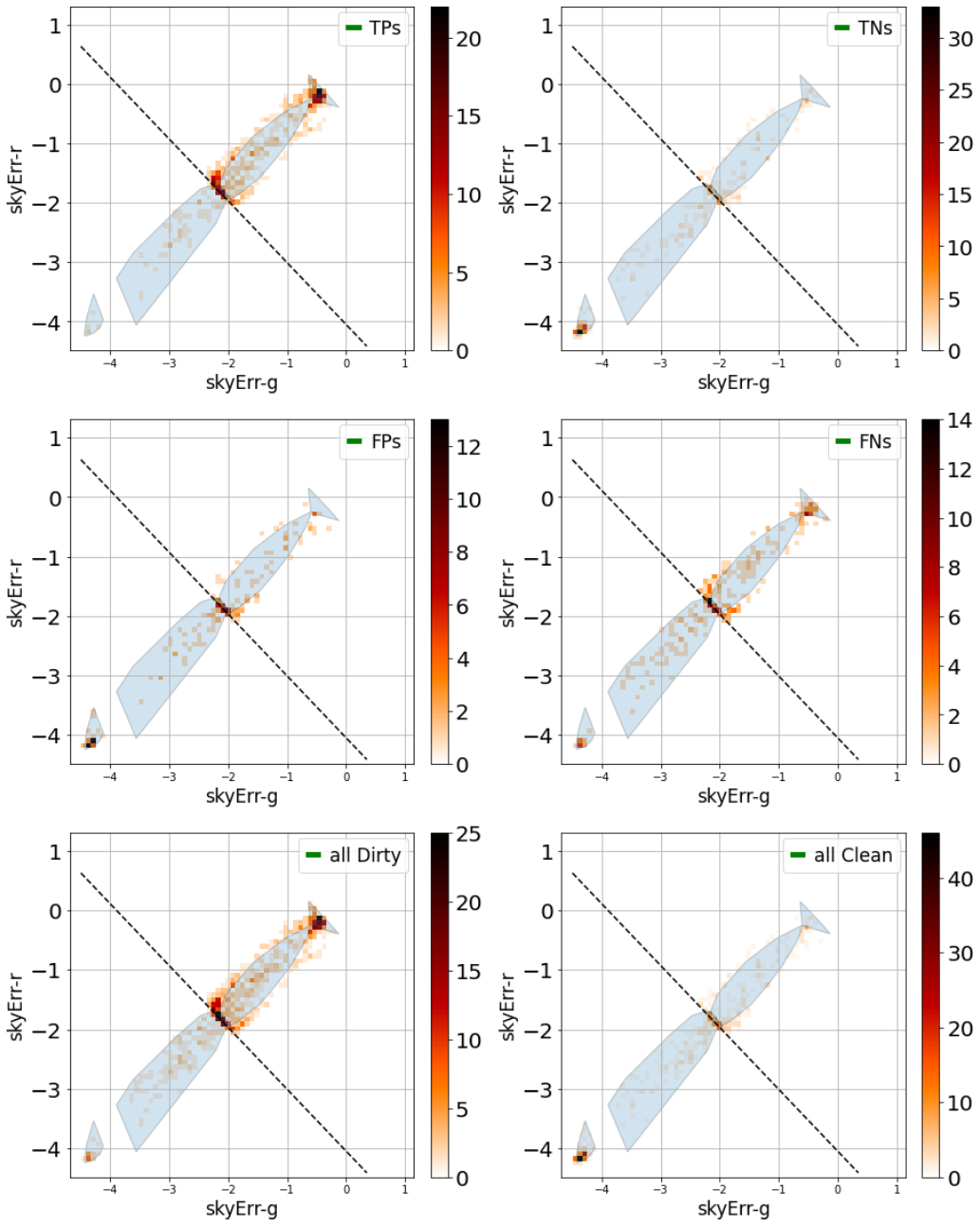


Figure 5.8: These six panels show the @D histograms for the dataset galaxies that have been used as input for the decision tree. The two bottom panels show the distribution of all dirty and clean galaxies in the left and right, respectively. The top left panel show the TPs, the top right the TNs, the central-left one the FPs and the centre right one the FNs. The colour bar goes from 0 galaxies in white, through 1 galaxy in light yellow, getting darker and darker to the maximum of galaxies per panels.

Many galaxies showing minor mergers or coalescing were found hidden in the subsets of Spiral galaxies with $f_m \in (0, 0.2]$ in the Right and Left wings. This was also found for Uncertain galaxies in the Bottom area with $f_m \in (0.2, 0.4]$, and for some sources in the Non-mergers area A that are close to the boundary, as suggested in the central-right panel of clean other-mergers in Fig. 5.6. This implies that the GZ DR1 question tree was not very efficient at differentiating spiral galaxies from irregular galaxies unless those were major merging pairs. Moreover, it also supports that mergers other than major ones can be classified through the sky error diagram.

The first conclusion from the contamination analysis is that, in the form calculated in SDSS DR6, the `skyErr` is sensible to the presence of contamination in a way that it confuses actual mergers. This can be expected because the sky background in SDSS DR6 did not mask sources, but instead calculated the clipped median for all pixels in every 128×128 box. This allows the PSF wings of any source near the target galaxy to have an effect on the background. From the difference of distribution between types of contaminations indicated in Fig. 5.7, it can be seen that indeed the stars induce a larger sky error. They are increasingly frequent towards the Top area and perpendicularly to the boundary.

The second conclusion from the contamination is that the contamination populates the diagram in a relatively uniform distribution, affecting non-mergers in any area, as depicted in Fig. 5.6. This complicates to create a sample of clean mergers in the leaf-shaped Merger area, where they appear more frequently.

Regarding the contamination of merging sources, it seems that the majority of contaminated mergers are affected both by nearby stars and visual pair galaxies. This is because they are galaxies located in more crowded portions of the sky.

Last but not least, the division of subsamples of galaxies depending on the four parameters proved very useful to narrow down the region-dependent behaviour of the sky error diagram. However, as shown in Table 5.1, it has introduced difficulties on extracting conclusions on the underlying distribution from the whole dataset, and therefore on the general capabilities of the methods.

5.2.2 Decision Tree

The preliminary results on the decision tree can provide two main conclusions on the capabilities of the sky background error methodology. One comes from its construction: it took into account only contamination by stellar or galactic sources. As described in Sect. 3.2.3, all sources within twice the Petrosian radius of a target galaxy that was a star or a galaxy not fulfilling the cross-pairs conditions, were considered to be contaminants. The fact that only a 5% of the sources were contaminated by artefacts, very small in comparison with stars and visual pairs, made us decide that cleaning them was less of a priority.

The recall and sensitivity for mergers plus the recall for non-mergers were found to be 71.82 %, 72.43 %, and 67.07 %, respectively. This is more relevant than the relatively low accuracy for both samples, 62.72 % and 64.14 % respectively, because the goal of the decision tree is to find contaminated non-mergers while keeping clean mergers. Moreover, the specificity of 72.43 % for mergers also implies a relatively high amount of clean mergers that were not discarded.

Finally, the `dt` results are quite consistent with the clustering in the distribution of contaminations, impliyng that contaminants can be found consistently. Thus, a further effort in determining them while differentiating them from the real merging features has the potential of creating clean and large catalogues of mergers.

5.3 Conclusions

The extension to the whole GZ DR1 sample of the sky error diagram, built in the g-versus-r-band `skyErr` plane using the training dataset of Sect. 2.1.3, had two main challenges: to make sure that GZ DR1 mergers not included in the training sample can be found, and to limit the non-mergers that intrude in the merger area. Besides, the galaxies in the training set were mainly major merging pairs, so that the extension aimed also to include other types of galaxy mergers, such as minor mergers or coalescing sources.

Because the GZ:DR1 galaxies populate the diagram more widely than the training set, as shown in Fig. 5.1, we wanted to narrow down what areas of the diagram are more likely to be populated by mergers. We used the alpha shape algorithm to delimit an area around the main distribution of mergers and two areas around clusters of non-mergers. Moreover, we also studied the statistics of the diagram when moving from the main merger area in all four directions. This resulted in defining four more areas in the diagram.

We thus separated the GZ DR1 in the seven diagram areas, plus in magnitude and morphologies based on the GZ results. This provided subsets of GZ DR1 defined by four parameters, and we randomly took nine galaxies for each subset, making the subsamples that we visually inspected. We split morphologically those galaxies in three classes: major merging pairs, non-merging galaxies, and a class gathering all other types of merging interactions. We also noted whether there was contamination around the galaxy. The possible contaminants were stars, image artefacts, or galaxies that we considered to not be interacting and that were too far appart to be identified as mergers through cross-pairs, i.e., calculating the distance to the target making use of their spectrometric redshift.

The main result from the visual inspection was that we found contaminated non-mergers all along the diagram. It also resulted that Spiral galaxies with GZ DR1's merger vote fractions f_m of moderate values – around $(0, 0.2]$ – were likely to be minor or coalescing merging galaxies. Because of the contamination, we decided to design a decision tree capable of finding those contaminated non-mergers.

The contaminants were mainly stars or visual pair galaxies near the visually inspected sources. The decision tree took as input all the detections from SDSS DR6 as long as they did not present some detection flags. If one of the detections was located at a certain distance from the target galaxies, then the target galaxies themselves were discarded and considered contaminated.

The performance of the decision tree was the following. It had a sensitivity of 72.43 % for clean merger identification, which indicates the amount of clean mergers that were recovered as clean ones – see the third row of Table 5.2. It also had a recall of 67.07 % of dirty non-mergers, which corresponds to the percentage of dirty non-mergers that the decision tree is capable of finding – see the second row of Table 5.3. This implies that, while the model still looses some of the mergers and also struggles to clean the sample of dirty non-mergers, it provides a well balanced and promising result.

We considered that the implementation of this project had a fundamental complication, the separation of the data by magnitudes and by GZ DR1 morphologies. It was initially done because of the clear dependence of the sky error values on them, that we introduced because they trace very well the variability of the galaxies across the `skyErr` diagram. However, it complicated the generalization. In future work, we would like to separate the more general decision tree from a more specific analysis of the magnitude and merger votes, which would provide insight on what are the best galaxy images that the sky error method could trace.

CHAPTER 6

Mergers in the North Ecliptic Pole

The extension of the sky error method to deeper images requires parametrizing the effect in the sky background of the Low Surface Brightness (LSB) merger features. Using the selection of mergers and non-mergers described in Sect. 2.2.3, we aim to find parameters that characterize mergers when applied to low Signal-to-noise (S/N) pixels. This text summarized my contribution to Pearson et al. (2022) and includes my work on the extension of the sky error method to deeper Subaru/Hyper Suprime-Cam (HSC) images. This chapter is based on a draft of a publication which is now being prepared and should be submitted soon after the submission of the thesis. This was again carried out by myself and the analysis and discussion was developed together with my auxiliary supervisor William J. Pearson and my supervisor prof. Agnieszka Pollo.

6.1 Results

As described in Chapter 3, Sect. 3.3, we visually confirmed or denied the merger classification by the Deep Learning (DL) applied to the HSC images on North Ecliptic Pole (NEP) – see Chapter 2, Sect. 2.2.2. The combined visual inspection was performed between my auxiliary supervisor dr. William Pearson and me, being the majority of the galaxies identified by him. The DL model combined a Convolutional Neural Network (CNN), applied on the *r*-band images, with a NN applied on morphological parameters, calculated from the images using `statmorph` (Rodriguez-Gomez et al., 2019), and it was trained separately for two redshift bins. This resulted in two catalogues of mergers with DL-based galaxy merger candidates and their visual classifications.

The two redshift intervals in which the dataset was split were a low-redshift bin of $z < 0.15$, and high-redshift bin of $0.15 \leq z < 0.30$. The model provided 1 477 merger candidates in the $z < 0.15$ redshift bin, 251 out of which were confirmed by our visual inspection. Regarding $0.15 \leq z < 0.30$ interval, 1 858 of the 8 718 candidates were confirmed.

As a consequence, only 17% and 21% of the merger candidates were confirmed in each bin. There are two possible arguments that can justify this, one is related to the performance of our inspection and the other comes from contaminations in the DL classification.

The visual inspection performances are described in the Appendix B of Pearson et al. (2022). They were tested by generating 100 mock major merging galaxies plus 100 non-merging ones from the Illustris TNG simulation (e.g. Marinacci et al., 2018). The galaxies had a redshift of $z = 0.15$, between the two datasets. It showed that, for both of us, the performance is worse than the DL model. The accuracy from dr. Pearson was 0.620, and my

accuracy was 0.630, while the model showed of 0.884 and 0.850 in the low and high-redshift intervals.

However, during the visual inspection, we observed multiple sources with properties that could confuse the model. Those were two blending galaxies that resembled a double nucleus, or visual pair galaxies with no apparent distortion and different redshift. It can be understood that this cases can be ambiguous for the model, and thus are natural sources of contamination.

6.1.1 Sky Background modifications

The initial result of this latest project was obtained by comparing the calibrated frames obtained from the two variations of the HSC data reduction described in Sect. 3.4.2.1. They differ in the treatment of the sky background. Figures 6.1 and 6.2 show and compare the results of the image processing for one calibrated Coadd frame and for one cutout around a galaxy, respectively. The frames named "Calibrated with sky" in the bottom-left, are those where the sky background created from the dithering between exposures was not modified, and the "Calibrated frames without sky" are those where the three dithering Header Data Units (HDUs) were set to zero.

The difference between the two coadds shows a defined structure. Because this difference is obtained by taking the image with normal sky and subtracting the image calibrated without sky from it, this means that negative pixels – shown in grey in the difference image of Fig. 6.1 – correspond to higher values in the image without the dithering-based background. Conversely, positive pixels – shown in white – have higher values when the background is treated normally. As a consequence, a negative value implies that a static background feature has been corrected in the frame with normal sky, but it remains in the image when the dither-based background is not subtracted.

The area populated by brighter stars in the left of the frame shows mainly pixels smaller than 0 in the difference frame. This means that the dithering reduced the glow surrounding the stars, and getting rid of the background treatment allows the bright haloes to stay in the image. There is also a substructure, shaped as a diagonal line in the region between the x-axis around [0, 500] and y-axis $\in$ [2500, 3100], that also shows negative difference. The origin of this structure does not seem clear to us.

The top area shows positive values instead. It is less populated by bright stars, and it clearly looks dimmer in the frame on the right of Fig. 6.1. This under-subtraction can be due to the lower density of sources, or due to the g-band filter itself. Overall, these differences are small but with a high spatial variability.

The clear substructure in the large coadd calibrated frame is nonetheless not so clear in the galaxy cutout of Fig. 6.2. The difference frame shows values one order of magnitude smaller than those in the pixels of each cutout, so that the contrast can barely be observed. There are flat and noise regions distributed across the Difference frame, although there is no clear correspondence between them and the sources of the image. While this supports the argument that the dithering-based background does not affect low Signal-to-noise features around galaxies, it does not give insights on the origin of the small size substructures.

6.1.2 Parameters

The parameters characterizing the low-surface brightness pixels around the galaxies of our catalogue are shown as histograms in Figs. 6.3 and 6.4. All were calculated through a clipped median background estimation within an aperture of 2.5 times the galaxies' effective radius (r_{eff}). The measurement used with a clipping factor of 2.7 and n = 10 iterations. Figure 6.3 shows the parameters with images reduced using the normal sky

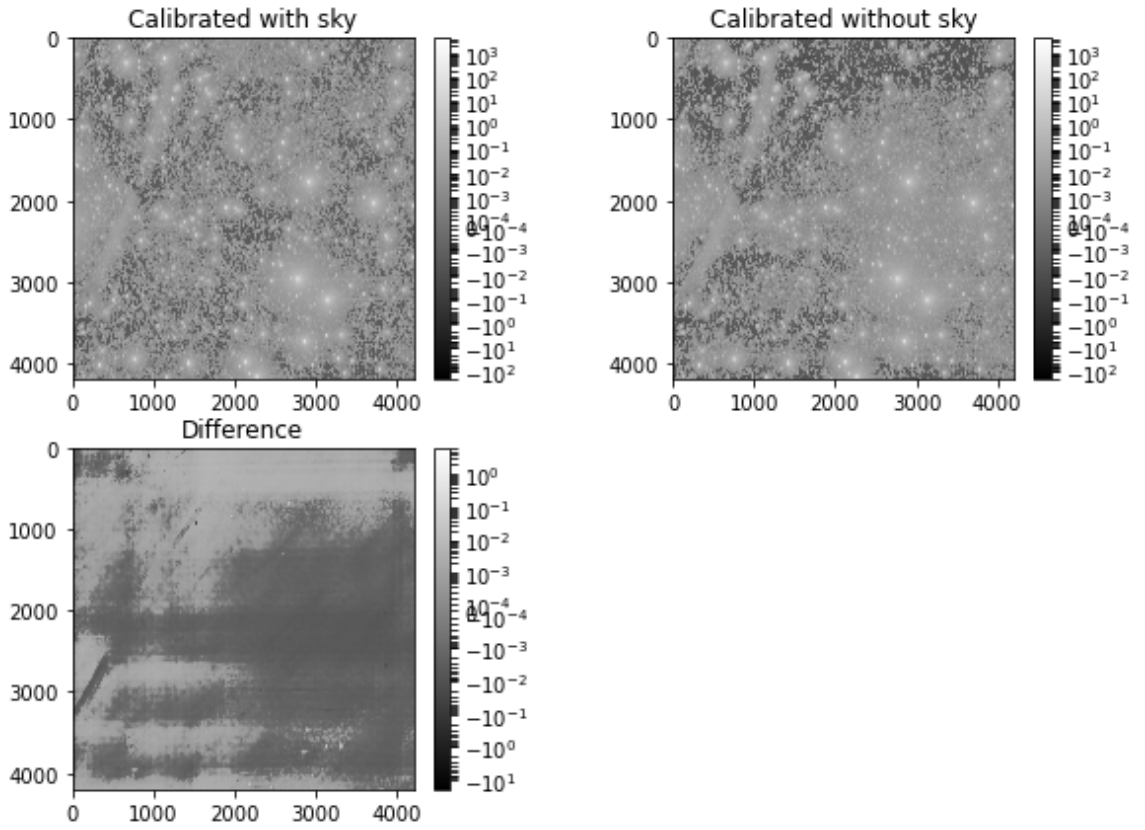


Figure 6.1: Comparison between the same Coadd Calibrated frame from the two different data reductions applied. The bottom left and bottom right images are the calibrated frames resulting from the pipeline runs when the last background HDUs were kept or when they were set into a 0 value, respectively. These are $4\,096 \times 4\,096$ images, displayed using a symmetric logarithmic grey scale. The Difference image in the bottom row is the difference between both frames, created as the image on the left minus the images on the right. According to the colour bar of the Difference image, positive numbers are in bright colours and negative in grey.

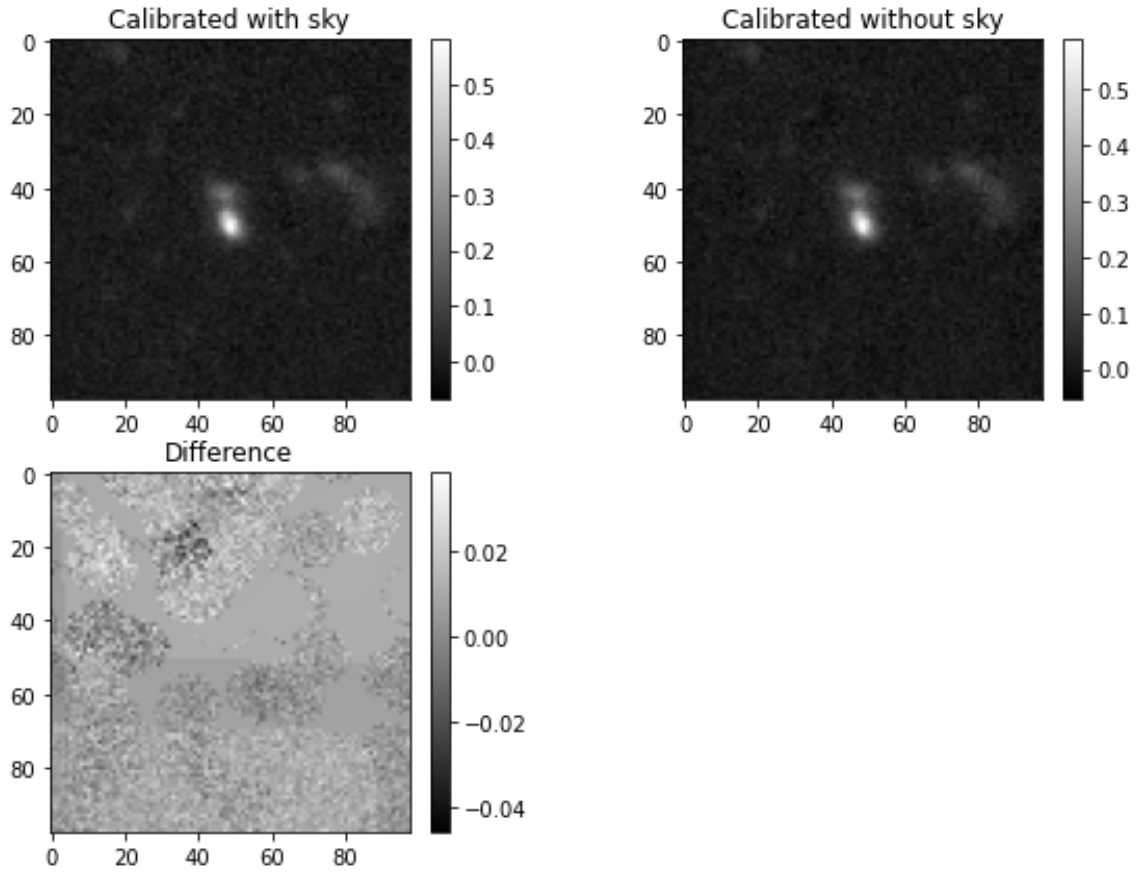


Figure 6.2: Comparison between the two cutouts of one of the galaxies of the GZ:CD-based catalogue. The two bottom cutouts were obtained with the same pipeline options as in 6.1, and the bottom panel is also the difference between the two. The colour scale in this case is linear.

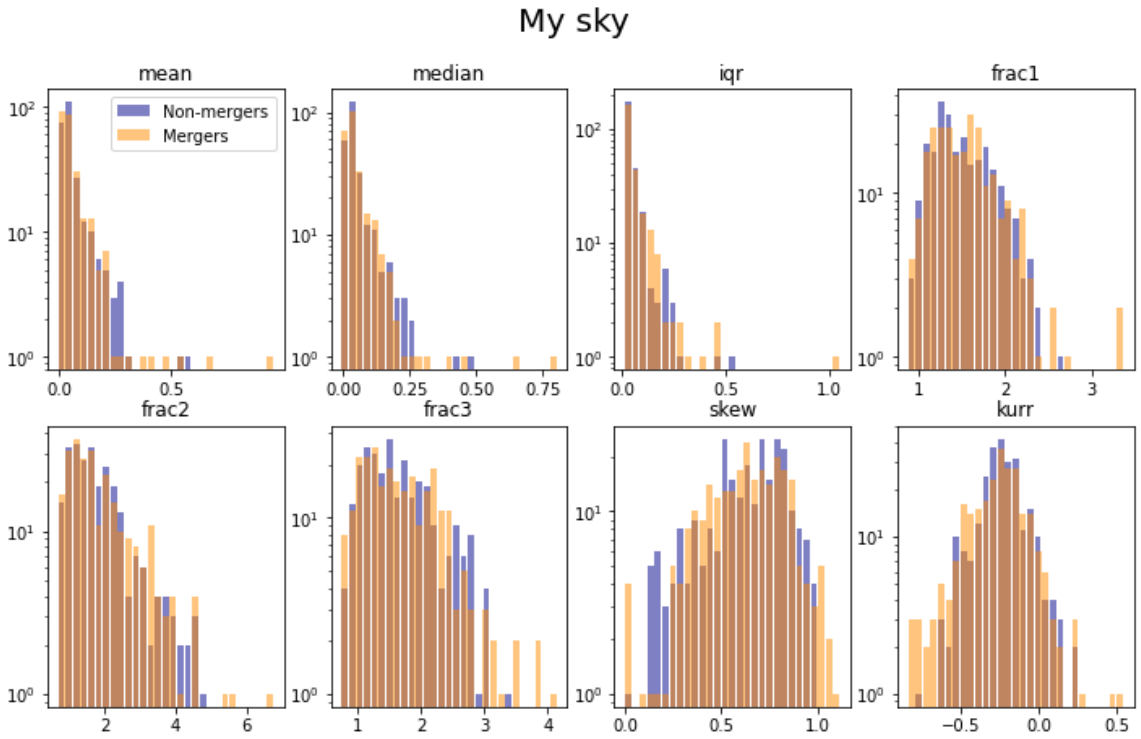


Figure 6.3: Eight panels showing the histograms of the eight parameters applied to the LSB pixel distribution around the GZ:CD HSC-NEP matched galaxies. The orange histogram bins correspond to mergers and the blue to non-mergers. The name of the parameters is indicated on top of each panel, and the bin heights are in logarithmic scale. The cutouts used for the analysis were obtained performing the HSC data reduction without any modifications in the *g*-band images.

treatment, and Fig. 6.4 corresponds to the pipeline without dithering-based background subtraction.

Firstly, there is quite small difference between Fig. 6.3 and Fig. 6.4. This is consistent with the results from Sect. 6.1.1, where no substructures at the level of Low-Surface-Brightness (LSB) are evident.

A panel by panel examination shows the main difference in the distribution of mergers and non-mergers to be in the lowermost tails of the skewness and kurtosis parameters. The higher end of the Mean, Median, and the IQR panels also hint some difference between classes. The Fractions 2 and 3 have slight differences in the distribution shapes. Overall, this is the same for both treatments of the sky background.

Because the parameter distributions indicate that the LSB pixels around mergers and non-mergers are not statistically the same, we applied a Neighbourhood Components Analysis dimensionality reduction to the eight-parameter set of Fig. 6.3.

6.1.3 Dimensionality reduction

NCA creates an embedding using the classification labels of the data, aiming to create something similar to the SDSS `skyErr` representation in Fig. 4.11. Figure 6.5 shows the result of applying NCA to the parameters plotted in 6.3. The mergers and non-mergers shown in the plot overlap in the more crowded area. However, there is some region of the embedding where mergers are more abundant than non-mergers. This is marked by the top-right red box.

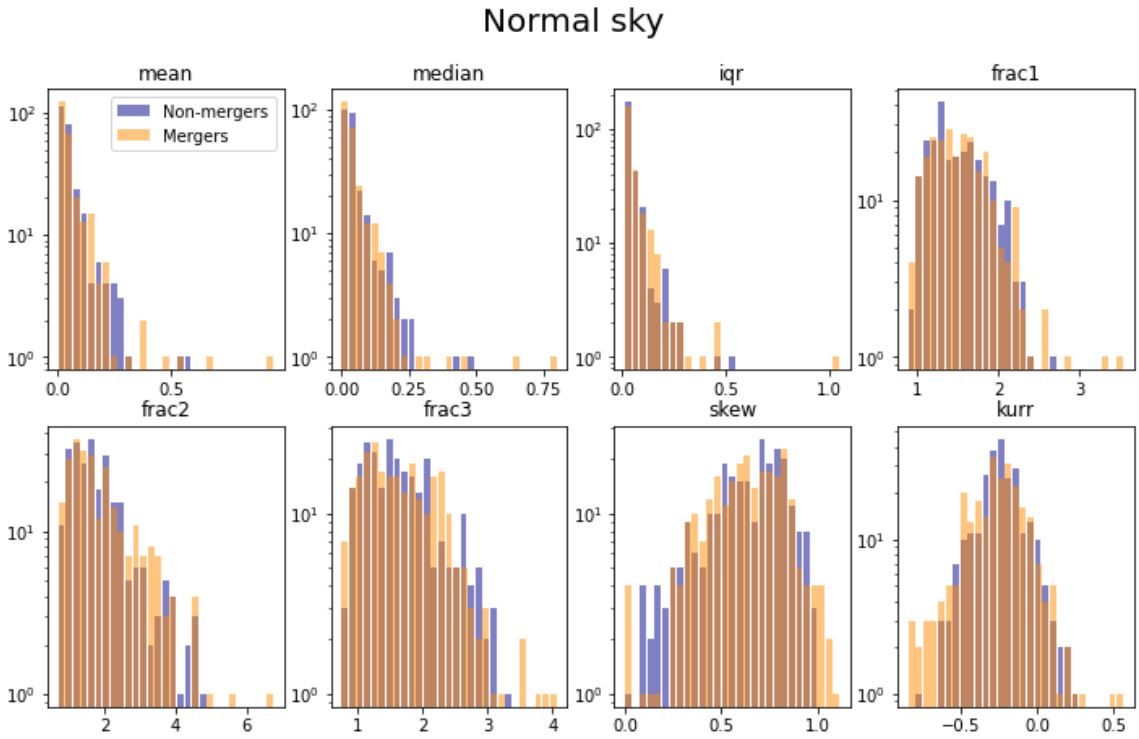


Figure 6.4: The histograms and panels are analogous to Fig. 6.3, except that the cutouts used were obtained from the HSC data reduction where the dithering-based sky background was not subtracted.

While there is no clear separation, as with the sky background error, the marked red box indicates the presence of a region in the parameter space with a majority of mergers. In this case it contains 49 mergers 30 and non-mergers, which means it is populated by ∼62 % of mergers.

6.2 Discussion

In the Results section of this chapter, many galaxy merger candidates were considered to have been incorrectly identified by the Deep Learning (DL) model. We observed some patterns during the visual inspection that were common for some of the rejected merger candidates. Double nucleus galaxies that looked more like a chance blend, or nearby galaxy pair without tidal distortion and at different redshifts.

One test was to apply an occlusion experiment on the galaxy images (e.g. Zeiler and Fergus, 2013). It consists on generating a map of the importance of the image pixels by setting to zero a kernel of pixels. The image with the occluded kernel is used as input of the model and the output is stored. The kernel then moves around the image, providing an average of the classification outputs per pixel. In this project, it was performed by moving a 16×16 one step at a time along the pixel grid, implying that the value in each pixel comes from the average of up to 16^2 results.

The occlusion was done for eight mergers and eight non-mergers in each redshift bin, all identified as mergers by the model and confirmed/rejected visually. It showed that the majority of non-mergers wrongly identified as mergers had a second galaxy nearby in the cutout. These companion galaxies showed redshifts not compatible with merging processes.

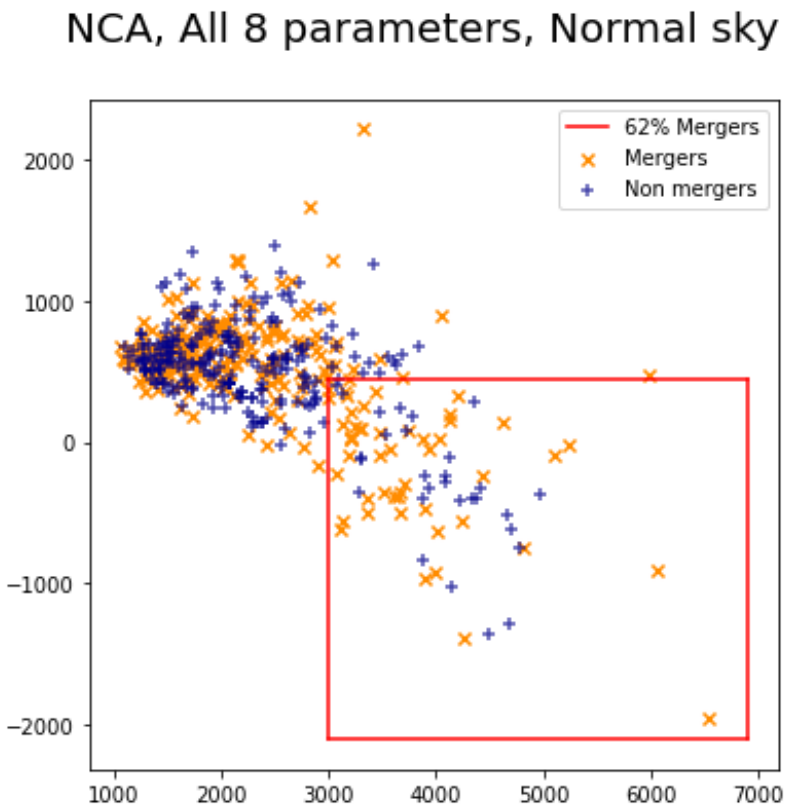


Figure 6.5: Scatter plot showing the result of applying an NCA dimensionality reduction to the parameters shown in Fig. 6.3. The dark orange crosses are merging galaxies, and the dark blue pluses non-merging ones. The red lines delimit a region in the embedding populated in a ∼62% by merging galaxies.

However, the redshift was not an input of the DL model, and therefore this contamination by visual pairs could be expected.

It is worth to note that the influence of the LSB surroundings in the merger classification was hinted by how obscuring the primary galaxy made some sources more likely to represent a merger. As stated in the text: "Hiding of the central source may make fainter structures around the galaxy more apparent and hence easier to identify as a merger, but this is speculation." (Pearson et al., 2022). It is quite interesting how this intuition from my auxiliary supervisor become a consistent argument along all my PhD.

The analysis of the relative importance between the image-based CNN and the morphological parameter-based NN in Pearson et al. (2022), concluded that the parameters contributed more than the images to the correct classifications. Nonetheless, the analysis of the visually rejected non-mergers did not lead to insights about contamination sources.

Overall, this initial experience during my PhD not only allowed me to learn the characteristic features of galaxy mergers, but also made us aware of the contamination in ML-based merger classifications. This discovery played a big role in motivating the analysis in Chapter 5.

6.2.1 Sky Background modifications: effect on merging features

The comparison of the two g-band data reductions shows that the dithering-based background subtraction is independent of the LSB features around galaxies. This can be interpreted by seeing how the LSB-based parameters in Figs. 6.3 and 6.4 are quite similar. The substructures in the Difference frame of Fig 6.2 are completely independent to the locations of the sources, supporting the independence of the LSB features from the background

The HSC commissioning team's motivation for introducing this sky background procedure for the HSC photometric pipeline was, in the first place, to avoid the over-subtraction of bright haloes. It is clearly depicted in the Fig. 5 of Aihara et al. (2019), that compares the over-subtraction in the previous pipeline of the bright halo around a nearby galaxy with the non-over-subtracted updated version. We find another effect of this new subtraction in Fig. 6.1, where the image obtained from the data reduction without the dithering-based background keeps the glow around the stars. Thus, the `hscPipe` v6.7 background subtraction is able to reduce the residual brightness from the extended PSF of the stars in the image. However, this corresponds to brighter sources than the majority of galaxies of our catalogue. Therefore, this is to our knowledge one of the first studies of the effect of the HSC sky background subtraction on LSB features around galaxy mergers in the intermediate galaxy size regime covered in this work.

The contrast between the two coadd images of Fig. 6.1 in the area populated by stars implies that the dithering is able to reduce the residual halo around stars. It has also been seen in Watkins et al. (2024), which simulated images applying the extended PSF obtained experimentally in Montes et al. (2021) for the HSC Subaru Strategic Program (HSC-SSP). They tested, among other background subtraction methods, a dithering strategy that included a smoothing of the resulting images. They found how the dithering sky subtraction reduced the halo glow of the images with extended PSF.

The research in Watkins et al. (2024) also found that the dithering pattern can infuse some structures, introduce noise back to the image, or produce over-subtraction if some sources passed unmasked during the detection. The substructures they found in the dithered frames were produced by the smoothing, so we did not find them in our images. The substructures in the Difference frame of Fig. 6.1 might have also been induced by the dithering pattern: they resemble a blurring around the position of the stars. In the Difference frame of Fig. 6.2, that alternates flat and noisy patches, the noise added back to the images seems also apparent.

The use of dithering for reducing the loss of LSB tidal features around galaxies has been explored and confirmed in previous works. The study in Trujillo and Fliri (2016) showed how it is possible to improve the lower limit in surface brightness by designing an optimized observation strategy. In their case, they used the 10-metre Gran Telescopio de Canarias (GTC) telescope to improve the imaging of LSB structures around the UGC00180 galaxy. They combined the dithering of the camera with rotations around its edge, and calculated the background for each exposure after masking the detections. They not only obtained high details on the stellar halo around UGC00180, but also managed to show the tidal stripped material around the two mergers, of redshifts z = 0.175 and z = 0.287, that were found in the nearby sky – see Fig. 5 in Trujillo and Fliri (2016). Such study managed to improve the surface brightness limit down to 31.5 mag/arcsec^2.

6.2.2 Dimensionality reduction on the LSB parameters

The main interpretation of the Neighbourhood Components Analysis (NCA) embedding obtained in Fig. 6.5 is that this eight-dimensional parameter space generated through the LSB pixel distribution indeed has the potential of providing a clean sample of galaxy mergers. The fact the embedding is still contaminated by non-mergers is an indication that this work is only on its preliminary stages. There are however some details that can be discussed.

First, the parameters we have shown were obtained applying an aperture with radius equal to 2.5 the galaxies effective radius and a clipped median with clipping factor of $2.7\times\sigma$. These were selected by a grid search, testing multiple values around them. Deviations from these parameters showed more similar distributions between mergers and non-mergers. Second, the dimensionality reduction we chose to show was the NCA because it created the embedding using the source labels. Dimensionality reductions such as PCA or tSNE did not create an area where mergers were more abundant, as the one within the red box of Fig. 6.5.

Regarding the parameter distributions of Figs. 6.3 and 6.4, the mergers and non-mergers overlap in the more dense areas. It is in the wings where there is more difference between distributions. This is somewhat artificially enhanced by the logarithmic scale in the y-axis, which makes smaller difference in low-number bins to look larger than big differences in high-number bins. However, the NCA area demonstrates that the distribution difference indeed exist.

There is still one main difference between these parameter representations and the original sky background error method: the `skyErr` decision boundary in Fig. 4.11 combines the g and r optical bands. Next iterations of this work will include the r band observations, and we also intend to add other HSC bands. This has been delayed and could not be included in this manuscript because we wanted to extend the dataset to all the H20-NEP images for both bands, which were carried out by a different HSC pipeline than `hscPipe` v6.7, and the new installation has not been implemented yet.

Finally, the eight parameters that we calculated were chosen to quantify the shape of the LSB-pixel histogram. Initially, we found that these histograms were different for merging and non-merging galaxies. The fact that our parameters provide small differences between sources is very encouraging to, in the future, test more complicated parametrization of the LSB distributions.

6.3 Conclusion

The main conclusion of this project was the clear presence of contamination in Machine Learning (ML) based classifications, among the galaxies of the Subaru/Hyper Suprime-Cam

(HSC) survey in the North Ecliptic Pole (NEP). We observed multiple sources that were not mergers but were mistaken by the model. As a result, the output catalogue published in Pearson et al. (2022) included the merger candidates obtained from ML model outputs, the visual inspection confirmations. Out of the 34 264 merger candidates, 10 195 were merger candidates, and 2 109 were visually confirmed: 251 out of 1 477 for galaxies with redshift z < 0.15; 1 858 out of 8 718 for those in the $0.15 \leq$ z <0.30 interval.

In this final work, we intended to combine all the previous work and define a `skyErr`-like parameter in HSC-NEP images taken by the AKARI-NEP collaboration, also used in Pearson et al. (2022). We took advantage of the Galaxy Zoo: Cosmic Dawn! (GZ:CD) morphological classifications, and built a catalogue of 256 mergers and 256 non-mergers, classified in GZ:CD and observed in the HSC-NEP g band images.

Because Low-Surface Brightness (LSB) features in astronomical images depend strongly on the sky background subtraction strategy, and the SDSS DR6 sky error was calculated during the sky subtraction itself, we modified the sky background in the HSC-NEP images. We run the data reduction twice, switching off and on the subtraction of sky background obtained from dithering, which we considered to be the most likely to influence the LSB surroundings. As a result, we found indications that the dithering-based background does not affect the LSB features. This was shown studying the relative difference between data reductions methods in the image themselves and in the parameter space created for our `skyErr`-like methodology, described in the following paragraph.

We defined a method to extract the LSB pixels around the observed sources, and generated parameters that describe the LSB distribution. We selected parameters that trace the distribution properties: the mean, the median, the interquartile range, three functions we defined by comparing quartile sections, the skewness and the kurtosis – see Sect. 3.4.1. The resulting parameter space showed some statistical differences between mergers and non-mergers. We performed a Neighbourhood Components Analysis (NCA) embedding shown in Fig. 6.5. The statistical differences appear enhanced in the red box of the bottom right side, where a 62 % of the galaxies are merging sources in contrast with the more homogeneous central distribution.

Further research will attempt to increase the dataset, the number of parameters, and the photometric bands reduced. For increasing the dataset, we will implement a new version of the HSC photometric pipeline, the `hscPipe` v7.9.1, which will allow the data reduction of the H20-NEP images, from which the GZ:CD catalogue was created. Once this is done, we will reduce the photometric bands other than g. Finally, better statistics and multi-band parameters will provide new insights in the LSB pixel distribution, making it possible to introduce new parameters that trace better the presence of LSB features and stripped material.

Last but not least, analysing the sky background-based merger identification with the increased image depth that HSC has compared to SDSS had an additional objective. We want to explore how this method could be applied to the Large Survey of Space and Time (LSST; Ivezić et al., 2019). This survey will make use of the Vera Rubin Observatory, observing the South Pole sky during 10 years, reaching even deeper imaging than HSC. The data commissioning team of those three surveys, SDSS, HSC, and LSST, has been the same. It has evolved and become bigger over the years, shifting their focus from one survey to the next. In fact, HSC was treated as a precursor survey to LSST, and the infrastructure of the LSST pipeline is based on the HSC one. Thus, the extension of our merger identification for LSST, and the possibility of including its results in the LSST data reduction pipeline, are the underlying goal of these four years of PhD. While the work still needs to be completed, the main steps have already been walked, and the results are just appearing on the horizon.

CHAPTER 7

Summary

This thesis has shed light in the current methodologies to find galaxy mergers in large optical surveys. There are two main areas in which we have advanced the field: the possibility of using photometry for finding mergers, and the assessment of contamination by non-interacting sources. We did this in the works presented through the Chapters 4, 5, and 6.

We demonstrated in Chapter 4 that photometric parameters can be used for finding merging sources. We arrived to this conclusion by the analysis of Sloan Digital Sky Survey Data Release 6 (SDSS DR6) galaxies through a Neural Network (NN). We first created a class-balanced catalogue of mergers and non-mergers to train the NN. For that, we selected mergers from Galaxy Zoo Data Release 1 (GZ DR1) that were confirmed in Darg et al. (2010a,b). This training catalogue had 2 930 mergers and 2 930 non-mergers for training and validations, and 250 mergers and 250 non-mergers for test.

We applied multiple sets of photometric parameters as inputs to the NN. We started with combining magnitude measurements in the five SDSS photometric bands, and in the end found that the error in the sky background measurement was capable of providing an accuracy of 92.64 ± 0.15 % in validation, and of 92.36 ± 0.21 % in the test set. Through attempting to understand the NN results, we arrived to the conclusion that the SDSS DR6 sky background error can be plotted to provide a decision boundary separating mergers and non-mergers. This can be done combining the SDSS r, g, and i bands, and as shown in Fig. 4.11, the g-versus-r plane gives a 91.59 % of accuracy.

The sky background error prowess to find mergers was interpreted to come from the influence of Low Surface-Brightness features around the mergers in the images. Thus, the stripped material, tidal arms, and other interaction-induced features, would have an imprint in the images that the SDSS DR6 sky error was capable of capturing. We made an initial evaluation of the method's dependence on image depth, and showed that it is capable of identifying other mergers than major merging pairs, which were the main type of mergers in Darg et al. (2010a,b) and thus in our training dataset.

The potential for finding mergers that the SDSS DR6 sky error showed encouraged the work presented in Chapter 5. The initial aim was to extend the decision boundary to all galaxies classified in GZ DR1. For narrowing down its applicability range, we visually inspected galaxies pre-selected in subsamples defined by magnitude, GZ DR1 morphological results, and location within the diagram. In the visual inspection, we classified the galaxies between merging major pairs, non-mergers, and other types of

mergers. Besides, we noted if they presented nearby potentially contaminating sources such as stars, visual pair galaxies with no physical relation, and image artefacts.

The main results of the visual inspection were two. The first was that merging galaxies that are not major pairs, such as minor mergers or coalescing mergers, can be found in the diagram but with lower sky error than the mergers. The second was that non-merging galaxies with contamination around could be found all around the diagram, an issue which we decided that needed to be addressed.

In order to address the contamination, we built a decision tree. The goal of this decision tree is to point out the contaminated sources, avoiding them to be considered as mergers. This decision tree discarded the target galaxies if they had an SDSS DR6 source detection at a certain distance. Moreover, if the merging galaxies could be identified by cross-pairs method, they were not discarded even in the presence of contamination. The tree resulted to discard a 67.07% of the contaminated non-mergers from the visual classification, while keeping 72.43% of clean mergers. Thus, with this initial results, we showed it is possible to obtain a clean sample of mergers from the sky error boundary.

We also worked with deeper imaging from the Subaru/Hyper Suprime-Cam (HSC) deep field images in the North Ecliptic Pole (NEP). I contributed to the catalogue of mergers in Pearson et al. (2022), where we confirmed visually mergers from the Machine Learning (ML) classification method built by dr. William J. Pearson. Some of the merger candidates identified by the ML-based method were contaminated non-mergers that we discarded.

Furthermore, we proved in the HSC-NEP images that it is possible to extend the methodology to the deeper images than SDSS in Chapter 6. We selected mergers and non-mergers from the Galaxy Zoo: Cosmic Dawn! (GZ CD) program. We analysed the low Signal-to-noise (S/N) pixels of galaxies observed at the g-band. For each source, we defined an aperture where we measured the background sky level, which we used to identify the low S/N pixels. By defining multiple parameters on the low (S/N) pixel distribution and applying dimensionality reduction, we found a preliminary set of mergers can be obtained, as shown in Fig. 6.5.

Moreover, we carried out our own data reduction of the g-band images. The SDSS DR6 sky error was itself calculated during the image calibration. Besides, some sky background models have been shown to have detrimental effect on LSB features. Among the sky background models made in the HSC pipeline we used, we only considered the potential effect background made out of the dithering of the images. We performed the calibration with this background switched on and off. The analysis of the difference between images and galaxies obtained with and without the background showed that the dithering-based background model does not affect the LSB structures around the galaxies of our dataset. In fact, the parameters we calculated for the low S/N pixel distribution changed only slightly.

Future work will address first the finalization of the HSC depth extension by calibrating a larger set of images from the H20 survey in the r and g bands, and exploring others bands. We expect to improve the identification due to the larger dataset, the inclusion of more bands, and the optimization of the parameters used to define the low S/N pixel distribution. Besides, further iterations will allow us to understand better the low S/N histogram and define extra parameters that can be applied to it. The HSC extension will allow applying all these discoveries to a new merger identification for the upcoming Large Survey of Space and Time (LSST), which will be carried by the Vera Rubin Observatory. I and my two supervisors are part of the LSST collaboration, and the success of the upcoming results will lead to adapting the method in HSC to LSST.

Other future work will be to improve our understanding in the range of application of the sky error methods. We found that it depends on the image depth of the sources, which itself is subjected to the brightness, size, and distance of the sources. We also have found that multiple types of merging sources have different imprints on the sky background.

Thus, further work on narrowing down the properties of mergers that can be optimally found will also be arranged in the future. The decision tree presented in the work and other techniques for cleaning the catalogues will also be addressed. We will quantify how the contamination effect depends on the magnitude of the contaminants, and will to add a branch of the tree that finds the contamination by image artefacts, using the artefact masking.

This thesis has shown how I have built methods for finding galaxy mergers that can be faster and more accurate than previous methods. Faster, because the sky error diagram is more efficient than visual inspection or NNs. A sky error parameter obtained directly from the data reduction of the survey would reduce to a minimum the effort put by any researcher with access to the image reduction outputs. It is also an accurate method because we have taken into account possible contaminations and addressed ways to clean them.

APPENDIX A

Reproducing the SDSS fibre magnitudes and errors

This appendix was also included as part of Suelves et al. (2023). In order to understand what information the fibre magnitude errors enclose, we attempted to reproduce them using the information in the SDSS DR6 documentation[1]. The documentation indicates that it is calculated as the aperture photometry inside a circle of 3 arc-seconds in diameter, the same angular size as the fibre, after the image is convolved with a 2 arc-second seeing to resemble what the fibre actually sees[2]. Therefore, by retrieving the correct catalogues and photometric parameters from the SDSS repository, one could reproduce the fibre errors and examine the properties that make them so relevant for the NN. This appendix is limited to showing how the uncalibrated fibre counts and count errors – found in the file fpObjc – relate to the fibre magnitude and errors, together with all the required calibration parameters.

Two main sets of information are required to reproduce the fibre magnitudes: first, the equations that connect the observational measurements with the magnitudes, and second, the files where these measurements are found. Table A.1 shows the files, and the equations – as found in the documentation website – are the following:

Magnitude

$$m = -\frac{2.5}{\ln(10)}\left[\sinh^{-1}\left(\frac{f}{f_0 2b}\right) + \ln(b)\right] \ , \tag{A.1}$$

[1] http://classic.sdss.org/dr6/algorithms/fluxcal.html
[2] http://classic.sdss.org/dr6/algorithms/photometry.html#mag_fiber

Table A.1: Files employed to recreate the aperture photometry that leads to the fibre magnitude data.

Name	Type	Contents
kfold_fibre	Input for NN	Fibre magnitudes and errors
fpObjc	Uncalibrated Catalogue, i.e. before converting counts to fluxes	Fibre raw `counts` and `counts errors` plus sky background (`sky`) and error (`skyErr`), interpolated to the source's centre
drField	Field calibration data	Calibration Parameters (per band) (e.g. `gain`, `airmass`, `zeropoints`)

Magnitude error

$$\sigma_m = \frac{2.5}{\ln 10} \frac{\sigma_{\texttt{counts}}}{\text{exposure time}} \frac{1}{f_0} \frac{1}{\sqrt{4b^2 + \left(\frac{f}{f_0}\right)^2}} , \tag{A.2}$$

Counts error

$$\sigma_{\texttt{counts}} = \left[\frac{\texttt{counts} + (\texttt{sky} \cdot N_{\text{pixels}})}{\text{gain}} + N_{\text{pixels}} \cdot (\texttt{dark variance} + \texttt{skyErr}) \right]^{1/2} . \tag{A.3}$$

Here, f is the pixel units in `counts` divided by the exposure time; f_0 is the zeropoint, that is, the flux of an object with zero magnitude, given the atmospheric conditions and the system's instrumentation; b is the softening parameter that indicates the flux level at which linear behaviour of m sets in; `sky` is the sky background estimation and `skyErr` its error, both interpolated on the source centroid; gain is the telescope's CCD's gain; N_{pixels} is the size in pixels of the aperture used, and `dark variance` is the dark current's variance calibrated for the given frame. It should be noted here that $\texttt{sky} \cdot N_{\text{pixels}}$ is the sky counts summed over the same area as the object counts, as indicated in the documentation count error[3].

These initial equations did not succeed in reproducing the fibre errors, and some modifications were found to be necessary. In order to justify the modification of these relations, we compared the fibre magnitudes, magnitude errors, and count errors obtained from the original and from our modified formulae. This study was made for ten arbitrary galaxies of the dataset in each of the five pass bands.

A.1 Magnitude error formula

First, to confirm that the fibre counts (`counts`) provided the fibre magnitudes, we took the fibre counts given in the fpObjc catalogue and applied Eq. A.1. Those fibre counts were the SDSS's counts extracted from the final seeing-convolved frame. Figure A.1 gives the five-band panels, where the y-axis shows the fibre magnitude and the x-axis shows the magnitude resulting from the fibre counts. The parameters f_0 and b were extracted from the field calibration dataset fpC. The straight black lines indicate the linear fits of the scatter plots, with the fit parameters and their errors in the legend. This linear fit confirms that Eq. A.1 does successfully relate counts and magnitudes.

Second, we applied Eq. A.2 to the fibre count errors in the fpObjc file. However, the fit showed a slope corresponding to the exposure time of the frames, as shown in the slopes of the five panels in Fig. A.2. Therefore, we modified the relation to Eq. A.4, and Fig. A.3 is the resulting fibre magnitude error, which lacks the exposure time slope and shows a one-to-one relation that confirms the presence of either a typing error or an inconsistent definition.

New magnitude error

$$\sigma_m^{\text{new}} = \frac{2.5}{\ln 10} \frac{\sigma_{\texttt{counts}}}{f_0} \frac{1}{\sqrt{4b^2 + \left(\frac{f}{f_0}\right)^2}} , \tag{A.4}$$

in comparison with Eq. A.2, the magnitude error is not divided by the exposure time.

[3] http://classic.sdss.org/dr6/algorithms/fluxcal.html#counterr

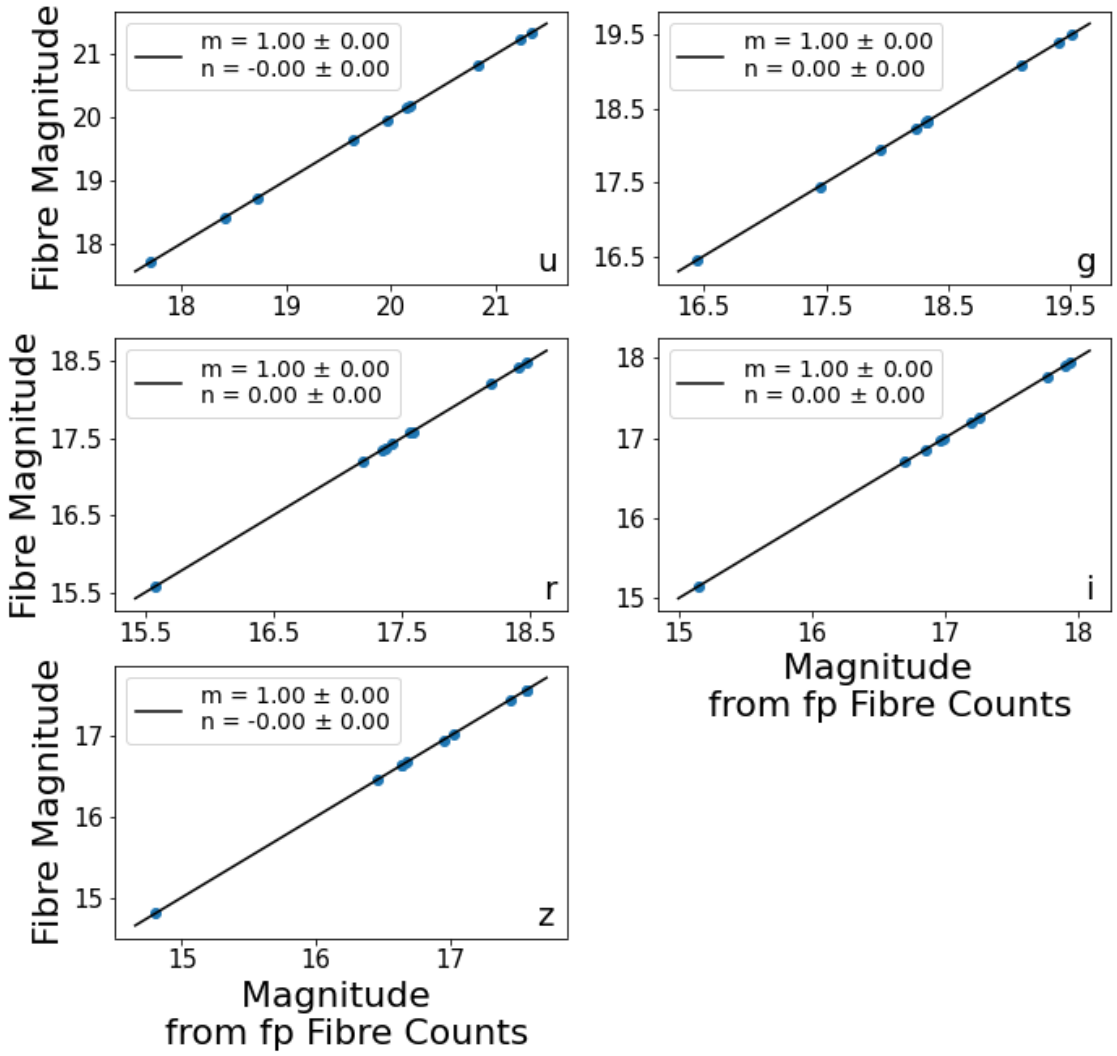


Figure A.1: Linear regression between the fibre magnitude extracted from the CasJobs portal (y-axis) compared to the magnitude calculated using Eq. A.1 from the fibre counts (`counts`) from the fpObjc catalogue (x-axis). The fit corresponds to ten galaxies pre-selected from our training dataset, and is done for all five SDSS bands, u, g, r, i, and z, shown in the five panels.

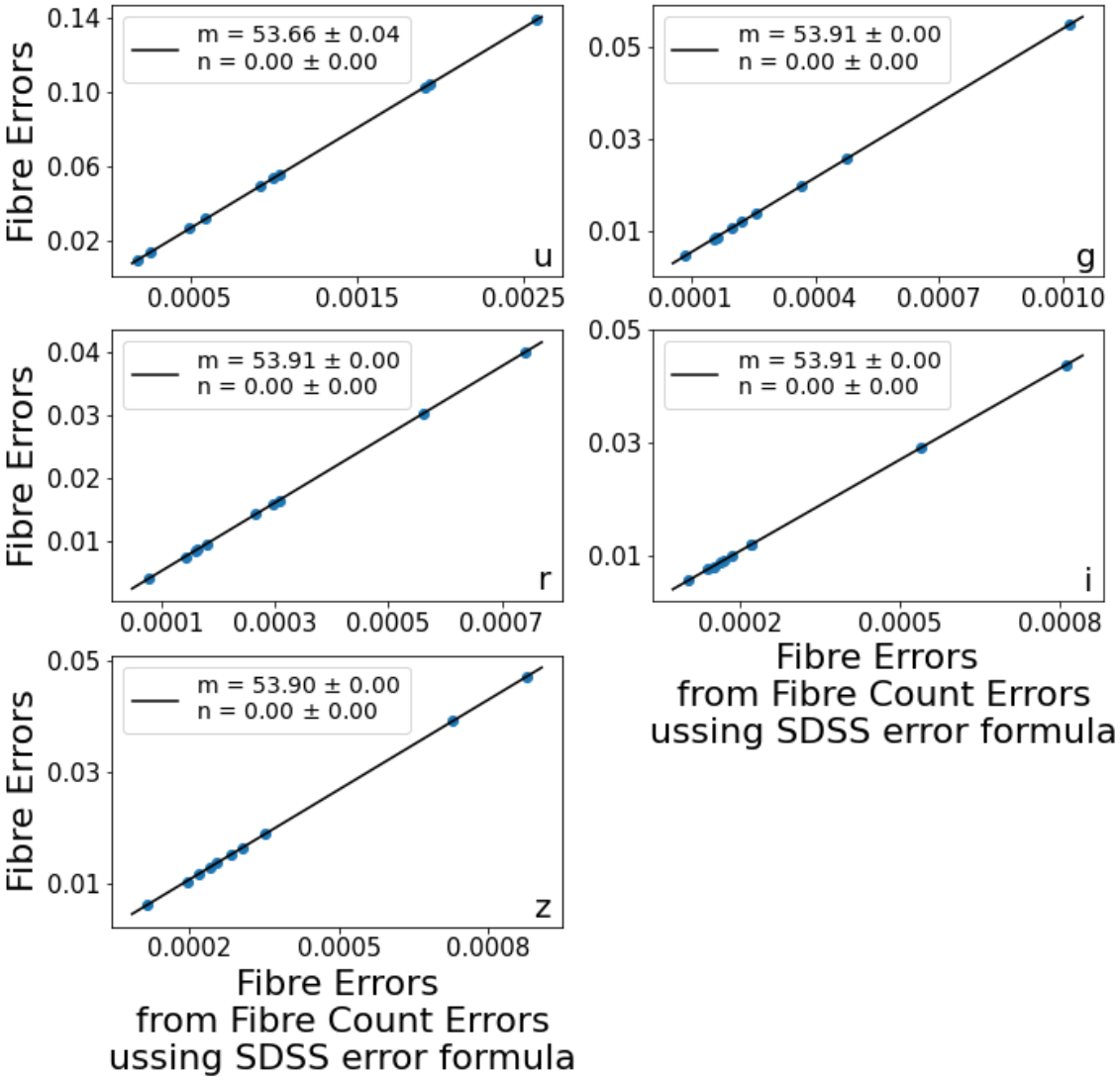


Figure A.2: Linear regression, for the same galaxies and bands as in Fig. A.1, between the fibre magnitude errors extracted from the CasJobs portal (y-axis) compared to the fibre magnitude errors calculated using Eq. A.2 from the fibre count errors from the fpObjc catalogue (x-axis).

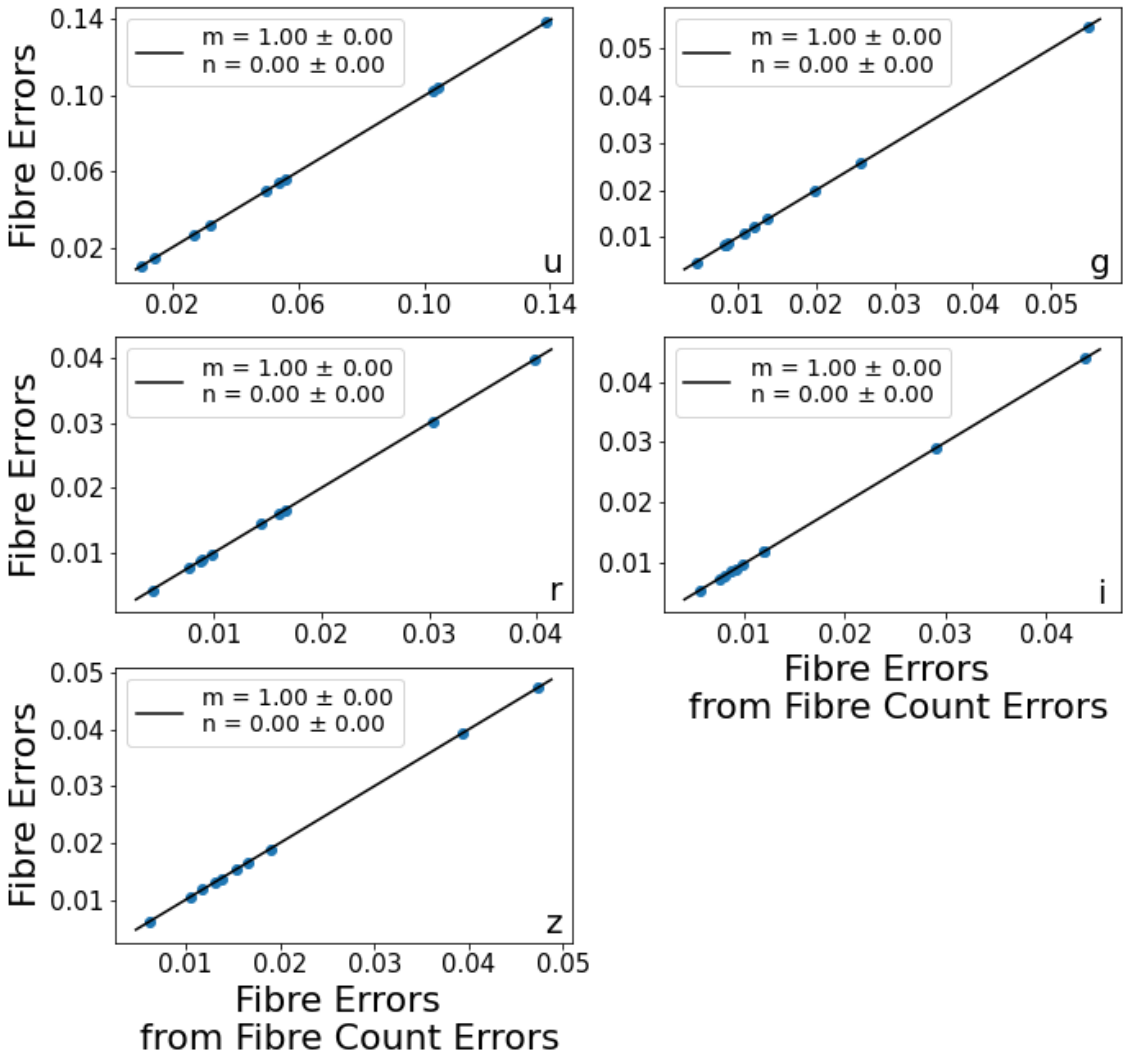


Figure A.3: Same as for Fig. A.2, but using the new equation for the magnitude errors (Eq. A.4).

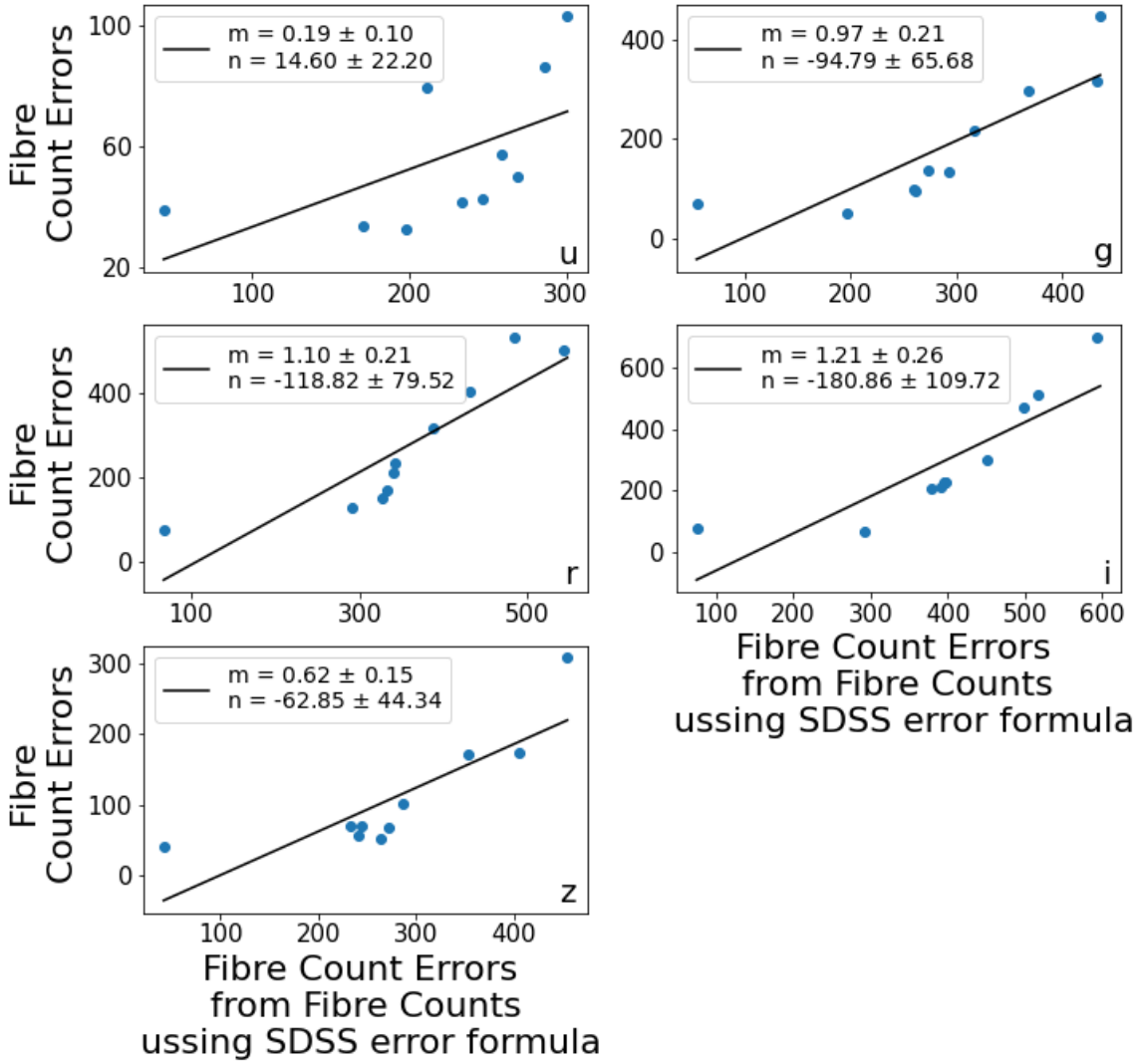


Figure A.4: Linear regression between the fibre count errors (y-axis) compared to the fibre count errors calculated using Eq. A.3 from the fibre counts (x-axis).

A.2 Count error formula

We have confirmed that the counts and count errors in fpObjc lead to the fibre magnitudes and errors, respectively. We have also corrected the exposure time factor in Eq. A.2. Nonetheless, when attempting to reproduce the fibre count errors using Eq. A.3, two problems arise. The first one is the ambiguity in the dimensional analysis. On the one hand, the units of `skyErr` are counts – in the fpObjc catalogue, although CasJobs contains them in maggies – but `dark variance` is given in counts squared. The formula is wrongly adding an error with a variance. On the other hand, the first and the second terms differ from each other in the gain denominator. Its units are [gain] = photo-electrons/counts, implying the first term is in counts^2/photo-electrons and the second in counts^2. This supported applying the correction `skyErr`^{2} over `dark variance`$^{1/2}$. The second problem is illustrated in Fig. A.4. The linear fits between the original fibre count errors and those calculated using Eq. A.3 are either quite deviated, as appears to be the case in bands u and z, or with a big variance and apparently a non-linear shape, as in the g, r, and i bands.

We defined Eq. A.5 improving Eq. A.3:

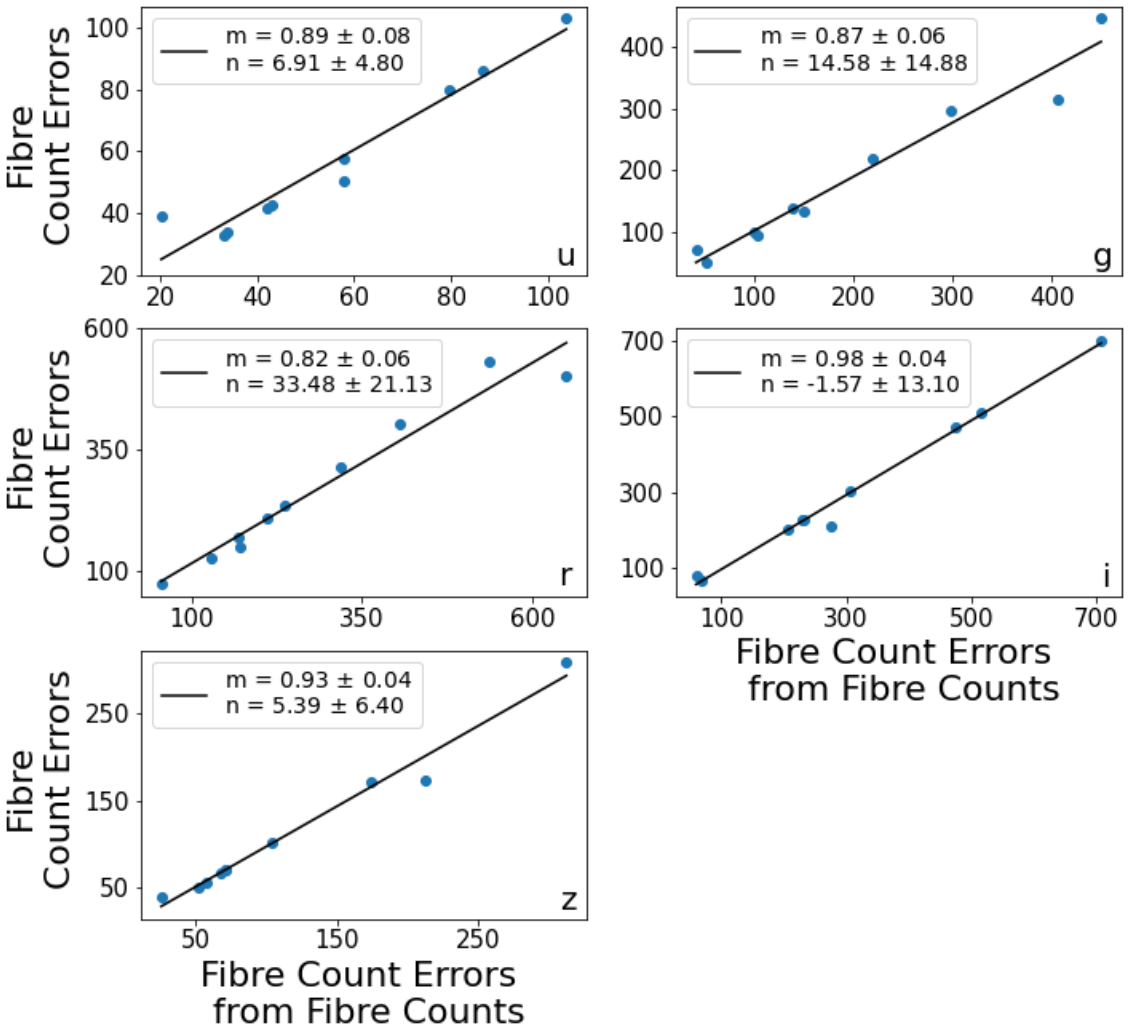


Figure A.5: Same as for Fig. A.4, but using the new equation for the count errors (Eq. A.5).

New counts error

$$\sigma^{\text{new}}_{\texttt{counts}} = \left[\frac{\texttt{counts} + (\texttt{sky} \cdot A_{\text{pixels}})}{\text{gain}} + A_{\text{pixels}} \cdot (\texttt{dark variance} + \texttt{skyErr}^2)\right]^{1/2} . \quad \text{(A.5)}$$

The improvement comes mainly from reducing the dimensional analysis ambiguity in the second term. We also changed the name of N_{pixels} to A_{pixels} so that it illustrates better that it is the area covered by the fibre aperture in pixel^2 units.

Figure A.5 compares the count errors with the result of Eq. A.5 on the counts. In contrast to Fig. A.4, a better linearity of the fit can be observed both visually and in the parameter's errors. The slope for all five bands is more uniform, and for the g, r, and i bands, a value of 1 for the slope is within the error bars, although it still deviates from the identity for u and z. Nonetheless, the uniformity of the fits supports Eq. A.5.

To finalize, using the calculated count errors in the x-axis of Fig. A.5, we applied the new magnitude error formula Eq. A.4 and compared the result with the fibre magnitude errors in Fig. A.6. The linear fit supported quite strongly the equation. The intercept was null for all bands and the slope differed from one only for the z band, showing a large relative error only for u. Some outliers seemed to spoil the results – such as the top one in the u-band panel, or the two separated ones in the top right area of the z-band panel. Figure A.7 shows, in contrast to Fig. A.6, the magnitude errors when applying Eq. A.3. From the slopes and the visual scatter of A.7, it is evident that the fibre errors were incorrectly written in the SDSS documentation. We note that the intercepts were all almost zero due to the nature of the asinh magnitude formula.

Our purpose in understanding the fibre errors was to identify its inputs to move forwards in Sect. 4.1.4. We did not study it further since we considered that the results confirmed that the `counts`, `sky`, `skyErr`, and `dark variance` were those inputs.

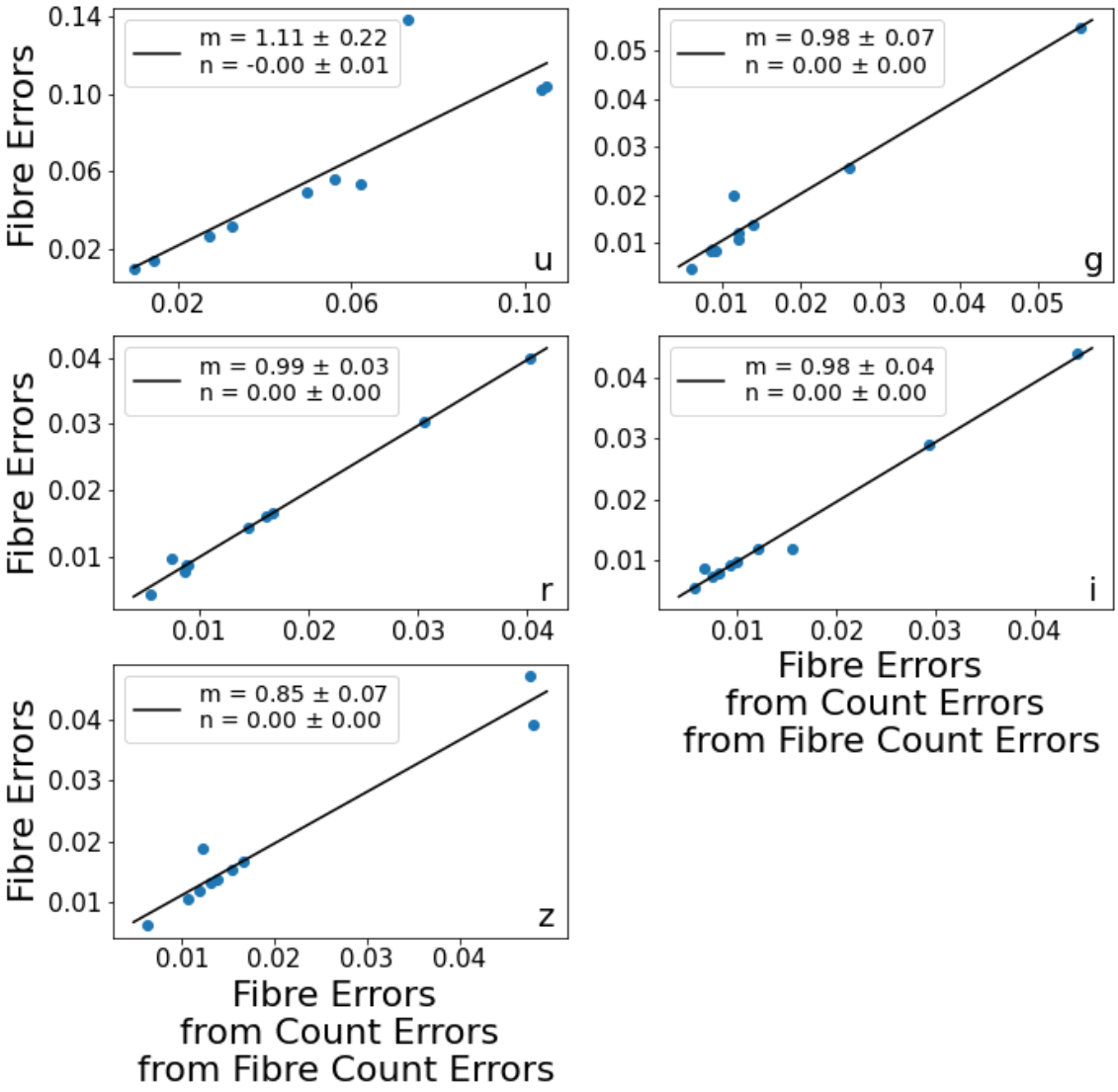


Figure A.6: Linear regression between fibre magnitude errors (y-axis), compared to the fibre magnitude errors calculated subsequently using Eqs. A.4 and A.5 on the fibre counts (x-axis).

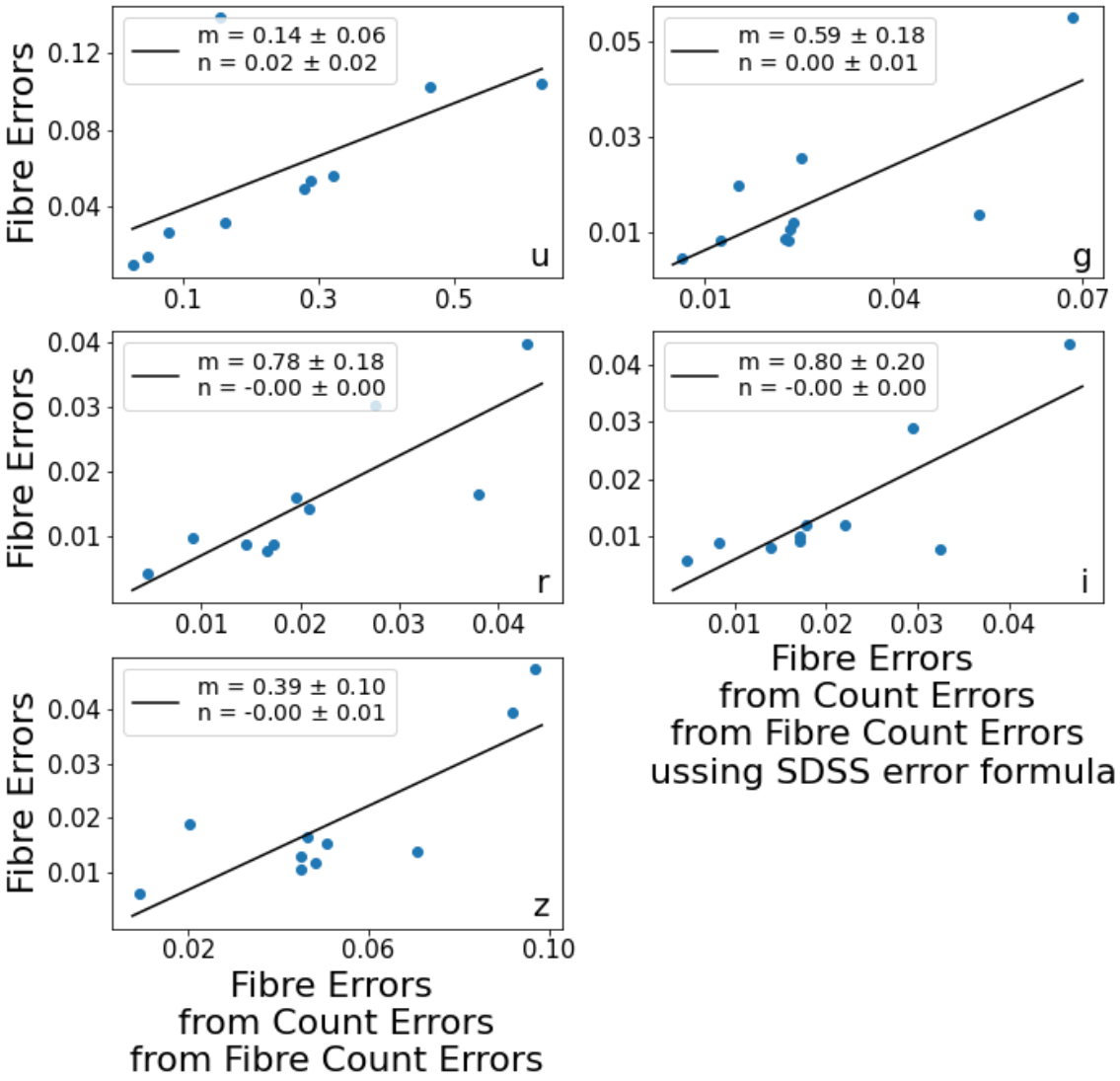


Figure A.7: Same as for Fig. A.6, but using the old equation for the count errors (Eq. A.3).

Bibliography

Abbott, T. M. C. et al. (Jan. 2022). Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. Phys. Rev. D 105.2, 023520, p. 023520. DOI: 10.1103/PhysRevD.105.023520. arXiv: 2105.13549 [astro-ph.CO].

Abraham, Roberto G. et al. (Sept. 1994). The Morphologies of Distant Galaxies. I. an Automated Classification System. ApJ 432, p. 75. DOI: 10.1086/174550.

Abraham, Roberto G. et al. (Nov. 1996). The Morphologies of Distant Galaxies. II. Classifications from the Hubble Space Telescope Medium Deep Survey. ApJS 107, p. 1. DOI: 10.1086/192352.

Abraham, Roberto G. et al. (May 2003). A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. ApJ 588.1, pp. 218–229. DOI: 10.1086/373919. arXiv: astro-ph/0301239 [astro-ph].

Ackermann, Sandro et al. (Sept. 2018). Using transfer learning to detect galaxy mergers. MNRAS 479.1, pp. 415–425. DOI: 10.1093/mnras/sty1398. arXiv: 1805.10289 [astro-ph.IM].

Adelman-McCarthy, Jennifer K. et al. (Apr. 2008). The Sixth Data Release of the Sloan Digital Sky Survey. ApJS 175.2, pp. 297–313. DOI: 10.1086/524984. arXiv: 0707.3413 [astro-ph].

Aihara, Hiroaki et al. (Jan. 2018). The Hyper Suprime-Cam SSP Survey: Overview and survey design. PASJ 70, S4, S4. DOI: 10.1093/pasj/psx066. arXiv: 1704.05858 [astro-ph.IM].

Aihara, Hiroaki et al. (Dec. 2019). Second data release of the Hyper Suprime-Cam Subaru Strategic Program. PASJ 71.6, 114, p. 114. DOI: 10.1093/pasj/psz103. arXiv: 1905.12221 [astro-ph.IM].

Angeloudi, Eirini et al. (Aug. 2023). ERGO-ML: towards a robust machine learning model for inferring the fraction of accreted stars in galaxies from integral-field spectroscopic maps. MNRAS 523.4, pp. 5408–5429. DOI: 10.1093/mnras/stad1669. arXiv: 2306.01056 [astro-ph.GA].

Annis, James et al. (Oct. 2014). The Sloan Digital Sky Survey Coadd: 275 deg^2 of Deep Sloan Digital Sky Survey Imaging on Stripe 82. ApJ 794.2, 120, p. 120. DOI: 10.1088/0004-637X/794/2/120. arXiv: 1111.6619 [astro-ph.CO].

Arp, Halton (Nov. 1966). Atlas of Peculiar Galaxies. ApJS 14, p. 1. DOI: 10.1086/190147.

Baldry, Ivan K. et al. (Jan. 2004). Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. ApJ 600.2, pp. 681–694. DOI: 10.1086/380092. arXiv: astro-ph/0309710 [astro-ph].

Balick, B. and R. L. Brown (Dec. 1974). Intense sub-arcsecond structure in the galactic center. ApJ 194, pp. 265–270. DOI: 10.1086/153242.

Bamford, Steven P. et al. (Mar. 2009). Galaxy Zoo: the dependence of morphology and colour on environment*. MNRAS 393.4, pp. 1324–1352. DOI: 10.1111/j.1365-2966.2008.14252.x. arXiv: 0805.2612 [astro-ph].

Barton, Elizabeth J. et al. (Feb. 2000). Tidally Triggered Star Formation in Close Pairs of Galaxies. ApJ 530.2, pp. 660–679. DOI: 10.1086/308392. arXiv: astro-ph/9909217 [astro-ph].

Bernardi, Mariangela et al. (Mar. 2011). Curvature in the colour-magnitude relation but not in colour-σ: major dry mergers at $M_* > 2 \times 10^{11}$ $M_\odot$? MNRAS 412.1, pp. 684–704. DOI: 10.1111/j.1365-2966.2010.17984.x. arXiv: 1005.3770 [astro-ph.CO].

Bershady, Matthew A. et al. (June 2000). Structural and Photometric Classification of Galaxies. I. Calibration Based on a Nearby Galaxy Sample. AJ 119.6, pp. 2645–2663. DOI: 10.1086/301386. arXiv: astro-ph/0002262 [astro-ph].

Bertin, E. and S. Arnouts (June 1996). SExtractor: Software for source extraction. A&AS 117, pp. 393–404. DOI: 10.1051/aas:1996164.

Bilicki, M. et al. (Aug. 2018). Photometric redshifts for the Kilo-Degree Survey. Machine-learning analysis with artificial neural networks. A&A 616, A69, A69. DOI: 10.1051/0004-6361/201731942. arXiv: 1709.04205 [astro-ph.CO].

Blanton, Michael R. et al. (Sept. 2005). The Properties and Luminosity Function of Extremely Low Luminosity Galaxies. ApJ 631.1, pp. 208–230. DOI: 10.1086/431416. arXiv: astro-ph/0410164 [astro-ph].

Blanton, Michael R. et al. (July 2011). Improved Background Subtraction for the Sloan Digital Sky Survey Images. AJ 142.1, 31, p. 31. DOI: 10.1088/0004-6256/142/1/31. arXiv: 1105.1960 [astro-ph.IM].

Bonnarel, F. et al. (Apr. 2000). The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. A&AS 143, pp. 33–40. DOI: 10.1051/aas:2000331.

Bosch, James et al. (Jan. 2018). The Hyper Suprime-Cam software pipeline. PASJ 70, S5, S5. DOI: 10.1093/pasj/psx080. arXiv: 1705.06766 [astro-ph.IM].

Bothun, G. et al. (July 1997). Low-Surface-Brightness Galaxies: Hidden Galaxies Revealed. PASP 109, pp. 745–758. DOI: 10.1086/133941.

Bottrell, Connor et al. (Dec. 2019). Deep learning predictions of galaxy merger stage and the importance of observational realism. MNRAS 490.4, pp. 5390–5413. DOI: 10.1093/mnras/stz2934. arXiv: 1910.07031 [astro-ph.GA].

Broadfoot, A. Lyle and Kenneth R. Kendall (Jan. 1968). The airglow spectrum, 3100-10,000 A. J. Geophys. Res. 73.1, p. 426. DOI: 10.1029/JA073i001p00426.

Calzetti, Daniela et al. (Apr. 2000). The Dust Content and Opacity of Actively Star-forming Galaxies. ApJ 533.2, pp. 682–695. DOI: 10.1086/308692. arXiv: astro-ph/9911459 [astro-ph].

Chambers, K. C. et al. (Dec. 2016). The Pan-STARRS1 Surveys. *arXiv e-prints*, arXiv:1612.05560, arXiv:1612.05560. DOI: 10.48550/arXiv.1612.05560. arXiv: 1612.05560 [astro-ph.IM].

Charmandaris, V. et al. (Apr. 2000). First detection of molecular gas in the shells of CenA. A&A 356, pp. L1–L4. DOI: 10.48550/arXiv.astro-ph/0003175. arXiv: astro-ph/0003175 [astro-ph].

Conroy, Charlie (Aug. 2013). Modeling the Panchromatic Spectral Energy Distributions of Galaxies. ARA&A 51.1, pp. 393–455. DOI: 10.1146/annurev-astro-082812-141017. arXiv: 1301.7095 [astro-ph.CO].

Conselice, Christopher J. (July 2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. ApJS 147.1, pp. 1–28. DOI: 10.1086/375001. arXiv: astro-ph/0303065 [astro-ph].

— (Feb. 2006). Early and Rapid Merging as a Formation Mechanism of Massive Galaxies: Empirical Constraints. ApJ 638.2, pp. 686–702. DOI: 10.1086/499067. arXiv: astro-ph/0507146 [astro-ph].

Conselice, Christopher J. et al. (Feb. 2000). The Asymmetry of Galaxies: Physical Morphology for Nearby and High-Redshift Galaxies. ApJ 529.2, pp. 886–910. DOI: 10.1086/308300. arXiv: astro-ph/9907399 [astro-ph].

Conselice, Christopher J. et al. (Sept. 2003). A Direct Measurement of Major Galaxy Mergers at z<∼3. AJ 126.3, pp. 1183–1207. DOI: 10.1086/377318. arXiv: astro-ph/0306106 [astro-ph].

Corcho-Caballero, Pablo et al. (Jan. 2023). Ageing and quenching through the ageing diagram: predictions from simulations and observational constraints. *Monthly Notices of the Royal Astronomical Society* 520.1, pp. 193–209. ISSN: 0035-8711. DOI: 10.1093/mnras/stad147. eprint: https://academic.oup.com/mnras/article-pdf/520/1/193/48956633/stad147.pdf. URL: https://doi.org/10.1093/mnras/stad147.

Darg, D. W. et al. (Jan. 2010a). Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies. MNRAS 401.2, pp. 1043–1056. DOI: 10.1111/j.1365-2966.2009.15686.x. arXiv: 0903.4937 [astro-ph.GA].

Darg, D. W. et al. (Jan. 2010b). Galaxy Zoo: the properties of merging galaxies in the nearby Universe - local environments, colours, masses, star formation rates and AGN activity. MNRAS 401.3, pp. 1552–1563. DOI: 10.1111/j.1365-2966.2009.15786.x. arXiv: 0903.5057 [astro-ph.GA].

Dark Energy Survey and Kilo-Degree Survey Collaboration et al. (Oct. 2023). DES Y3 + KiDS-1000: Consistent cosmology combining cosmic shear surveys. *The Open Journal of Astrophysics* 6, 36, p. 36. DOI: 10.21105/astro.2305.17173. arXiv: 2305.17173 [astro-ph.CO].

de Jong, J. T. A. et al. (Dec. 2013a). The Kilo-Degree Survey. *The Messenger* 154, pp. 44–46.

de Jong, Jelte T. A. et al. (Jan. 2013b). The Kilo-Degree Survey. *Experimental Astronomy* 35.1-2, pp. 25–44. DOI: 10.1007/s10686-012-9306-1. arXiv: 1206.1254 [astro-ph.CO].

De Propris, Roberto et al. (Oct. 2005). The Millennium Galaxy Catalogue: Dynamically Close Pairs of Galaxies and the Global Merger Rate. AJ 130.4, pp. 1516–1523. DOI: 10.1086/433169. arXiv: astro-ph/0506635 [astro-ph].

de Vaucouleurs, G. (Apr. 1963). Revised Classification of 1500 Bright Galaxies. ApJS 8, p. 31. DOI: 10.1086/190084.

de Vaucouleurs, Gerard (Jan. 1959). Classification and Morphology of External Galaxies. *Handbuch der Physik* 53, p. 275. DOI: 10.1007/978-3-642-45932-0_7.

Dieleman, Sander et al. (June 2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. MNRAS 450.2, pp. 1441–1459. DOI: 10.1093/mnras/stv632. arXiv: 1503.07077 [astro-ph.IM].

Dolgov, A. D. (Jan. 2000). Dark matter in the universe. *Astronomical and Astrophysical Transactions* 19.3, pp. 305–326. DOI: 10.1080/10556790008238580. arXiv: hep-ph/9910532 [hep-ph].

Domínguez Sánchez, H. et al. (Feb. 2018). Improving galaxy morphologies for SDSS with Deep Learning. MNRAS 476.3, pp. 3661–3676. DOI: 10.1093/mnras/sty338. arXiv: 1711.05744 [astro-ph.GA].

Domínguez Sánchez, H. et al. (May 2023). Identification of tidal features in deep optical galaxy images with convolutional neural networks. MNRAS 521.3, pp. 3861–3872. DOI: 10.1093/mnras/stad750. arXiv: 2303.03407 [astro-ph.IM].

Driver, Simon P. et al. (Oct. 2009). GAMA: towards a physical understanding of galaxy formation. *Astronomy and Geophysics* 50.5, pp. 5.12–5.19. DOI: 10.1111/j.1468-4004.2009.50512.x. arXiv: 0910.5123 [astro-ph.CO].

Duc, P. A. et al. (Sept. 2000). Formation of a Tidal Dwarf Galaxy in the Interacting System Arp 245 (NGC 2992/93). AJ 120.3, pp. 1238–1264. DOI: 10.1086/301516. arXiv: astro-ph/0006038 [astro-ph].

Duc, Pierre-Alain and Florent Renaud (2013). Tides in Colliding Galaxies. *Lecture Notes in Physics, Berlin Springer Verlag*. Ed. by Jean Souchay et al. Vol. 861, p. 327. DOI: 10.1007/978-3-642-32961-6_9.

Duc, Pierre-Alain et al. (Jan. 2015). The ATLAS^{3D} project - XXIX. The new look of early-type galaxies and surrounding fields disclosed by extremely deep optical images. MNRAS 446.1, pp. 120–143. DOI: 10.1093/mnras/stu2019. arXiv: 1410.0981 [astro-ph.GA].

Duncan, Kenneth et al. (May 2019). Observational Constraints on the Merger History of Galaxies since z $\approx$ 6: Probabilistic Galaxy Pair Counts in the CANDELS Fields. ApJ 876.2, 110, p. 110. DOI: 10.3847/1538-4357/ab148a. arXiv: 1903.12188 [astro-ph.GA].

Dye, S. and S. J. Warren (2005). Decomposition of the Visible and Dark Matter in the Einstein Ring 0047–2808 by Semilinear Inversion. *The Astrophysical Journal* 623.1, p. 31. DOI: 10.1086/428340. URL: https://dx.doi.org/10.1086/428340.

Edelsbrunner, Herbert et al. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory* 29.4, pp. 551–559. DOI: 10.1109/TIT.1983.1056714.

Einasto, J. (Mar. 1969). The andromeda galaxy M 31: I. A preliminary model. *Astrophysics* 5.1, pp. 67–80. DOI: 10.1007/BF01013353.

Eisenstein, Daniel J. et al. (Nov. 2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. ApJ 633.2, pp. 560–574. DOI: 10.1086/466512. arXiv: astro-ph/0501171 [astro-ph].

Ekers, R. D. et al. (Oct. 1975). A full synthesis map of Sgr A at 5 GHz. A&A 43, pp. 159–166.

Eliche-Moral, M. C. et al. (Oct. 2018). Formation of S0 galaxies through mergers. Morphological properties: tidal relics, lenses, ovals, and other inner components. A&A 617, A113, A113. DOI: 10.1051/0004-6361/201832911. arXiv: 1806.06070 [astro-ph.GA].

Ellison, Sara L. et al. (May 2008). Galaxy Pairs in the Sloan Digital Sky Survey. I. Star Formation, Active Galactic Nucleus Fraction, and the Mass-Metallicity Relation. AJ 135.5, pp. 1877–1899. DOI: 10.1088/0004-6256/135/5/1877. arXiv: 0803.0161 [astro-ph].

Ellison, Sara L. et al. (Aug. 2019). A definitive merger-AGN connection at z $\sim$ 0 with CFIS: mergers have an excess of AGN and AGN hosts are more frequently disturbed. MNRAS 487.2, pp. 2491–2504. DOI: 10.1093/mnras/stz1431. arXiv: 1905.08830 [astro-ph.GA].

Etherington, I.M.H. (1933). LX. On the definition of distance in general relativity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 15.100, pp. 761–773. DOI: 10.1080/14786443309462220. eprint: https://doi.org/10.1080/14786443309462220. URL: https://doi.org/10.1080/14786443309462220.

Event Horizon Telescope Collaboration et al. (May 2022). First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way. ApJ 930.2, L12, p. L12. DOI: 10.3847/2041-8213/ac6674.

Feldmann, Robert et al. (Feb. 2016). The formation of massive, quiescent galaxies at cosmic noon. *Monthly Notices of the Royal Astronomical Society: Letters* 458.1, pp. L14–L18. ISSN: 1745-3925. DOI: 10.1093/mnrasl/slw014. eprint: https://academic.oup.com/mnrasl/article-pdf/458/1/L14/56942721/mnrasl_458_1_l14.pdf. URL: https://doi.org/10.1093/mnrasl/slw014.

Fensch, Jérémy et al. (Dec. 2020). Shedding light on the formation mechanism of shell galaxy NGC 474 with MUSE. *A&A* 644, A164, A164. DOI: 10.1051/0004-6361/202038550. arXiv: 2007.03318 [astro-ph.GA].

Fernie, J. D. (Dec. 1969). The Period-Luminosity Relation: A Historical Review. PASP 81.483, p. 707. DOI: 10.1086/128847.

Ferreira, Leonardo et al. (June 2020). Galaxy Merger Rates up to z $\sim$ 3 Using a Bayesian Deep Learning Model: A Major-merger Classifier Using IllustrisTNG Simulation Data. ApJ 895.2, 115, p. 115. DOI: 10.3847/1538-4357/ab8f9b. arXiv: 2005.00476 [astro-ph.GA].

Flewelling, H. A. et al. (Nov. 2020). The Pan-STARRS1 Database and Data Products. ApJS 251.1, 7, p. 7. DOI: 10.3847/1538-4365/abb82d. arXiv: 1612.05243 [astro-ph.IM].

Fox, David C. and Abraham Loeb (Dec. 1997). Do the Electrons and Ions in X-Ray Clusters Share the Same Temperature? ApJ 491.2, pp. 459–466. DOI: 10.1086/305007. arXiv: astro-ph/9706266 [astro-ph].

Gaia Collaboration et al. (Nov. 2016). The Gaia mission. A&A 595, A1, A1. DOI: 10.1051/0004-6361/201629272. arXiv: 1609.04153 [astro-ph.IM].

Gail, Hans-Peter and Peter Hoppe (2010). The Origins of Protoplanetary Dust and the Formation of Accretion Disks. *Protoplanetary Dust: Astrophysical and Cosmochemical Perspectives*. Ed. by D. A. Apai and D. S. Lauretta, pp. 27–65. DOI: 10.1017/CBO9780511674662.003.

Goldberger, Jacob et al. (Jan. 2004). Neighbourhood Components Analysis. Vol. 17.

Goto, Tomotsugu et al. (Mar. 2017). Hyper Suprime-Camera Survey of the Akari NEP Wide Field. *Publication of Korean Astronomical Society* 32.1, pp. 225–230. DOI: 10.5303/PKAS.2017.32.1.225. arXiv: 1505.00012 [astro-ph.GA].

Gunn, J. E. et al. (Dec. 1998). The Sloan Digital Sky Survey Photometric Camera. AJ 116.6, pp. 3040–3081. DOI: 10.1086/300645. arXiv: astro-ph/9809085 [astro-ph].

Hamed, M. et al. (Nov. 2023a). Decoding the IRX-β dust attenuation relation in star-forming galaxies at intermediate redshift. A&A 679, A26, A26. DOI: 10.1051/0004-6361/202346976. arXiv: 2309.01819 [astro-ph.GA].

Hamed, M. et al. (June 2023b). The slippery slope of dust attenuation curves. Correlation of dust attenuation laws with star-to-dust compactness up to z = 4. A&A 674, A99, A99. DOI: 10.1051/0004-6361/202245818. arXiv: 2304.13713 [astro-ph.GA].

Helmi, Amina (Aug. 2020). Streams, Substructures, and the Early History of the Milky Way. ARA&A 58, pp. 205–256. DOI: 10.1146/annurev-astro-032620-021917. arXiv: 2002.04340 [astro-ph.GA].

Hernquist, Lars and P. J. Quinn (Aug. 1988). Formation of Shell Galaxies. I. Spherical Potentials. *ApJ* 331, p. 682. DOI: 10.1086/166592.

Heymans, Catherine et al. (Feb. 2021). KiDS-1000 Cosmology: Multi-probe weak gravitational lensing and spectroscopic galaxy clustering constraints. A&A 646, A140, A140. DOI: 10.1051/0004-6361/202039063. arXiv: 2007.15632 [astro-ph.CO].

Hildebrandt, H. et al. (Nov. 2010). PHAT: PHoto-z Accuracy Testing. A&A 523, A31, A31. DOI: 10.1051/0004-6361/201014885. arXiv: 1008.0658 [astro-ph.CO].

Holwerda, Benne W. et al. (Sept. 2019). The Frequency of Dust Lanes in Edge-on Spiral Galaxies Identified by Galaxy Zoo in KiDS Imaging of GAMA Targets. AJ 158.3, 103, p. 103. DOI: 10.3847/1538-3881/ab2886. arXiv: 1909.07461 [astro-ph.GA].

Hood, Callie E. et al. (Apr. 2018). The Origin of Faint Tidal Features around Galaxies in the RESOLVE Survey. *ApJ* 857.2, 144, p. 144. DOI: 10.3847/1538-4357/aab719. arXiv: 1803.05447 [astro-ph.GA].

Hoskin, M. A. (Jan. 1976). The 'Great Debate': What Really Happened. *Journal for the History of Astronomy* 7, p. 169. DOI: 10.1177/002182867600700302.

Hotelling, Harold (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, pp. 498–520.

Hubble, E. P. (Jan. 1925). Cepheids in Spiral Nebulae. *Popular Astronomy* 33, pp. 252–255.

— (May 1926). A spiral nebula as a stellar system: Messier 33. ApJ 63, pp. 236–274. DOI: 10.1086/142976.

— (Mar. 1929). A spiral nebula as a stellar system, Messier 31. ApJ 69, pp. 103–158. DOI: 10.1086/143167.

Hwang, H. S. et al. (Nov. 2011). GOODS-Herschel: the impact of galaxy-galaxy interactions on the far-infrared properties of galaxies. A&A 535, A60, A60. DOI: 10.1051/0004-6361/201117476. arXiv: 1109.1937 [astro-ph.CO].

Hyde, Joseph B. and Mariangela Bernardi (Apr. 2009). Curvature in the scaling relations of early-type galaxies. MNRAS 394.4, pp. 1978–1990. DOI: 10.1111/j.1365-2966.2009.14445.x. arXiv: 0810.4922 [astro-ph].

Ioffe, Sergey and Christian Szegedy (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

Ivezić, Željko et al. (Mar. 2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. ApJ 873.2, 111, p. 111. DOI: 10.3847/1538-4357/ab042c. arXiv: 0805.2366 [astro-ph].

Iye, Masanori (July 2021). Subaru Telescope —History, active/adaptive optics, instruments, and scientific achievements—. *Proceedings of the Japan Academy, Series B* 97, pp. 337–370. DOI: 10.2183/pjab.97.019.

Jackson, R A et al. (Jan. 2022). Extremely massive disc galaxies in the nearby Universe form through gas-rich minor mergers. *Monthly Notices of the Royal Astronomical Society* 511.1, pp. 607–615. ISSN: 0035-8711. DOI: 10.1093/mnras/stac058. eprint: https://academic.oup.com/mnras/article-pdf/511/1/607/48246790/stac058.pdf. URL: https://doi.org/10.1093/mnras/stac058.

Jeans, J. H. (Jan. 1902). The Stability of a Spherical Nebula. *Philosophical Transactions of the Royal Society of London Series A* 199, pp. 1–53. DOI: 10.1098/rsta.1902.0012.

Joseph, R. D. and G. S. Wright (May 1985). Recent star formation in interacting galaxies - II. Super starbursts in merging galaxies. MNRAS 214, pp. 87–95. DOI: 10.1093/mnras/214.2.87.

Keel, W. C. et al. (May 1985). The effects of interactions on spiral galaxies. I. Nuclear activity and star formation. AJ 90, pp. 708–730. DOI: 10.1086/113779.

Kent, S. M. (Oct. 1985). CCD surface photometry of field galaxies. II. Bulge/disk decompositions. ApJS 59, pp. 115–159. DOI: 10.1086/191066.

Kim, S. J. et al. (Dec. 2012). The North Ecliptic Pole Wide survey of AKARI: a near- and mid-infrared source catalog. A&A 548, A29, A29. DOI: 10.1051/0004-6361/201219105. arXiv: 1208.5008 [astro-ph.CO].

Kim, Seong Jin et al. (Jan. 2021). Identification of AKARI infrared sources by the Deep HSC Optical Survey: construction of a new band-merged catalogue in the North Ecliptic Pole Wide field. MNRAS 500.3, pp. 4078–4094. DOI: 10.1093/mnras/staa3359. arXiv: 2012.00750 [astro-ph.GA].

Kingma, Diederik P. and Jimmy Ba (Dec. 2014). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980, arXiv:1412.6980. arXiv: 1412.6980 [cs.LG].

Kitzbichler, M. G. and S. D. M. White (Dec. 2008). A calibration of the relation between the abundance of close galaxy pairs and the rate of galaxy mergers. MNRAS 391.4, pp. 1489–1498. DOI: 10.1111/j.1365-2966.2008.13873.x. arXiv: 0804.1965 [astro-ph].

Knapen, J. H. et al. (Dec. 1995). The Central Region in M100: Observations and Modeling. ApJ 454, p. 623. DOI: 10.1086/176516. arXiv: astro-ph/9506098 [astro-ph].

Kocevski, Dale D. et al. (Jan. 2012). CANDELS: Constraining the AGN-Merger Connection with Host Morphologies at z ~2. ApJ 744.2, 148, p. 148. DOI: 10.1088/0004-637X/744/2/148. arXiv: 1109.2588 [astro-ph.CO].

Krajnović, Davor et al. (July 2011). The ATLAS^{3D} project - II. Morphologies, kinemetric features and alignment between photometric and kinematic axes of early-type galaxies. MNRAS 414.4, pp. 2923–2949. DOI: 10.1111/j.1365-2966.2011.18560.x. arXiv: 1102.3801 [astro-ph.CO].

Lambas, Diego G. et al. (Dec. 2003). Galaxy pairs in the 2dF survey - I. Effects of interactions on star formation in the field. MNRAS 346.4, pp. 1189–1196. DOI: 10.1111/j.1365-2966.2003.07179.x. arXiv: astro-ph/0212222 [astro-ph].

Lazar, I. et al. (Apr. 2023). Relaxed blue ellipticals: accretion-driven stellar growth is a key evolutionary channel for low mass elliptical galaxies. MNRAS 520.2, pp. 2109–2120. DOI: 10.1093/mnras/stad224. arXiv: 2302.06631 [astro-ph.GA].

Lazar, I. et al. (Mar. 2024). The morphological mix of dwarf galaxies in the nearby Universe. MNRAS 529.1, pp. 499–518. DOI: 10.1093/mnras/stae510. arXiv: 2402.12440 [astro-ph.GA].

Le Fèvre, O. et al. (Jan. 2000). Hubble Space Telescope imaging of the CFRS and LDSS redshift surveys - IV. Influence of mergers in the evolution of faint field galaxies from z∼1. MNRAS 311.3, pp. 565–575. DOI: 10.1046/j.1365-8711.2000.03083.x. arXiv: astro-ph/9909211 [astro-ph].

Leavitt, Henrietta S. and Edward C. Pickering (Mar. 1912). Periods of 25 Variable Stars in the Small Magellanic Cloud. *Harvard College Observatory Circular* 173, pp. 1–3.

Lee, Hyung Mok et al. (Feb. 2009). North Ecliptic Pole Wide Field Survey of AKARI: Survey Strategy and Data Characteristics. PASJ 61, p. 375. DOI: 10.1093/pasj/61.2.375. arXiv: 0901.3256 [astro-ph.GA].

Leinert, Ch. et al. (Jan. 1998). The 1997 reference of diffuse night sky brightness. A&AS 127, pp. 1–99. DOI: 10.1051/aas:1998105.

Lemaître, Georges (1933). L'univers en expansion. *Annales de la Société scientifique de Bruxelles*. Vol. 53, p. 51.

Lin, Lihwai et al. (Dec. 2004). The DEEP2 Galaxy Redshift Survey: Evolution of Close Galaxy Pairs and Major-Merger Rates up to z ∼1.2. ApJ 617.1, pp. L9–L12. DOI: 10.1086/427183. arXiv: astro-ph/0411104 [astro-ph].

Lintott, Chris et al. (Jan. 2011). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. MNRAS 410.1, pp. 166–178. DOI: 10.1111/j.1365-2966.2010.17432.x. arXiv: 1007.3265 [astro-ph.GA].

Lintott, Chris J. et al. (Sept. 2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. MNRAS 389.3, pp. 1179–1189. DOI: 10.1111/j.1365-2966.2008.13689.x. arXiv: 0804.4483 [astro-ph].

Lo, K. Y. et al. (Dec. 1975). VLBI observations of the compact radio source in the center of the Galaxy. ApJ 202, pp. L63–L65. DOI: 10.1086/181982.

Lofthouse, E. K. et al. (Mar. 2017). Major mergers are not significant drivers of star formation or morphological transformation around the epoch of peak cosmic star formation. MNRAS 465.3, pp. 2895–2900. DOI: 10.1093/mnras/stw2895. arXiv: 1608.03892 [astro-ph.GA].

Lorenzon, G. et al. (Apr. 2024). Tracing the evolutionary pathways of dust and cold gas in high-z quiescent galaxies with SIMBA. *arXiv e-prints*, arXiv:2404.10568, arXiv:2404.10568. DOI: 10.48550/arXiv.2404.10568. arXiv: 2404.10568 [astro-ph.GA].

Lotz, Jennifer M. et al. (July 2004). A New Nonparametric Approach to Galaxy Morphological Classification. AJ 128.1, pp. 163–182. DOI: 10.1086/421849. arXiv: astro-ph/0311352 [astro-ph].

Lotz, Jennifer M. et al. (Nov. 2008). Galaxy merger morphologies and time-scales from simulations of equal-mass gas-rich disc mergers. *Monthly Notices of the Royal Astronomical Society* 391.3, pp. 1137–1162. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2008.14004.x. eprint: https://academic.oup.com/mnras/article-pdf/391/3/1137/13143271/mnras391-1137.pdf. URL: https://doi.org/10.1111/j.1365-2966.2008.14004.x.

Lotz, Jennifer M. et al. (Dec. 2011). The Major and Minor Galaxy Merger Rates at z < 1.5. ApJ 742.2, 103, p. 103. DOI: 10.1088/0004-637X/742/2/103. arXiv: 1108.2508 [astro-ph.CO].

Lupton, Robert H. et al. (Sept. 1999). A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements. AJ 118.3, pp. 1406–1410. DOI: 10.1086/301004. arXiv: astro-ph/9903081 [astro-ph].

Maaten, Laurens van der and Geoffrey Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Magnier, Eugene. A. et al. (Nov. 2020a). Pan-STARRS Photometric and Astrometric Calibration. ApJS 251.1, 6, p. 6. DOI: 10.3847/1538-4365/abb82a. arXiv: 1612.05242 [astro-ph.IM].

Magnier, Eugene A. et al. (Nov. 2020b). Pan-STARRS Pixel Analysis: Source Detection and Characterization. ApJS 251.1, 5, p. 5. DOI: 10.3847/1538-4365/abb82c. arXiv: 1612.05244 [astro-ph.IM].

Magnier, Eugene A. et al. (Nov. 2020c). The Pan-STARRS Data-processing System. ApJS 251.1, 3, p. 3. DOI: 10.3847/1538-4365/abb829. arXiv: 1612.05240 [astro-ph.IM].

Mahajan, Smriti et al. (Mar. 2018). Galaxy And Mass Assembly (GAMA): blue spheroids within 87 Mpc. MNRAS 475.1, pp. 788–799. DOI: 10.1093/mnras/stx3202. arXiv: 1712.03644 [astro-ph.GA].

Małek, K. et al. (May 2010). Star forming galaxies in the AKARI deep field south: identifications and spectral energy distributions. A&A 514, A11, A11. DOI: 10.1051/0004-6361/200913419. arXiv: 0911.5598 [astro-ph.CO].

Małek, K. et al. (Apr. 2024). Attenuation proxy hidden in surface brightness - colour diagrams. A new strategy for the LSST era. A&A 684, A30, A30. DOI: 10.1051/0004-6361/202348432. arXiv: 2401.12831 [astro-ph.GA].

Margalef-Bentabol, B. et al. (Mar. 2024). Galaxy merger challenge: A comparison study between machine learning-based detection methods. *arXiv e-prints*, arXiv:2403.15118, arXiv:2403.15118. DOI: 10.48550/arXiv.2403.15118. arXiv: 2403.15118 [astro-ph.GA].

Marinacci, Federico et al. (Nov. 2018). First results from the IllustrisTNG simulations: radio haloes and magnetic fields. MNRAS 480.4, pp. 5113–5139. DOI: 10.1093/mnras/sty2206. arXiv: 1707.03396 [astro-ph.CO].

Martel, Hugo et al. (May 2013). The connection between star formation and metallicity evolution in barred spiral galaxies. MNRAS 431.3, pp. 2560–2575. DOI: 10.1093/mnras/stt354. arXiv: 1302.6211 [astro-ph.CO].

Martin, G. et al. (May 2019). The formation and evolution of low-surface-brightness galaxies. MNRAS 485.1, pp. 796–818. DOI: 10.1093/mnras/stz356. arXiv: 1902.04580 [astro-ph.GA].

Maschmann, Daniel et al. (Sept. 2020). Double-peak emission line galaxies in the SDSS catalogue. A minor merger sequence. A&A 641, A171, A171. DOI: 10.1051/0004-6361/202037868. arXiv: 2007.14410 [astro-ph.GA].

Massey, Philip and Craig B. Foltz (Apr. 2000). The Spectrum of the Night Sky over Mount Hopkins and Kitt Peak: Changes after a Decade. PASP 112.770, pp. 566–573. DOI: 10.1086/316552.

Matsuhara, Hideo et al. (Aug. 2006). Deep Extragalactic Surveys around the Ecliptic Poles with AKARI (ASTRO-F). *Publications of the Astronomical Society of Japan* 58.4, pp. 673–694. ISSN: 0004-6264. DOI: 10.1093/pasj/58.4.673. eprint: https://academic.oup.com/pasj/article-pdf/58/4/673/54711615/pasj_58_4_673.pdf. URL: https://doi.org/10.1093/pasj/58.4.673.

Mayor, Michel and Didier Queloz (Nov. 1995). A Jupiter-mass companion to a solar-type star. Nature 378.6555, pp. 355–359. DOI: 10.1038/378355a0.

McConnachie, Alan W. (2012). THE OBSERVED PROPERTIES OF DWARF GALAXIES IN AND AROUND THE LOCAL GROUP. *The Astronomical Journal* 144.1, p. 4. DOI: 10.1088/0004-6256/144/1/4. URL: https://dx.doi.org/10.1088/0004-6256/144/1/4.

McGaugh, Stacy S. (May 1996). The number, luminosity and mass density of spiral galaxies as a function of surface brightness. MNRAS 280.2, pp. 337–354. DOI: 10.1093/mnras/280.2.337. arXiv: astro-ph/9511010 [astro-ph].

McKee, C. F. and J. P. Ostriker (Nov. 1977). A theory of the interstellar medium: three components regulated by supernova explosions in an inhomogeneous substrate. ApJ 218, pp. 148–169. DOI: 10.1086/155667.

Melchior, Peter et al. (Aug. 2021). The challenge of blending in large sky surveys. *Nature Reviews Physics* 3.10, pp. 712–718. DOI: 10.1038/s42254-021-00353-y.

Merritt, David et al. (Dec. 2006). Empirical Models for Dark Matter Halos. I. Nonparametric Construction of Density Profiles and Comparison with Parametric Models. AJ 132.6, pp. 2685–2700. DOI: 10.1086/508988. arXiv: astro-ph/0509417 [astro-ph].

Mihos, J. Christopher and Lars Hernquist (June 1996). Gasdynamics and Starbursts in Major Mergers. ApJ 464, p. 641. DOI: 10.1086/177353. arXiv: astro-ph/9512099 [astro-ph].

Montes, Mireia et al. (Mar. 2021). The Buildup of the Intracluster Light of A85 as Seen by Subaru's Hyper Suprime-Cam. ApJ 910.1, 45, p. 45. DOI: 10.3847/1538-4357/abddb6. arXiv: 2101.08290 [astro-ph.GA].

Morgan, W. W. (Aug. 1958). A Preliminary Classification of the Forms of Galaxies According to Their Stellar Population. PASP 70.415, p. 364. DOI: 10.1086/127243.

Mundy, Carl J. et al. (Sept. 2017). A consistent measure of the merger histories of massive galaxies using close-pair statistics - I. Major mergers at z < 3.5. MNRAS 470.3, pp. 3507–3531. DOI: 10.1093/mnras/stx1238. arXiv: 1705.07986 [astro-ph.GA].

Nagai, Daisuke et al. (Jan. 2007). Testing X-Ray Measurements of Galaxy Clusters with Cosmological Simulations. ApJ 655.1, pp. 98–108. DOI: 10.1086/509868. arXiv: astro-ph/0609247 [astro-ph].

Nair, Vinod and Geoffrey E. Hinton (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. ICML'10. Haifa, Israel: Omnipress, 807–814. ISBN: 9781605589077.

Nanni, A. et al. (Sept. 2020). The gas, metal, and dust evolution in low-metallicity local and high-redshift galaxies. A&A 641, A168, A168. DOI: 10.1051/0004-6361/202037833. arXiv: 2006.15146 [astro-ph.GA].

Navarro, Julio F. et al. (Dec. 1997). A Universal Density Profile from Hierarchical Clustering. ApJ 490.2, pp. 493–508. DOI: 10.1086/304888. arXiv: astro-ph/9611107 [astro-ph].

Nevin, R. et al. (Feb. 2019). Accurate Identification of Galaxy Mergers with Imaging. ApJ 872.1, 76, p. 76. DOI: 10.3847/1538-4357/aafd34. arXiv: 1901.01975 [astro-ph.GA].

Newton, Oliver et al. (Apr. 2023). The Undiscovered Ultradiffuse Galaxies of the Local Group. ApJ 946.2, L37, p. L37. DOI: 10.3847/2041-8213/acc2bb. arXiv: 2212.05066 [astro-ph.GA].

Nightingale, James W et al. (Dec. 2023). Scanning for dark matter subhaloes in Hubble Space Telescope imaging of 54 strong lenses. *Monthly Notices of the Royal Astronomical Society* 527.4, pp. 10480–10506. ISSN: 0035-8711. DOI: 10.1093/mnras/stad3694. eprint: https://academic.oup.com/mnras/article-pdf/527/4/10480/54943520/stad3694.pdf. URL: https://doi.org/10.1093/mnras/stad3694.

Oi, Nagisa et al. (Jan. 2021). Subaru/HSC deep optical imaging of infrared sources in the AKARI North Ecliptic Pole-Wide field. MNRAS 500.4, pp. 5024–5042. DOI: 10.1093/mnras/staa3080.

Oke, J. B. and J. E. Gunn (Mar. 1983). Secondary standard stars for absolute spectrophotometry. ApJ 266, pp. 713–717. DOI: 10.1086/160817.

Omori, Kiyoaki Christopher et al. (Nov. 2023). Galaxy mergers in Subaru HSC-SSP: A deep representation learning approach for identification, and the role of environment on

merger incidence. A&A 679, A142, A142. DOI: 10.1051/0004-6361/202346743. arXiv: 2309.15539 [astro-ph.GA].

Ostriker, J. P. (Jan. 1980). Elliptical Galaxies are not Made by Merging Spiral Galaxies. *Comments on Astrophysics* 8, p. 177.

Park, Minjung et al. (2023). Rapid Quenching of Galaxies at Cosmic Noon. *The Astrophysical Journal* 953.1, p. 119. DOI: 10.3847/1538-4357/acd54a. URL: https://dx.doi.org/10.3847/1538-4357/acd54a.

Patton, D. R. et al. (Jan. 1997). Close Pairs of Field Galaxies in the CNOC1 Redshift Survey. ApJ 475.1, pp. 29–42. DOI: 10.1086/303535. arXiv: astro-ph/9608016 [astro-ph].

Patton, D. R. et al. (Jan. 2002). Dynamically Close Galaxy Pairs and Merger Rate Evolution in the CNOC2 Redshift Survey. ApJ 565.1, pp. 208–222. DOI: 10.1086/324543. arXiv: astro-ph/0109428 [astro-ph].

Patton, D. R. et al. (Nov. 2005). A Hubble Space Telescope Snapshot Survey of Dynamically Close Galaxy Pairs in the CNOC2 Redshift Survey. AJ 130.5, pp. 2043–2057. DOI: 10.1086/491672. arXiv: astro-ph/0507417 [astro-ph].

Pearson, W. J. et al. (Nov. 2019a). Effect of galaxy mergers on star-formation rates. A&A 631, A51, A51. DOI: 10.1051/0004-6361/201936337. arXiv: 1908.10115 [astro-ph.GA].

Pearson, W. J. et al. (June 2019b). Identifying galaxy mergers in observations and simulations with deep learning. A&A 626, A49, A49. DOI: 10.1051/0004-6361/201935355. arXiv: 1902.10626 [astro-ph.GA].

Pearson, W. J. et al. (May 2022). North Ecliptic Pole merging galaxy catalogue. A&A 661, A52, A52. DOI: 10.1051/0004-6361/202141013. arXiv: 2202.10780 [astro-ph.GA].

Pearson, W. J. et al. (Apr. 2024a). Determining the time before or after a galaxy merger event. *arXiv e-prints*, arXiv:2404.11166, arXiv:2404.11166. DOI: 10.48550/arXiv.2404.11166. arXiv: 2404.11166 [astro-ph.GA].

Pearson, W. J. et al. (Mar. 2024b). Effects of galaxy environment on merger fraction. *arXiv e-prints*, arXiv:2403.11615, arXiv:2403.11615. DOI: 10.48550/arXiv.2403.11615. arXiv: 2403.11615 [astro-ph.GA].

Penzias, A. A. and R. W. Wilson (July 1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. ApJ 142, pp. 419–421. DOI: 10.1086/148307.

Petrosian, V. (Dec. 1976). Surface Brightness and Evolution of Galaxies. ApJ 210, p. L53. DOI: 10.1086/182301.

Pfleiderer, J. (Jan. 1963). Gravitationseffekte bei der Begegnung zweier Galaxien. Mit 5 Textabbildungen. ZAp 58, p. 12.

Pistis, F. et al. (Mar. 2024). A comparative study of the fundamental metallicity relation. The impact of methodology on its observed evolution. A&A 683, A203, A203. DOI: 10.1051/0004-6361/202346943. arXiv: 2312.00930 [astro-ph.GA].

Planck Collaboration et al. (Sept. 2020). Planck 2018 results. VI. Cosmological parameters. A&A 641, A6, A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

Press, William H. and Paul Schechter (Feb. 1974). Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. ApJ 187, pp. 425–438. DOI: 10.1086/152650.

Quinn, P. J. (Apr. 1984). On the formation and dynamics of shells around elliptical galaxies. ApJ 279, pp. 596–609. DOI: 10.1086/161924.

Reiprich, Thomas H. et al. (Aug. 2013). Outskirts of Galaxy Clusters. Space Sci. Rev. 177.1-4, pp. 195–245. DOI: 10.1007/s11214-013-9983-8. arXiv: 1303.3286 [astro-ph.CO].

Riccio, G. et al. (Sept. 2021). Preparing for LSST data. Estimating the physical properties of $z < 2.5$ main-sequence galaxies. A&A 653, A107, A107. DOI: 10.1051/0004-6361/202140854. arXiv: 2106.12573 [astro-ph.GA].

Rodrigues, Myriam et al. (Jan. 2018). Testing the hierarchical assembly of massive galaxies using accurate merger rates out to z 1.5. *Monthly Notices of the Royal Astronomical Society* 475.4, pp. 5133–5143. ISSN: 0035-8711. DOI: 10.1093/mnras/sty098. eprint: https://academic.oup.com/mnras/article-pdf/475/4/5133/24111478/sty098.pdf. URL: https://doi.org/10.1093/mnras/sty098.

Rodriguez, Juan D. et al. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.3, pp. 569–575. DOI: 10.1109/TPAMI.2009.187.

Rodriguez-Gomez, Vicente et al. (May 2016). The stellar mass assembly of galaxies in the Illustris simulation: growth by mergers and the spatial distribution of accreted stars. MNRAS 458.3, pp. 2371–2390. DOI: 10.1093/mnras/stw456. arXiv: 1511.08804 [astro-ph.GA].

Rodriguez-Gomez, Vicente et al. (Mar. 2019). The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to Pan-STARRS observations. MNRAS 483.3, pp. 4140–4159. DOI: 10.1093/mnras/sty3345. arXiv: 1809.08239 [astro-ph.GA].

Rutherford, Tomas H. et al. (Apr. 2024). The SAMI Galaxy Survey: using tidal streams and shells to trace the dynamical evolution of massive galaxies. MNRAS 529.2, pp. 810–830. DOI: 10.1093/mnras/stae398. arXiv: 2402.02728 [astro-ph.GA].

Salpeter, Edwin E. (Jan. 1955). The Luminosity Function and Stellar Evolution. ApJ 121, p. 161. DOI: 10.1086/145971.

Sanders, D. B. and I. F. Mirabel (Jan. 1996). Luminous Infrared Galaxies. ARA&A 34, p. 749. DOI: 10.1146/annurev.astro.34.1.749.

Sazonova, Elizaveta et al. (Apr. 2024). RMS asymmetry: a robust metric of galaxy shapes in images with varied depth and resolution. *arXiv e-prints*, arXiv:2404.05792, arXiv:2404.05792. DOI: 10.48550/arXiv.2404.05792. arXiv: 2404.05792 [astro-ph.GA].

Schindler, S. and A. Diaferio (Feb. 2008). Metal Enrichment Processes. Space Sci. Rev. 134.1-4, pp. 363–377. DOI: 10.1007/s11214-008-9321-8. arXiv: 0801.1061 [astro-ph].

Schlafly, E. F. et al. (Sept. 2012). Photometric Calibration of the First 1.5 Years of the Pan-STARRS1 Survey. ApJ 756.2, 158, p. 158. DOI: 10.1088/0004-637X/756/2/158. arXiv: 1201.2208 [astro-ph.IM].

Schmidt, M. (Mar. 1963). 3C 273 : A Star-Like Object with Large Red-Shift. Nature 197.4872, p. 1040. DOI: 10.1038/1971040a0.

Scott, Caroline and Sugata Kaviraj (Jan. 2014). Star formation and AGN activity in interacting galaxies: a near-UV perspective. MNRAS 437.3, pp. 2137–2145. DOI: 10.1093/mnras/stt2014. arXiv: 1310.5148 [astro-ph.GA].

Smith, R. W. and R. Berendzen (Jan. 1982). Book-Review - the Expanding Universe - Astronomy's Great Debate 1900-1931. *Journal for the History of Astronomy* 13, p. 209.

Snyder, Gregory F. et al. (June 2015). Diverse structural evolution at z
gt; 1 in cosmologically simulated galaxies. *Monthly Notices of the Royal Astronomical Society* 451.4, pp. 4290–4310. ISSN: 0035-8711. DOI: 10.1093/mnras/stv1231. eprint: https://academic.oup.com/mnras/article-pdf/451/4/4290/3891442/stv1231.pdf. URL: https://doi.org/10.1093/mnras/stv1231.

Snyder, Gregory F. et al. (Dec. 2015). Galaxy morphology and star formation in the Illustris Simulation at z = 0. MNRAS 454.2, pp. 1886–1908. DOI: 10.1093/mnras/stv2078. arXiv: 1502.07747 [astro-ph.GA].

Sofue, Yoshiaki and Vera Rubin (Jan. 2001). Rotation Curves of Spiral Galaxies. ARA&A 39, pp. 137–174. DOI: 10.1146/annurev.astro.39.1.137. arXiv: astro-ph/0010594 [astro-ph].

Sola, Elisabeth et al. (June 2022). Characterization of low surface brightness structures in annotated deep images. A&A 662, A124, A124. DOI: 10.1051/0004-6361/202142675. arXiv: 2203.03973 [astro-ph.GA].

Somerville, Rachel S. and Romeel Davé (Aug. 2015). Physical Models of Galaxy Formation in a Cosmological Framework. ARA&A 53, pp. 51–113. DOI: 10.1146/annurev-astro-082812-140951. arXiv: 1412.2712 [astro-ph.GA].

Srivastava, Nitish et al. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15.56, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2, pp. 111–147. ISSN: 00359246. URL: http://www.jstor.org/stable/2984809 (visited on 05/31/2022).

Stoughton, Chris et al. (Jan. 2002). Sloan Digital Sky Survey: Early Data Release. AJ 123.1, pp. 485–548. DOI: 10.1086/324741.

Suelves, L. E. et al. (Jan. 2023). Merger identification through photometric bands, colours, and their errors. A&A 669, A141, A141. DOI: 10.1051/0004-6361/202244509. arXiv: 2211.07489 [astro-ph.GA].

Sureshkumar, U. et al. (Feb. 2024b). Do galaxy mergers prefer under-dense environments? *arXiv e-prints*, arXiv:2402.18520, arXiv:2402.18520. DOI: 10.48550/arXiv.2402.18520. arXiv: 2402.18520 [astro-ph.GA].

— (June 2024a). Do galaxy mergers prefer under-dense environments? A&A 686, A40, A40. DOI: 10.1051/0004-6361/202347705. arXiv: 2402.18520 [astro-ph.GA].

Takamiya, M. (1999). Galaxy Structural Parameters: Star Formation Rate and Evolution with Redshift. *The Astrophysical Journal Supplement Series* 122.1, pp. 109–150. DOI: 10.1086/313216. URL: https://doi.org/10.1086/313216.

Tamm, Antti et al. (July 2007). Visible and dark matter in M31 - I. Properties of stellar components. *arXiv e-prints*, arXiv:0707.4375, arXiv:0707.4375. DOI: 10.48550/arXiv.0707.4375. arXiv: 0707.4375 [astro-ph].

Tasca, L. A. M. et al. (May 2014). Evidence for major mergers of galaxies at $2 \lesssim z < 4$ in the VVDS and VUDS surveys. A&A 565, A10, A10. DOI: 10.1051/0004-6361/201321507. arXiv: 1303.4400 [astro-ph.CO].

Taylor, M. B. (Dec. 2005). TOPCAT & STIL: Starlink Table/VOTable Processing Software. *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell et al. Vol. 347. Astronomical Society of the Pacific Conference Series, p. 29.

Tempel, Elmo et al. (July 2007). Visible and dark matter in M 31 - II. A dynamical model and dark matter density distribution. *arXiv e-prints*, arXiv:0707.4374, arXiv:0707.4374. DOI: 10.48550/arXiv.0707.4374. arXiv: 0707.4374 [astro-ph].

Thuruthipilly, H. et al. (Feb. 2024). Shedding light on low-surface-brightness galaxies in dark energy surveys with transformer models. A&A 682, A4, A4. DOI: 10.1051/0004-6361/202347649. arXiv: 2310.13543 [astro-ph.GA].

Thuruthipilly, Hareesh et al. (Aug. 2022). Finding strong gravitational lenses through self-attention. Study based on the Bologna Lens Challenge. A&A 664, A4, A4. DOI: 10.1051/0004-6361/202142463. arXiv: 2110.09202 [cs.CV].

Tonry, J. L. et al. (May 2012). The Pan-STARRS1 Photometric System. ApJ 750.2, 99, p. 99. DOI: 10.1088/0004-637X/750/2/99. arXiv: 1203.0297 [astro-ph.IM].

Toomre, Alar and Juri Toomre (Dec. 1972). Galactic Bridges and Tails. ApJ 178, pp. 623–666. DOI: 10.1086/151823.

Tous, J. L. et al. (Jan. 2023). The Origin of Star-forming Rings in S0 Galaxies. ApJ 942.1, 48, p. 48. DOI: 10.3847/1538-4357/aca484. arXiv: 2211.09697 [astro-ph.GA].

Trimble, Virginia (Dec. 1995). The 1920 Shapley-Curtis Discussion: Background, Issues, and Aftermath. PASP 107, p. 1133. DOI: 10.1086/133671.

Trujillo, Ignacio and Jüergen Fliri (June 2016). Beyond 31 mag arcsec^{-2}: The Frontier of Low Surface Brightness Imaging with the Largest Optical Telescopes. ApJ 823.2, 123, p. 123. DOI: 10.3847/0004-637X/823/2/123. arXiv: 1510.04696 [astro-ph.GA].

Tumlinson, Jason et al. (Aug. 2017). The Circumgalactic Medium. ARA&A 55.1, pp. 389–432. DOI: 10.1146/annurev-astro-091916-055240. arXiv: 1709.09180 [astro-ph.GA].

Tyson, J. A. (Jan. 1990). The shift-and-stare technique and a large area CCD mosaic. *CCDs in astronomy*. Ed. by George H. Jacoby. Vol. 8. Astronomical Society of the Pacific Conference Series, pp. 1–10.

Van Der Maaten, Laurens et al. (2009). Dimensionality reduction: a comparative review. *J Mach Learn Res* 10, pp. 66–71.

Vega-Ferrero, J. et al. (Sept. 2021). Pushing automated morphological classifications to their limits with the Dark Energy Survey. MNRAS 506.2, pp. 1927–1943. DOI: 10.1093/mnras/stab594. arXiv: 2012.07858 [astro-ph.GA].

Vulpiani, Angelo (2014). *Lewis Fry Richardson: scientist, visionary and pacifist*. Springer.

Walmsley, Mike et al. (Oct. 2019). Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society* 491.2, pp. 1554–1574. ISSN: 0035-8711. DOI: 10.1093/mnras/stz2816. eprint: https://academic.oup.com/mnras/article-pdf/491/2/1554/31144873/stz2816.pdf. URL: https://doi.org/10.1093/mnras/stz2816.

Walmsley, Mike et al. (Mar. 2019). Identification of low surface brightness tidal features in galaxies using convolutional neural networks. MNRAS 483.3, pp. 2968–2982. DOI: 10.1093/mnras/sty3232. arXiv: 1811.11616 [astro-ph.GA].

Walmsley, Mike et al. (Jan. 2022). Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. MNRAS 509.3, pp. 3966–3988. DOI: 10.1093/mnras/stab2093. arXiv: 2102.08414 [astro-ph.GA].

Walmsley, Mike et al. (Dec. 2023). Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys. MNRAS 526.3, pp. 4768–4786. DOI: 10.1093/mnras/stad2919. arXiv: 2309.11425 [astro-ph.GA].

Wang, L. et al. (Dec. 2020). Towards a consistent framework of comparing galaxy mergers in observations and simulations. A&A 644, A87, A87. DOI: 10.1051/0004-6361/202038084. arXiv: 2009.02974 [astro-ph.GA].

Waters, C. Z. et al. (Nov. 2020). Pan-STARRS Pixel Processing: Detrending, Warping, Stacking. ApJS 251.1, 4, p. 4. DOI: 10.3847/1538-4365/abb82b. arXiv: 1612.05245 [astro-ph.IM].

Watkins, Aaron E. et al. (Mar. 2024). Strategies for optimal sky subtraction in the low surface brightness regime. MNRAS 528.3, pp. 4289–4306. DOI: 10.1093/mnras/stae236. arXiv: 2401.12297 [astro-ph.GA].

West, Andrew A. et al. (Feb. 2010). H I-Selected Galaxies in the Sloan Digital Sky Survey. I. Optical Data. AJ 139.2, pp. 315–328. DOI: 10.1088/0004-6256/139/2/315. arXiv: 0910.4965 [astro-ph.CO].

West, Andrew Alan (Nov. 2005). HI selected galaxies in the Sloan Digital Sky Survey. PhD thesis. University of Washington, Seattle.

White, S. D. M. and M. J. Rees (May 1978). Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. MNRAS 183, pp. 341–358. DOI: 10.1093/mnras/183.3.341.

White, Simon D. M. and Carlos S. Frenk (Sept. 1991). Galaxy Formation through Hierarchical Clustering. ApJ 379, p. 52. DOI: 10.1086/170483.

Willett, Kyle W. et al. (Nov. 2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. MNRAS 435.4, pp. 2835–2860. DOI: 10.1093/mnras/stt1458. arXiv: 1308.3496 [astro-ph.CO].

Wolszczan, A. and D. A. Frail (Jan. 1992). A planetary system around the millisecond pulsar PSR1257 + 12. Nature 355.6356, pp. 145–147. DOI: 10.1038/355145a0.

York, Donald G. et al. (Sept. 2000). The Sloan Digital Sky Survey: Technical Summary. AJ 120.3, pp. 1579–1587. DOI: 10.1086/301513. arXiv: astro-ph/0006396 [astro-ph].

Ysard, N. et al. (Nov. 2019). From grains to pebbles: the influence of size distribution and chemical composition on dust emission properties. A&A 631, A88, A88. DOI: 10.1051/0004-6361/201936089. arXiv: 1909.05015 [astro-ph.GA].

Yu, Lean et al. (2005). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering* 18.2, pp. 217–230.

Zalesky, L. M. (Jan. 2021). The Hawaii Two-0 Twenty Square Degree Survey. *American Astronomical Society Meeting Abstracts*. Vol. 237. American Astronomical Society Meeting Abstracts, 215.05, p. 215.05.

Zeiler, Matthew D and Rob Fergus (Nov. 2013). Visualizing and Understanding Convolutional Networks. *arXiv e-prints*, arXiv:1311.2901, arXiv:1311.2901. DOI: 10.48550/arXiv.1311.2901. arXiv: 1311.2901 [cs.CV].

Zwicky, F. (Oct. 1937). On the Masses of Nebulae and of Clusters of Nebulae. ApJ 86, p. 217. DOI: 10.1086/143864.