

UNIVERSITY OF THE WESTERN CAPE

Anomaly Detection With Machine Learning In Astronomical Images



A thesis submitted in fulfilment for the
degree of Master of Science

in the
Centre for Radio Cosmology
Department of Physics and Astronomy

March 2022

Declaration of Authorship

I, VERLON ETSEBETH, declare that this thesis titled, 'ANOMALY DETECTION WITH MACHINE LEARNING IN ASTRONOMICAL IMAGES' 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 this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University 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: _____

Vetsebeth

Date: 10 March 2022

“Once you’ve got a task to do, it’s better to do it than live with the fear of it.”

Joe Abercrombie



UNIVERSITY *of the*
WESTERN CAPE

Abstract

Centre for Radio Cosmology
Department of Physics and Astronomy

Master of Science

by Verlon Etsebeth

Observations that push the boundaries have historically fuelled scientific breakthroughs, and these observations frequently involve phenomena that were previously unseen and unidentified. Data sets have increased in size and quality as modern technology advances at a record pace. Finding these elusive phenomena within these large data sets becomes a tougher challenge with each advancement made. Fortunately, machine learning techniques have proven to be extremely valuable in detecting outliers within data sets. Astronomy is a framework that utilises machine learning techniques for anomaly detection in astronomy and incorporates active learning to provide target specific results. It is used here to evaluate whether machine learning techniques are suitable to detect anomalies within the optical astronomical data obtained from the Dark Energy Camera Legacy Survey. Using the machine learning algorithm isolation forest, Astronomy is applied on subsets of the Dark Energy Camera Legacy Survey (DECaLS) data set. The pre-processing stage of Astronomy had to be significantly extended to handle real survey data from DECaLS, with the changes made resulting in up to 10% more sources having their features extracted successfully. For the top 500 sources returned, 292 were ordinary sources, 86 artefacts and masked sources and 122 were interesting anomalous sources. A supplementary machine learning algorithm known as active learning enhances the identification probability of outliers in data sets by making it easier to identify target specific sources. The addition of active learning further increases the amount of interesting sources returned by almost 40%, with 273 ordinary sources, 56 artefacts and 171 interesting anomalous sources returned. Among the anomalies discovered are some merger events that have been successfully identified in known catalogues and several candidate merger events that have not yet been identified in the literature. The results indicate that machine learning, in combination with active learning, can be effective in detecting anomalies in actual data sets. The extensions integrated into Astronomy pave the way for its application on future surveys like the Vera C. Rubin Observatory Legacy Survey of Space and Time.

Acknowledgements

National Research Foundation

We acknowledge support from South African Radio Astronomy Observatory and the National Research Foundation (NRF) towards this research. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

Personal Thanks

I would like to thank my supervisor, Dr Michelle Lochner, for her patience and guidance through each step of the journey and without whom this project could not have been done. It has been a long journey, but one that I have thoroughly enjoyed.

A special thanks to the Centre for Radio Cosmology, for accepting me and granting me this wonderful opportunity. I would also like to thank Dr Michael Walmsley for his input with regard to Dr Lochner's Astronomy programme.

DECaLS

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; Proposal ID 2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NOAO Prop. ID 2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z-band Legacy Survey (MzLS; Prop. ID 2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Center for Cosmology and Astro-Particle Physics at the Ohio State


University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas AM University, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenossische Technische Hochschule (ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciencies de l'Espai (IEEC/CSIC), the Institut de Fisica d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig Maximilians Universitat Munchen and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas AM University.

BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program "The Emergence of Cosmological Structures" Grant XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. The BASS is also supported by the External Cooperation Program of Chinese Academy of Sciences (Grant 114A11KYSB20160057), and Chinese National Natural Science Foundation (Grant 11433005).

The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration.

The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

Contents

Declaration of Authorship	i
Abstract	iii
Acknowledgements	iv
List of Figures	ix
List of Tables	xi
Abbreviations	xii
	
1 Making Discoveries In Optical Surveys	1
1.1 Introduction	1
1.2 Overview Of Galaxy Morphology And Evolution	3
1.2.1 Galaxy Morphologies And Classification	3
1.2.2 Basic Concepts Of Galaxy Evolution	7
1.3 Scientific Discoveries With Anomalous Sources	9
1.3.1 Galaxy Mergers	10
1.3.2 Gravitational Lenses	12
1.4 Large Optical Surveys	15
2 An Overview Of Machine Learning	18
2.1 Introduction	18
2.1.1 Basic Procedure Of Machine Learning Algorithms	19
2.2 Machine Learning Categories	19
2.2.1 Supervised Learning	20
2.2.2 Reinforcement Learning	21
2.2.3 Semi-supervised Learning	21
2.2.4 Unsupervised Learning	22
2.2.4.1 Outlier Detection Algorithms	23
2.2.5 Active Learning	24
2.3 Isolation Forest	24
2.3.1 Algorithm	26
2.3.2 Anomaly Score	28

2.4	Determining The Performance Of Machine Learning Algorithms	29
2.4.1	Rank Weighted Score	32
2.5	Machine Learning Applications In Astronomy	33
2.5.1	Advantages And Disadvantages of Machine Learning	37
3	Methodology I: Applying Base Astronomy to DECaLS Data	39
3.1	Introduction	39
3.2	Brief Overview Of The Legacy Surveys	40
3.2.1	Sky Coverage Of The Legacy Surveys	40
3.2.2	Depths Of The Legacy Surveys	42
3.2.3	DECaLS Data	44
3.2.4	DECaLS Data Format	44
3.2.4.1	Data Bricks	45
3.2.4.2	Data Cutouts	46
3.3	General Introduction to Astronomy	46
3.3.1	General Steps Followed When Applying Astronomy	47
3.4	Procedure Used To Apply Astronomy On DECaLS	50
3.4.1	Accessing Data And Setting Up The Parameters To Use	50
3.4.2	Image Processing Techniques Applied To The DECaLS Data	50
3.4.2.1	Image Transform: Scale	51
3.4.2.2	Image Transform: Axis Shift	51
3.4.2.3	Image Transform: Greyscale	52
3.4.2.4	Image Transform: Sigma Clipping	52
3.4.3	Ellipse Fitting Feature Extraction Method For Optical Data	53
3.4.4	Machine Learning	56
3.4.5	Astronomy: Frontend	56
3.5	How And Where The Feature Extraction Process Fails	57
3.5.1	Feature Extraction Failures Due To Sources	57
3.5.2	Feature Extraction Failures Due To Image Problems	59
4	Methodology II: Extending Astronomy	62
4.1	Introduction	62
4.2	Data Format Adjustments And Selection Cuts	64
4.3	Changes Made For Image Based Failures	65
4.3.1	Ellipse Fitting Errors Due To Small Image Sizes	66
4.3.2	Implementing Adaptive Image Scaling In Astronomy	67
4.3.3	Ellipse Fitting Errors Due To Large Image Sizes	70
4.4	Changes Made For Source Based Failures	71
4.4.1	Effect Of Nearby Bright Sources And Masked Sources	71
4.4.2	Faint Sources Causing Failures	72
4.4.3	The Impact of Band Weightings On Images	73
4.4.3.1	Visual Inspections Of Different Band Weightings	75
4.4.3.2	Single Channel Benefits For Feature Extraction	76
4.4.3.3	How Different Weightings Affect Feature Extraction	77
4.4.3.4	Finding An Optimal Function	79
4.4.3.5	Discussion On The Band Weightings Used	81
4.5	Limitations Of Outlier Detection Algorithms	82

4.5.1	iForest	83
4.5.2	Local Outlier Factor	86
4.5.3	RWS - Variations Between Runs	89
4.6	Improvements Due To The Changes Made	95
5	Results of Applying Improved Astronomy On DECaLS Data	98
5.1	Introduction	98
5.2	Data Sets Investigated	98
5.3	Testing Procedures And Performance Measurements	100
5.4	Impact Of Adjustments Made To Astronomy	101
5.5	Improved Performance From Active Learning	104
5.6	Interesting Sources Identified	108
6	Discussion	114
6.1	Changes Made To Astronomy	115
6.2	Active Learning	116
6.3	Computational Performance Of Algorithms Used	118
6.4	Performance On The DECaLS Data	118
6.5	Future Work	120

Bibliography		122
---------------------	--	------------



List of Figures

1.1	Examples of different morphologies	5
1.2	Hubble Tuning Fork diagram	6
1.3	Example of a galaxy merger	12
1.4	Basic concept of gravitational lensing	14
2.1	Basic Decision Tree	25
2.2	Steps of iForest	27
2.3	Confusion Matrix	30
3.1	DESI magnitude depth colour map	42
3.2	Histograms of band depths	43
3.3	Front end of Astronomy	49
3.4	Sigma clipping example	53
3.5	Example of contours and ellipses fit to a source.	55
3.6	Feature extraction failures due to bright sources nearby.	58
3.7	Feature extraction failures due to masked sources.	58
3.8	Feature extraction failures due to faint sources.	59
3.9	Feature extraction problems caused by open ellipses.	60
3.10	Feature extraction problems caused by large images.	61
4.1	Example of varying angular diameters of sources	63
4.2	Results from upscaling function	70
4.3	Different band weightings.	76
4.4	Comparison between the two display functions	79
4.5	Grid search ellipses found and RWS results	80
4.6	Representation of the Gaussian distributions	82
4.7	iForest runtime and memory usage for number of sources	83
4.8	iForest performance for number of sources	84
4.9	iForest runtime and memory usage for dimensions	85
4.10	iForest performance for dimensions	85
4.11	LOF runtime and memory usage for number of sources	86
4.12	LOF performance for number of sources	87
4.13	iForest runtime and memory usage for dimensions	88
4.14	LOF performance for dimensions	88
4.15	RWS for each parameter	92
4.16	Number of anomalies for each run	93
4.17	Recall and precision values for each run	94
4.18	Improvements from changes made	96

5.1 Histogram showing the increased features fit successfully 102

5.2 Base Astronomy versus the adaptations made 103

5.3 Base Astronomy versus the adaptations made for the top 2000 sources . 103

5.4 Active learning increments versus bulk training 105

5.5 Histogram of RWS for the top 2000 sources of the Lens Set 106

5.6 Active learning compared to machine learning only 107

5.7 Histograms of the RWS values for all of the methods 108

5.8 Cross matched images of galaxy mergers 110

5.9 Sources similar to the galaxy merger catalogue 112

5.10 Issues with the Bright subset 113



List of Tables

3.1	Bands used and sky coverage of the Legacy Surveys	41
4.1	Adaptive Scaling results	69
4.2	Varying scores obtained from iForest	90
4.3	Varying scores obtained when using different parameters	91
4.4	Increase in the number of sources detected	97
5.1	Sources that overlap from the galaxy merger catalogue	109
5.2	Sources similar to those in the galaxy merger catalogue	111



Abbreviations

BASS	Beijing- Arizona Sky Survey
CCD	Charged Coupled Devices
CNNs	convolutional neural networks
Dec	Declination
DECaLS	Dark Energy Camera Legacy Survey
DES	Dark Energy Survey
DESI	Dark Energy Spectroscopic Instrument
DR8	Data Release 8
EMD	Earth Mover's Distance
FITS	Flexible Image Transport System
HST	Hubble Space Telescope
GANs	General Adversarial Networks
HSC-SSP	Hyper Suprime-Cam Subaru Strategic Program
Λ CDM	Cold Dark Matter
LOF	Local Outlier Factor
LSST	Legacy Survey of Space and Time
MzLS	Mayall z-band Legacy Survey
RA	Right Ascension
ROC-curves	Receiver Operating Characteristic curves
RWS	Rank Weighted Score
SDSS	Sloan Digital Sky Survey
SNR	Signal to Noise Ratio
SVM	Support to Vector Machine
t-SNE	t-Distributed Stochastic Neighbour Embedding
WISE	Wide-field Infrared Survey Explorer

Dedicated to my family, for always standing by my side and supporting me through the good times and the bad times.



UNIVERSITY *of the*
WESTERN CAPE

Chapter 1

Making Discoveries In Optical Surveys

1.1 Introduction

Astronomical surveys are key to making new scientific discoveries, since they contain the observational data needed for these discoveries. All of the information that is analysed, processed and interpreted is included in the data obtained from the surveys. Technical innovation has always been a driving force in pushing the boundaries of science, perhaps more so in Astronomy than in any other field due to the vast increases in observations made with improved technology. Astronomy is currently experiencing a major increase in new surveys being undertaken. The amount of data that some of these surveys are expected to produce has increased rapidly and, with it, the potential for unexpected discoveries [1]. However, a disadvantage of having so much data is that the few key sources required to make new scientific discoveries can easily be overlooked.

New telescopes are designed with specific scientific goals in mind, but they collect data far beyond what is required for these goals alone. Discoveries thus tend to be made in two separate ways. The first is due to the goals as set out for the telescope, with the discovery being made directly as a result of the technology used and the methods applied. The second stems from alternative methods and searches applied to the collected data with the idea of mining it for new discoveries. These tend to be different in nature to the goals set out for the survey itself, but are often important in their own right. An example of this was the application of an outlier detection machine learning algorithm on spectra from the Sloan Digital Sky Survey (SDSS) [2], which found numerous new anomalous galaxies based on their spectra [3].

Surprisingly though, a significant number of important discoveries in astronomy have been made unexpectedly. While scientific goals are set for telescopes, it sometimes happens that new discoveries are made unexpectedly along the way. A well known example of this is the accidental discovery of pulsars [4]. Studies done also indicate that about half of the scientific discoveries made in astronomy are unexpected and arise from scouring data sets for purposes other than that set out for the telescope [5, 6].

It is crucial that the data obtained from a survey is fully scrutinised using various methods to maximise the potential to make new discoveries. Increased data volumes and survey sensitivities increases the number of discoveries that can *potentially* be made, but with an increase in volume and complexity comes an increase in difficulty to detect the required data to make the discoveries.

Astronomy is already at the point where it is no longer feasible to mine such large data sets manually. Even citizen science projects such as Galaxy Zoo [7], which utilises thousands of people, are still unable to mine large scale surveys effectively. This problem is not unique to current and upcoming surveys, but exist within older data sets as well which have not been completely examined for discoveries. Often, the majority of the techniques applied to these data sets are ones available around the time that the data sets were published, yet newer techniques could provide valuable new insights. An example of this is a fairly recent scientific discovery made in 2018. A group of astronomers discovered that there is a possibility that thousands of black holes are likely to exist near the galactic centre of the Milky Way [8]. This is possibly a new discovery but makes use of data that is 20 years old. The need for more advanced and complex methods that are capable of handling large amounts of data and finding the relevant data required for discoveries is clear.

Such a solution lies with the application of *machine learning*. Although the concept of computer intelligence and machine learning is not new [9], it is only due to recent advances in computational capabilities, availability of data and development of novel approaches such as deep learning that machine learning has become more common [10]. Machine learning is applied on optical astronomical data with the aim of detecting interesting and anomalous sources that provide better opportunities to make scientific discoveries.

Before diving in to the details of machine learning and its application on the data set, a basic understanding of galaxies is required and is covered in section 1.2. This section explains why it is important to find anomalies such as gravitational mergers. Section 1.3 details some anomalous sources and what can be learned from them. The focus of the thesis utilises optical surveys, which are covered in more depth in section 1.4, wherein the need for more advanced techniques becomes clear. It should be noted that

the majority of the information covered within the following sections is obtained from the literature review by R. Buta [11], unless where otherwise indicated.

1.2 Overview Of Galaxy Morphology And Evolution

Galaxies are systems of stars, gas, dust, remnants of dead stars and dark matter all bound together gravitationally. Their characteristics and morphologies can differ significantly from each other. Not only do they vary in size and shape, but they also vary in age, colour, star formation rate and luminosity, among other things. The structure of galaxies is one of the basic ways in which these properties are depicted and through which the evolution of galaxies is determined.

Early observations of galaxies were not able to fully quantify all of these properties, however, they were able to identify visual distinctions between galaxies. From these visual differences, various classification systems were proposed for the different types of galaxies that were observed. The most successful classification system is that proposed by Edwin Hubble [12], which has proven to be mostly reliable under certain specifications and with some adjustments made. Classification is thus an ideal starting point in understanding galaxies and the properties that they typically possess.

Understanding galaxy evolution is one of the most active topics in extragalactic astronomy today. There are several indications that some important evolutionary steps include galaxy merger events [13]. In the early universe, galaxy mergers and collisions were much more common compared to the late universe [14]. This is simply because the early universe was smaller than the late universe and galaxies were much closer to each other. The overall matter density of the Universe was higher during the early universe. However, due to the large distances, observing the early universe and making accurate measurements can be difficult. Finding these merger events in order to study the role they play in galaxy evolution is therefore important.

This section starts off by covering the basic classification scheme, including adjustments made throughout history, in order to provide an understanding of different galaxy types. The second part of the section focuses on the connections between these classifications along with the basic theories of evolution and how merger events fit in with the process.

1.2.1 Galaxy Morphologies And Classification

The history of galaxy morphology predates the knowledge that galaxies are in fact extragalactic. First mentions of galaxy morphology even predates the telescope, with

some descriptions dating back as early as the 10th century by the Persian astronomer Abd al-Rahman al-Sufi [15]. Since then there have been numerous observations made that clearly distinguish galaxies from stars, based on the resolved structure that is visible. Yet throughout this time, even with advances in telescopes and observational techniques, the structures and morphology of galaxies have continued to be one of the most common ways to describe galaxies.

Although numerous improvements in classification and morphology were carried out over the years by the likes of Charles Messier, William Herschel and John Herschel, it was only with the invention of photography that the structures and morphology of galaxies could be studied in more detail. This led to the well known *Hubble classification*, first published in 1926 by Edwin Hubble. The classification, along with the *Tuning Fork* diagram, was published in the book, “The Realm of the Nebulae”, in 1936 [16]. This classification is still used today due to its simplicity and easier to apply.

Hubble classified the morphology of galaxies into three main classes: ellipticals, lenticulars and spirals, [12]. Along with these were a few irregular galaxies that did not quite fit in with the main classes. The spiral galaxy class has subdivisions for galaxies with bars and those without. Most of the nearby, bright galaxies were classified according to this scheme. The main classes of the Hubble classification system are as follows:

- **Spiral Galaxies:** also called disk galaxies and typically, but not always, contain a bulge that is reminiscent of an elliptical galaxy but with an outer, thinner disk of stars. This thin disk often contains spiral arms. *Barred Spirals* are spiral galaxies that contain a bar near the center.
- **Lenticular Galaxies:** are similar to spiral galaxies except that they do not contain any spiral arms.
- **Elliptical Galaxies:** are elliptical in shape and usually do not contain many features. The brightness tends to decrease the further out from the centre the galaxy is viewed. Unlike spiral galaxies where most of the stars rotate around the core in the same direction, the orbits of stars in elliptical galaxies vary significantly. Ellipticals are believed to be highly evolved galaxies created by galaxy merger events; either single, large events, or multiple smaller ones [17].



FIGURE 1.1: These three images illustrate the three different classes proposed by Edwin Hubble. The first image is a spiral galaxy, NGC 1376, with clear spiral arms and dust lanes (Acknowledgment: R. Thompson (University of Arizona)). The second is a lenticular, NGC 6861, which also contains dust lanes in this instance, but there are no spiral arms visible (Acknowledgement: J. Barrington). The last is that of an elliptical galaxy, NGC 4150, which appears much more uniform throughout (R.M. Crockett (University of Oxford, U.K.)). Image credit: ESA/Hubble NASA.

Any galaxy that did not quite fit into these classes was labelled as irregular. With improvements in technology and observations, it was found that the Hubble classification scheme was not sufficient to cover all types of galaxies observed. However, 90% of bright, luminous galaxies that are relatively close to the Milky Way fit into the Hubble classification scheme. Fainter dwarf galaxies do not fit into the scheme, and observations reveal that these dwarf galaxies greatly outnumber the luminous galaxies that do fall in line with the classification scheme. It is evident that an updated classification scheme is needed [18].

Figure 1.2 illustrates the Hubble Tuning Fork diagram. The lenticular galaxies are not clearly indicated on the diagram, although they are similar to the spiral galaxies. Irregular galaxies would also not fit into any specific group on their own as they contain all galaxies that do not fit in with the specified classes.

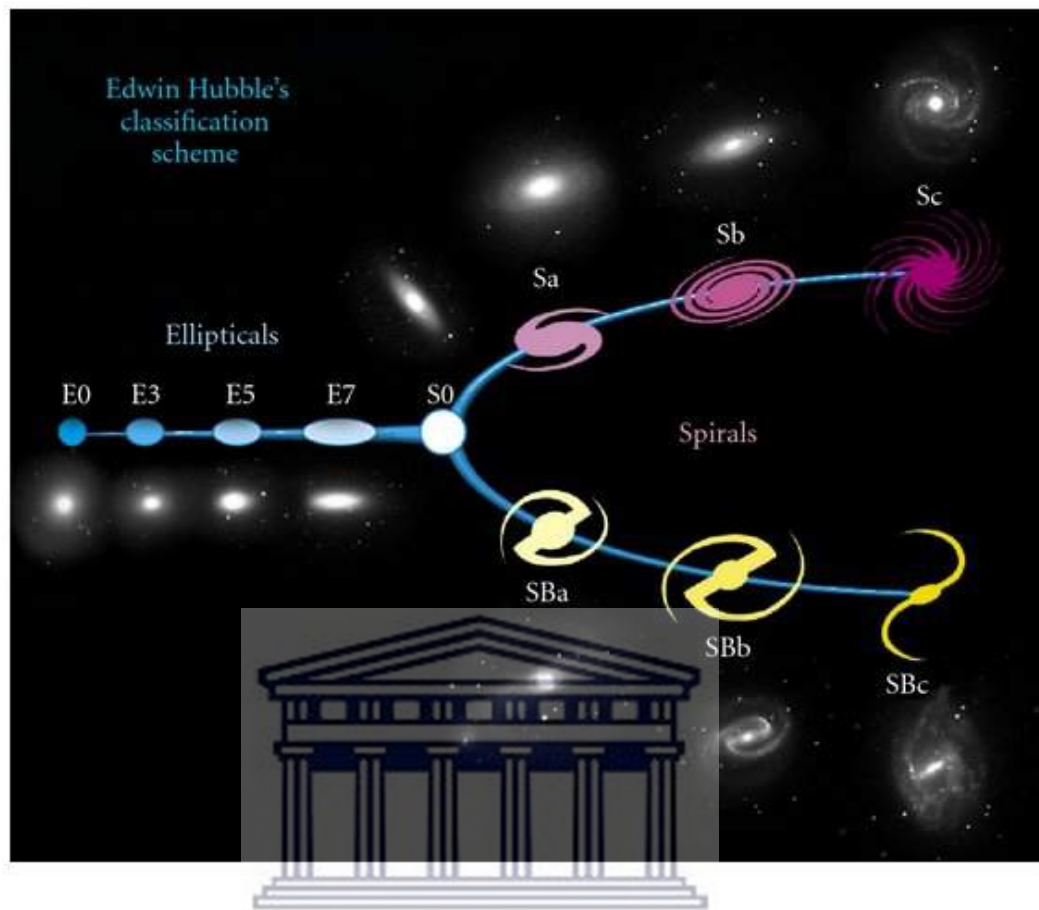


FIGURE 1.2: A replication of the original Hubble Tuning Fork Diagram. The distinction between spiral galaxies and elliptical galaxies is clear, although the evolutionary path is not. Image credit: NASA and ESA¹.

In 1959, Gerard de Vaucouleurs published a revised edition of Hubble's scheme that included additional visual properties that could be discerned with newer technology [19]. While classification, morphology and developments thereof are always useful, it does not provide any physical measurement as to why the galaxies differ. It is not clear which physical features of the galaxies are key in defining their morphology or their evolutionary path.

At around the same time, Erik Holmberg discovered that the morphologies of nearby galaxies were related to some physical properties of the galaxies [20]. Holmberg found that spiral galaxies tended to be smaller in mass than ellipticals and usually contained younger, brighter stars in active star formation regions. Ellipticals were typically found to be larger, redder and generally older than spirals, but not always [21]. This correlation

¹<https://en.wikipedia.org/w/index.php?title=File:HubbleTuningFork.jpg>

is an important part in understanding the process of galaxy formation and evolution as it makes a clear, physical distinction between the morphologies.

An important breakthrough in technology in the form of charged coupled devices (CCDs) allowed the light distributions of galaxies to be measured in a detailed way. de Vaucouleurs determined that massive elliptical galaxies have very similar light distributions, known today as the de Vaucouleurs profile [22]. Coan et al. utilised CCDs in conjunction with deep photographic images to reinforce the light distribution relation [23, 24]. This highlighted the fact that while galaxy morphologies vary, galaxies in each class were similar in nature.

The problem arises with the observations being done, which were mainly focused on nearby galaxies. A significant amount of effort was put into studying the local universe, but the early universe remained difficult to study. [25]. Even with modern technology, observing more than the basic structures and morphology of early universe galaxies remains a challenge. Telescopes have limited resolving power, thus even large galaxies at a high redshift would be harder to identify. The images of such galaxies have a poor resolution, making it difficult to identify unique features and characteristics. Distant galaxies are also fainter and redder due to the redshift, causing faint morphological features to be lost.

The basic classification of the morphology of a galaxy is thus still instrumental in understanding the properties of the galaxy. Simple classifications of distant galaxies allow it to be compared to nearby galaxies to gain insight into its properties. It is a key starting point when investigating galaxies as it provides information between different types of galaxies. Additionally, any theory that is created regarding galaxy evolution or formation will have to account for the different morphologies observed today.

Another key aspect of galaxy morphology is that it is strongly related to the galactic star formation history. The current morphology of a galaxy is dependent upon its star formation history and classification illustrates these properties directly. This can also be used to study the evolutionary paths of galaxies since changes in star formation rates usually occur due to external events occurring.

1.2.2 Basic Concepts Of Galaxy Evolution

It is known that galaxies evolve over large amounts of time. This is clearly seen by the rapid variations in the stellar mass density, the total solar masses (M_{\odot}) located within a unit volume, of galaxies located at the redshift range $1 < z < 3$, where almost half of the stellar mass is formed by redshift $z = 1$, [26]. The redshift of a galaxy is the shift of

the spectrum of an object towards the longer wavelengths, caused by the Doppler effect as the source recedes away from the observer due to the expansion of the Universe. The redshift increases with the distance of the object and can be considered to be a distance measurement. The star formation rate, the total mass of stars formed per year, usually in M_{\odot} per year, in the universe appears to peak at a redshift of $z = 2.5$, despite the large variations in star formation history of individual galaxies [27]. Although galaxies evolve over time, it is not clear what physical factors are the cause of this. Some of the theories of galaxy formation are as follows:

- **Top Down theories:** This was based on principles similar to stellar formation [28]. The basic theory is that disk galaxies formed through the collapse of gigantic gas clouds. In the early universe, matter was distributed in clumps consisting mostly of dark matter. These clumps interacted gravitationally, creating tidal forces that lead to increased angular momentum in the surrounding matter. The baryonic matter cooled and started contracting towards the centre, forming a disk. This disk broke into smaller sections which contracted individually to form stars and so the galaxy is formed.
- **Bottom Up theories:** The theory is that tiny quantum fluctuations occurred shortly after the Big Bang [29]. Matter started off in clumps formed by these fluctuations and merged with other matter due to gravitational forces. This resulted in disk shaped distributions which would contain dark matter halos similar to that which is observed today.
- **Λ CDM Model:** The current *standard model of cosmology*, the Λ CDM model, theorises that the universe was created in the Big Bang and is now composed of roughly 5% baryonic matter, 27% dark matter and 68% dark energy [30, 31]. The model is based on *inflation* and the *general theory of relativity* along with the *standard model of particle physics* [32, 33]. The Λ CDM model also assumes the universe is homogeneous and isotropic. The basis of structure formation within the model stems from gravitational collapses in over-dense regions, whereby structures grow and merge due to gravitational forces. However, an unsolved problem with the model is that it underestimates the number of thin disk galaxies that we see [34]. This is because the model predicts a large number of galaxy mergers which typically results in galaxies without a thin disk. There are some open questions in the Λ CDM model, such as those involving cosmological simulations that fail to accurately predict the population of galaxies observed today for example [35], but it is still the generally accepted cosmological model as it best explains current data.

The actual formation of galaxies is almost impossible to observe directly and is largely theorised. However, the evolution of galaxies can be observed as mentioned earlier. The different stages of a galaxy can be observed by investigating galaxies that are similar in nature, but which differ slightly. This allows a *timeline* of the evolution to be created by viewing the similar galaxies as different steps along the way. Most of this has been possible due to the Hubble Space Telescope (HST) and extensive surveys such as the SDSS [2].

Such surveys have allowed measurements to be made of the structures of galaxies, allowing the morphology to be used to determine its evolution. Modern space observations along with new ground based telescopes have allowed galaxies to be observed at much higher redshifts, [36]. These observations reveal that the early universe differs significantly from the late universe. They also reveal that there is a general trend stemming from small galaxies with high star formation rates in the early universe to much larger and quieter galaxies found in the late universe.

Investigations into the general trend suggests that it starts at smaller galaxies with active star forming disks which evolve into more massive ellipticals that lack significant star formation. However, galaxies do not just expand in size and mass. The main theory behind such an evolutionary path is that galaxies must interact with other galaxies at some point in their life time in order to grow to the much larger galaxies that we see in the late universe. These interactions are commonly seen in the early universe in the form of collisions between galaxies, sometimes even complete galaxy mergers that result in a single, larger galaxy. This theory is supported by the Λ CDM cosmology model that proposes a hierarchical method of structure formation through mergers of the dark matter halos around galaxies [37].

Not only can these collisions or merger events alter the appearance and properties of galaxies, they might actually be *required* in the formation of large elliptical galaxies that can be seen in the late universe. It is therefore crucial to make accurate observations of such events. It is estimated that roughly 10 - 20% of star-forming galaxies undergo some form of galaxy interactions [38]. This is only one of the anomaly types sought after in the thesis, in order to identify unknown galaxy collision and merger events to allow follow up investigations to be done into their properties.

1.3 Scientific Discoveries With Anomalous Sources

No two galaxies, stars or planets are the same, but given the vast quantity of each, it becomes possible to make classifications statistically. Finding a source that does not fit in

with the rest becomes an important scientific goal as these lead to a better understanding of the source in question. These objects can be considered to be “anomalies” in the sense that they do not conform to regular or typical patterns. However, there are anomalies that are known but are rare. As mentioned previously, such a rare, yet important, event that is of scientific significance is that of a galaxy collision or merger.

Another important anomaly that is of great interest and importance is that of a gravitational lens. Not only are they extremely rare, but they provide unique information that can not be determined in any other way known. Gravitational lenses allow a range of properties to be investigated, from dark matter to estimates of the Hubble constant.

This section covers these anomalies in more detail, starting off with galaxy merger events before moving on to gravitational lenses. A broad overview is given of each anomaly, from how they are formed to what significance they have. Additionally, any unusual or unidentifiable source would also be of interest, especially if they are unexpected.

1.3.1 Galaxy Mergers

Studies of distant galaxies indicate that galaxy interactions, collisions and mergers influence the size and shape of galaxies seen today, as covered in [11]. During galaxy collisions, the stars themselves do not get directly affected much due to the distances between them. Instead, their orbits change due to the changes in the gravitational forces that act on them. Collision events could be minor, creating a few tidal streams between galaxies, or they could be more involved. Observations and simulations indicate that mergers involving more massive galaxies, or a higher mass fraction, are usually more major than mergers involving less massive galaxies [39]. When two similarly sized galaxies collide it is referred to as a *galaxy merger* and usually results in a much larger, single galaxy. If the collision involves a galaxy that is much smaller than the other one, then the larger galaxy consumes the smaller one in a process called *galactic cannibalism*. In such an event, the larger galaxy could remain largely undisturbed by the entire event.

There are two separate merger types based on the galaxies that are involved within the merger event. The first type is called a “dry” merger, where the galaxies that collide have already progressed past the star formation period and are devoid of gas. These dry mergers typically result in the larger, elliptical galaxies as seen in the late universe. The second type is the “wet” merger, where the galaxies are relatively gas-rich, resulting in high star formation periods when they collide [40]. These mergers are typically responsible for changing a younger and bluer galaxy into an older, redder one.

The following are indications that galaxy collisions play an important role in galaxy evolution:

- Evidence can often be seen in very large elliptical galaxies, which often possess multiple nuclei, a strong indication of having consumed multiple galaxies.
- The shapes of peculiar galaxies are also caused by past interactions with other galaxies.
- Elliptical galaxies in the late universe are much larger than galaxies observed in the early universe. They must have gained additional material during their lifetime.

These are all important factors to consider in understanding galaxy evolution. It is not just the shape of the galaxy or galaxies that are affected by the collisions. For galaxies that contain significant interstellar matter regions, a collision can cause these regions to compress, triggering star formation. Estimates show that collisions can increase the star formation rate of the galaxies by as much as a factor of 10 [41]. Galaxies that show signs of this are called starburst galaxies. Starburst galaxies tend to be much brighter than normal galaxies and so they become easier to spot at large distances. Identifying merger events, or galaxies that have had recent interactions with other galaxies, is thus made easier by identifying starburst galaxies.

Younger galaxies, located a very large distance away at around 12 billion light years, resemble closely starburst galaxies that are involved in mergers [42]. Most of them have peculiar shapes with multiple nuclei and contain larger and brighter stars usually only found in regions with high star formation rates [43]. This clearly indicates a relation between galaxies that are currently merging, and younger galaxies that also appear to be merging. It is clear that galaxy interactions were much more common in the distant past and it is indicative that collisions and interactions are *required* to create the more evolved galaxies that we see today [44].

In the late universe, galaxy merger events are less common with only an estimated 4% of nearby bright galaxies showing some form of interaction [45]. These events can thus be considered to be rare and anomalous, yet they are vital in understanding the evolutionary paths of galaxies.

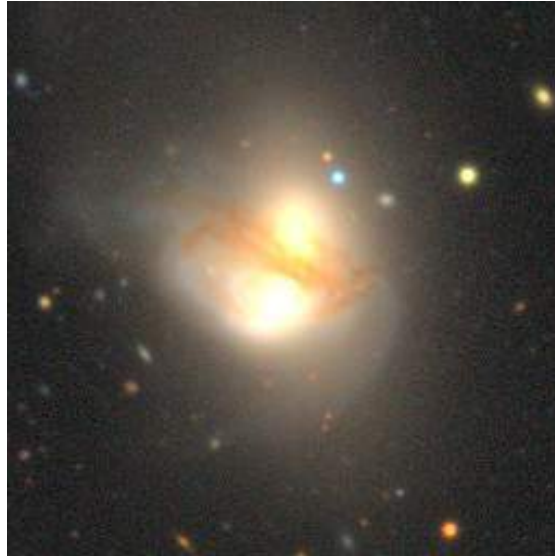


FIGURE 1.3: The image here is an example of a galaxy merger event that is a part of the Dark Energy Camera Legacy Survey (DECaLS) data set, also known as NGC0061². The two separate galactic nuclei are clearly visible in the image. It is also evident that a merger completely distorts the shape of both constituent galaxies, a property that will be utilised to locate them.

Figure 1.3 shows an example of a galaxy merger event. While it may be clear to classify the individual galaxies according to their morphology, it is unclear as to what the outcome will look like. Regions of high star formation rates are also seen at the parts where the galaxies make contact.

1.3.2 Gravitational Lenses

A gravitational lens, also referred to as the foreground source or object, is an object that is massive enough to visibly affect the path of light that emanates from a background source as it travels by the lens to the observer [46]. There are many factors that influence whether a lens can be seen. Higher mass systems are more likely to be detected by ground-based surveys such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST), as higher mass increases the lens cross-section. However, the likelihood of detecting a particular lens system is also a function of the geometry of the system as well as the seeing conditions during the observation [47].

Figure 1.4 illustrates the basic concept of gravitational lensing, where a massive source is located along the line of sight to a distant object. The light from the distant source travels along different paths around the lens, often resulting in multiple images of the same background source being observed if the lens is massive enough. Different variations

²<https://www.legacysurvey.org/viewer>

are possible whereby the foreground source is a cluster of galaxies or even a star in the case of microlensing. The different variations can be found as lenses, rings, arcs, series of arcs or a series of bright spots all depending on the alignment and viewing angle of the objects in question.

Gravitational lenses were first theorised by Albert Einstein more than a century ago and observed in 1919 by Frank Watson Dyson and Arthur Stanley Eddington when they measured the gravitational deflection of starlight passing near the Sun during a solar eclipse [48]. However, it was only in 1979 that the first extragalactic lens was observed by Dennis Walsh, Robert F. Carswell and Ray J. Weymann, due to advances in technology that made such observations possible [49]. This large gap between theory and observations is due to the nature of gravitational lenses themselves. The alignment of sources is crucial, as is the ability to observe such events.

There are several factors that affect the ability to observe lenses. First is the optical depth; which is a measure of the amount of absorption that occurs when light travels through an absorbing medium, measured as the ratio of incident to transmitted light so that a high value means less transmitted light. Because of the vast distances between the lens and the background object, the optical depth for the observation may be larger, making observations more challenging. Second is the actual distances involved, whereby the background source needs to be significantly more distant than the foreground lensing object in order for the lensing to take place. The other significant factor is the surface brightness of the source, which is the amount of apparent brightness per angular area of the source. The surface brightness of the background source is conserved in the lensing process, but dims with increased redshift as $(1 + z)^4$ due to the expansion of the Universe [50]. As such, lensed sources were too faint to observe until technology reached the sufficient capacity to detect them.

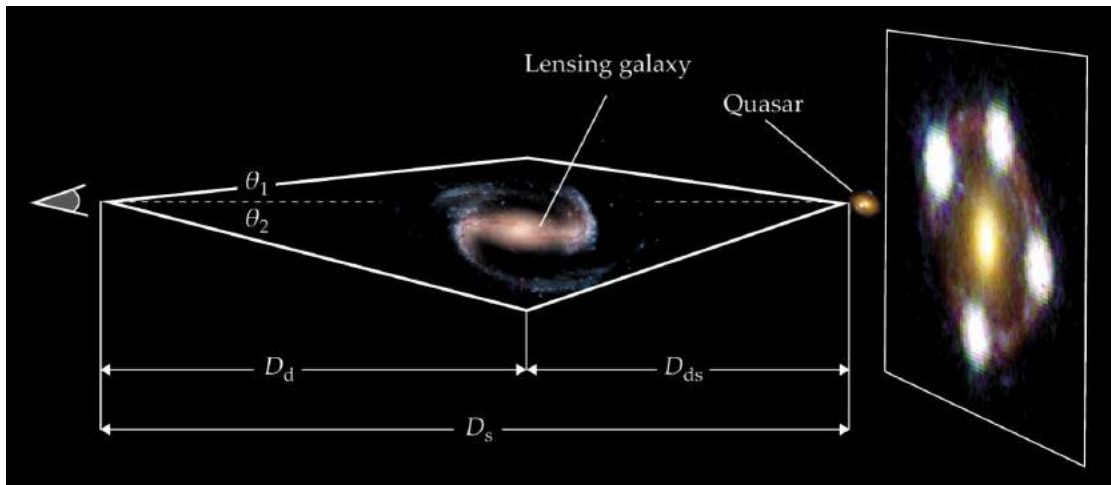


FIGURE 1.4: Example of gravitational lensing. In the image the different pathways that the light from the background source, the quasar, travels is illustrated. This is an instance of strong lensing where the background source is replicated multiples times from the observers point of view. Image credit: Freddie Pagani [51].

There are three different types of gravitational lensing observed today, each one varying by how the light is affected and what the foreground source consists of. Weak lensing occurs when the density of the lens, amongst other factors, is not high enough to bend the light so as to form multiple lenses. Weak lensing illustrates the *distribution of dark matter on large scales* and can be used to constrain the density of dark matter within a given volume [52]. Microlensing occurs when the background object, or background *and* foreground objects, are unresolved. As a result of this, the actual distortions are not observed directly. Instead, the brightness of the sources, stars within the Milky Way itself, is amplified temporarily as the objects align. In this instance, light curves are used to measure the distances as well as the motions of the objects in question. Microlensing is thus a transient anomaly and will not be covered in detail in the thesis. However, it is worth mentioning that microlensing is used to search for dark matter within the Milky Way itself, as well as to find extra-solar planets, since it is observed with stars within the Milky Way itself [53]. The last type is strong lensing, which is covered in more detail in the following section.

Strong Lensing

If the density of the lens is high enough, and if the geometry of the system allows, it will create multiple images of the background source, as seen in Figure 1.4. Strong lensing results in distinct visual features that are not seen elsewhere which, coupled with their rarity, causes them to be considered anomalous.

The light originates from a single source, but travels along different paths. These path variations cause time delays that depend on the geometry of the Universe and can be used to measure the expansion rate of the Universe, also known as the Hubble constant.

Elliptical galaxies are examples of good lens candidates since they are massive and compact. Strong lensing provides measurements of the *distribution of dark matter* in elliptical galaxies since the light paths are affected by the mass distribution of the lens. The baryonic matter of the lens can be observed and measured directly and the effect on the path that the light travels can also be measured.

Galaxy clusters can cause lensing that magnifies distant galaxies to such an extent that they can be observed directly, which might otherwise not be possible due to the large distances to such sources. This magnification allows *observations of the earliest galaxies* to be made due to magnifications up to a factor of 30, at a time when the Universe was only about 10-15% of its current age [54].

Galaxy clusters cover a larger area compared to single galaxies and are much more massive, thus providing a larger chance to observe lensing of background sources. However, they have additional complications in the form of matter distributions within them that need to be taken into account. Galaxy clusters that create lenses magnify the background sources in angular size as well as integrated brightness. The earliest galaxies observed reveal that they are very small in nature, growing in size through accretion of surrounding hydrogen gas and through mergers with other galaxies. Strong lenses thus assist in observing parts of the evolutionary path of galaxies, improving our understanding of galaxy formation and evolution.

1.4 Large Optical Surveys

Often an astronomical survey is comprised of observations of a specific region of the sky without there being any specific observational targets. These are referred to as sky surveys. However, surveys can also often be focused solely on a specific type of target source, with the goal being to gain additional information about these targets. These second types of surveys are often subsets of sky survey types, but with additional observations made on the target sources. An example of such is the Extremely Luminous Quasar Survey [55], a subset of the Sloan Digital Sky Survey (SDSS) [2]. In general, surveys are used to produce astronomical catalogues.

This thesis utilises optical data and as such, optical surveys will be discussed here only. In the following few paragraphs, a few of the optical surveys from the past to future

surveys are briefly looked at to gain a better understanding of the increases in data volumes and sources observed.

DPOSS : The Digital Palomar Observatory Sky Survey

The Digitized Palomar Observatory Sky Survey [56], is the result of digitising one of the last major photographic sky surveys, POSS-II [57]. It was one of the largest sky surveys done towards the end of the 20th century, containing roughly 3TB of data in three bands excluding catalogues created for the sources therein.

SDSS : Sloan Digital Sky Survey

In recent years, numerous discoveries have been made by studying millions of galaxies that span over a large area of the sky. These discoveries were based on the SDSS, which images over 200 million galaxies over a quarter of the sky using five different wavelength bands. The SDSS has made it possible to study various aspects of the universe, from the structure of the Milky Way, to large scale cosmological structure and the study of dark matter and dark energy.

First started in 1998, the SDSS has utilised a 120 megapixels CCD camera to make precise photometric measurements for over 900 million objects and has been a revolution in astronomy. Thanks to its multi-fibre spectroscopy, the SDSS has reached data volumes of around 40TB. With further expansions, the Data Release 12 done in July 2014 reached 116TB in total.

DESI : Dark Energy Spectroscopic Instrument Legacy Surveys

The DESI Legacy Imaging Surveys currently includes over 1 billion galaxies within a data set well over 1000TB in size [58]. Twenty five times larger than SDSS, the DESI Legacy Imaging Surveys took six years to complete using three different telescopes. The data set has only been released recently and has not been completely studied yet. The map produced by the survey covers half of the sky to the deepest magnitude depths observed to date.

Vera C. Rubin Observatory

The Vera C. Rubin Observatory will soon be operational and will produce roughly 30TB each night over the planned 10 year observational period [59]. It is expected to

reach data volumes totalling 70PB (1PB = 1000TB) with the corresponding catalogue reaching almost 20PB on its own. It is estimated that the LSST will reach depths 100 times fainter than SDSS in 6 different bands and will create a catalogue containing an estimated 20 billion galaxies and 20 billion stars.

These include some of the largest optical surveys that have been done, as well as ongoing and future surveys. The volume of data available is increasing at a rapid rate and future surveys will be much larger still [60]. Large surveys have become the primary method for astronomers to study the universe. The unprecedented data quality and quantity has produced new scientific opportunities but at the same time the vast volumes of data has created challenges that must be overcome. Astronomy has moved into a data intensive, computationally driven era that requires techniques capable of producing results that match the quality of the surveys.

Not only are methods required to handle and process such large volumes of data, but simple procedures like classifying the morphology of a galaxy becomes extremely time consuming for large amounts of data. Making scientific discoveries often depends on finding the few sources that differ considerably from others, but they become increasingly difficult to find in larger and larger data sets.

One solution to tackling large volumes of data is crowd sourcing, whereby a large number of volunteers are enlisted to work through the data to achieve a goal. This was successfully implemented in the Galaxy Zoo project [7], in which the morphology of nearly 1 million galaxies from SDSS was classified. The project have proven to be successful in classifying galaxies and has since been expanded upon. However, even with a large number of volunteers, 1 million sources is still dwarfed by the estimated 20 billion galaxies expected from the LSST. Even if only a small number of these galaxies can be resolved, it will dwarf existing catalogues like the SDSS.

This necessitates the need for a more automated technique that utilises computational power to speed up the process. This is the goal of the thesis, to expand on and implement such a technique on a data set.

Chapter 2

An Overview Of Machine Learning

2.1 Introduction

As mentioned in the previous chapter, astronomical data contains the typical challenges that come with big data sets: large volumes and high dimensional data along with some more unique issues including gaps in observations of a target, or seeing variations for instance. In recent times however, it has been proven that *big data* can be successfully tackled with machine learning as can be seen in Longo, Cunshi and Sen [61–63].

The simplest definition of machine learning stems from the name itself; a machine that learns. Although the application of machine learning has seen a sharp increase in modern times, the first application of machine learning dates back to the early 1940's, [9]. It is only in recent times however, that the need for such techniques has been required and that the technology exists that is capable of applying such techniques.

One of the key aspects of a machine learning algorithm is that it can automatically improve its performance as it learns. Machine learning involves computers applying algorithms that utilise statistical methods to make predictions or to classify data. These algorithms differ in usage and ability, ranging from basic mappings that relate different data sets to creating complex functions that represent patterns inherent within data sets. In modern times, machine learning algorithms are found in almost every field, from medical diagnosis and biology to facial recognition software and many more.

The chapter starts off with a basic overview of how machine learning algorithms work in general. The typical procedure of applying a machine learning algorithm is illustrated in section 2.1.1. In the sections that follow, a broad outline of the different types of

machine learning algorithms is made in section 2.2. The algorithm predominantly used in this thesis is covered in section 2.3. Measuring the performance of machine learning algorithms plays a crucial role in evaluating the results. Thus section 2.4 is dedicated to typical performance measurement and evaluation techniques. The chapter finishes off with in section 2.5 with an overview of machine learning applications within astronomy, based mainly on similar works done to detect anomalies.

2.1.1 Basic Procedure Of Machine Learning Algorithms

The basic procedure that machine learning algorithms follow is fairly straightforward. The initial step, perhaps in any scientific study, is data collection and preparation. The data quality and even quantity affects the performance of machine learning algorithms. Machine learning is ideal for big data volumes since it is based on data; the more data there is and the better the quality of the data, the better the algorithms tend to perform. The data preparation step is often crucial as unwanted data entries can be removed that would otherwise affect the outcome of the algorithm. Some algorithms do not require a validation set, or even a training set.

Features are *extracted* from the data points themselves and are unique to each individual data point. They are lower dimensional representations of the data points, often representing single aspects of the data only that are relevant to the task at hand. For example, galaxy mergers are easily identified visibly due to distortions within the natural shape of the galaxies involved. The features that could be used to identify mergers could thus be derived from the shape of the galaxy. Feature selection and extraction is a vital step in order to achieve the required task. Not only does it reduce the amount of information used, decreasing computational times, but it also discards or down weighs less important information. This is also important since some machine learning algorithms can not deal with high dimensional data and require dimensionality reduction techniques.

2.2 Machine Learning Categories

Machine learning models or algorithms can be separated into various different categories. Each category functions in a specific way and has its own requirements. This determines which algorithms are suitable for which data sets and for which problems. For cases where predictions or classifications are to be made based on current, labelled data, then supervised learning would be the best. If no solutions are known for data points within the data set then unsupervised learning algorithms is required. The main different

categories are briefly covered here, with the focus being on their general workings and applications rather than any specific algorithms and how they function. However, some examples are briefly mentioned for each category.

Once the category or type of algorithm required is known, they can be investigated or tested to determine which algorithm, or even an ensemble of algorithms, would work best for the situation at hand. Finding an optimal algorithm can prove to be invaluable and is often necessary to do before tackling the goal directly. The *No Free Lunch Theorem* states that there is no single algorithm that will work for all tasks since each has its own idiosyncrasies, [64]. For this reason it is important to select the algorithm best suited to the data and the task at hand.

2.2.1 Supervised Learning

The main aspect that sets supervised machine learning apart from the other categories is that it requires a *training set* that consists of labelled data points. What this means is that there exists a subset of the data for which the true value or outcome is known. Supervised methods are given this training set with the labels from which they attempt to learn the pattern therein. This pattern is then applied by the machine learning algorithm upon a *testing set* for which the labels are not known. Based on the pattern learnt, the algorithm labels the unknown data. This is the fundamental process of supervised machine learning. Supervised machine learning differs from the traditional model fitting methods since it creates the relations between the input and output itself instead of relying on a predefined relation.

Mathematically this consists of the machine learning algorithm learning a function that maps the data to the known labels:

$$f : \mathbf{x} \rightarrow y \quad (2.1)$$

where f is the function created by the algorithm, \mathbf{x} is the input data, usually in the form of a vector consisting of multiple values that represent the features of the data, and y is the single label or output value. For supervised learning, x and y is known for a subset of the data and this subset is used to determine the function f . The function is then applied on unlabelled data to determine the corresponding y values. For classification tasks, y is a categorical variable and can take on a limited amount of values only. When y is real-valued, it is a regression. Supervised learning can thus be used to predict class,

be it binary or multiple classes. It can also be used to predict probability distributions and continuous quantities as in regression.

Most supervised learning algorithms can be adapted to either type. A few examples of supervised machine learning algorithms, specifically the *Support Vector Machines*, *Decision Trees* and *K-Nearest Neighbour* algorithms, can be found in Schölkopf, Rokach and Altman respectively [65–67].

2.2.2 Reinforcement Learning

Reinforcement learning is similar to supervised learning but it does not use a labelled training set. Instead, the model typically makes a small amount of predictions or classifications which are then *rated*. The algorithm adjusts based on how well it performed on the previous predictions or classifications by making a series of decisions, each one rewarded or penalised depending on how well it performs. The penalties and rewards are set up beforehand, but there is no information provided to the machine learning algorithm to complete the task. It is thus free of any outside influence. Since reinforcement learning constantly improves its rewards through trial and error.

The main challenge with reinforcement learning is in setting up the environment as it has to be very complex and detailed in order to produce quality results for the specific task at hand. The environment includes everything that the agent, or learner, can interact with in order to make its decisions. There is also no way of controlling the actions of the algorithm other than changing the rewards and penalties applicable.

Some examples of reinforcement learning algorithms include some neural networks often applied in video games as well as in navigation systems and robotics.

2.2.3 Semi-supervised Learning

Semi-supervised machine learning algorithms are similar in nature to supervised algorithms except that the training set contains labelled and unlabelled data. In a sense it falls in between supervised and unsupervised machine learning since it is trained on both known and unknown data. The training set typically contains a small amount of labelled data and a large amount of unlabelled data.

This can have multiple advantages. First, labelled data can often be hard to come by, or be expensive as it can require an expert to manually label the data. Using a combination of labelled and unlabelled data reduces this cost.

Additionally, using a combined training set provides the benefits of both supervised and unsupervised learning. Applying supervised learning to such a training set would require the unlabelled points to be removed. The opposite is true for an unsupervised algorithm that would require the labels to be removed. The semi-supervised algorithm is capable of doing both at once, learning new labels in the training set to a higher degree of accuracy due to the known labels. More labels would then be correctly identified for the testing set, increasing the performance of the algorithm compared to supervised learning based only on the same known labels.

However, this is also a disadvantage since semi-supervised learning can not always replace supervised learning. This is because the unlabelled data within the training set for semi-supervised learning must be directly correlated to the labelled data. If they are not, then the algorithm would label them incorrectly, reducing the performance on the testing set. The labelled data must be a true representation of the entire training set, including the unlabelled data.

2.2.4 Unsupervised Learning

Unsupervised learning utilises data that does not have any labels. The algorithm does not know, and does not have any way of knowing, what the true label is for a given data point. It creates its own solution based on the patterns it identifies in the data set. Unsupervised learning algorithms typically finds features that are common throughout the data set and groups the data points according to this.

These algorithms are data driven; the results are solely dependent on patterns identified within the data itself. Unsupervised learning is typically used to detect trends within data as well as unusual data points commonly referred to as outliers or anomalies. One advantage of this is that they can detect trends not known or expected to exist within the data. This makes them ideal for scientific research since they can identify new discoveries and/or provide new knowledge about the data set.

One of the disadvantages of unsupervised learning is the presence of free parameters that can not be optimised. Different parameter values may result in significantly different results since they are not optimised. While the performance might not be affected severely, the results can be seen as being different interpretations of the same data set.

The specific interpretation of the output of unsupervised learning must thus be done with care as these free parameters have different meanings when applied. In addition to this, some of the internal parameters could also be affected by the free parameters. This variability of results due to the change in parameters often produces a different outcome

even though the same algorithm is used. In some instances this can cause unwanted complications. However, as unsupervised learning is applied to unknown data sets, the scientific goal might not be precise and the variability might even assist in the data exploration process.

2.2.4.1 Outlier Detection Algorithms

Unsupervised learning is usually applied on data sets that are unexplored. Often the goal of this is to find unusual data points that are also referred to as outliers or anomalies. Outlier detection algorithms are designed to find these anomalies. They typically rank data points based on how *different* they are to the norm of the data set. Defining what is different and defining the norm of the data is often a debatable concept.

Finding the norm of the data, or how similar the data points are to each other, is the easier of the two. The data set can be modelled in various ways which would determine the common points and density fields can also be used to find similar points. The most common ways to determine which data points are outliers are to find the points which are least similar to the other points, or to find the points that lie in low density sections of the data set, or to find points that do not follow the model of the data well.

Outliers might be useful and interesting sources captured in the data set, or they may be unwanted defects that need to be identified during a process as they can affect the model. They might be a data point altogether unique, or they might be similar to others, but an extreme sample of the data type.

Common outlier detection algorithms include:

- **Isolation Forest:** iForest is the algorithm used throughout the thesis and is covered in more detail in section [2.3](#)
- **Local Outlier Factor:** LOF is an outlier detection algorithm. It works around the basis of finding data points that are located far away from other points in the feature space of the data set. Each data point is scored based on how isolated it is from its nearest neighbours. LOF works well for relatively low dimensional data, but struggles for high dimensional data, [\[68\]](#).
- **One Class Support Vector Machines:** The standard Support Vector Machine (SVM) is typically used for binary classifications, but can be adjusted to classify points as being either normal, or outliers, [\[65\]](#).

- **Clustering Analysis Algorithms:** Data points that are similar to each other are grouped together in the data set. This helps to identify similarities between sources as well as to identify differences between clusters themselves. It can also be used to detect outliers that do not belong to any specific cluster. The goal of clustering analysis algorithms is to find clusters that exist within the feature space of the data set. Unlike classification, it is not used to predict labels or classes, but is used to separate the data into natural groups or clusters. A cluster is an area in the feature space of the data set that contains a high density of data points. Points with similar features will naturally group together and points that do not lie within any specific cluster are considered to be outliers. There are numerous different types of clustering algorithms, a few examples can be seen in Achtert, Saquib Sarfraz and Wang [69–71].

2.2.5 Active Learning

Active learning is a different type of machine learning that is applied upon other machine learning techniques rather than on the extracted features themselves. Often, obtaining known labels for supervised learning can be expensive or difficult to obtain, or in the case of unsupervised learning, there are no known labels [72]. Active learning effectively creates known labels based on the input of the user. For supervised learning this increases the number of labels available, while it can change an unsupervised model into a supervised model.

The performance of the relevant model can thus be improved by applying active learning [73]. For anomaly detection, active learning allows the rare phenomena to be scored higher manually once detected and retrains the rest of the scores based on these labels, increasing the scores of similar sources. The caveat is that in order to increase the number of observations of a specific anomaly detected, samples of such an anomaly must be identified first. This is thus easier to apply for supervised learning if the anomalies form part of the known labelled data. However, for unsupervised learning, active learning can greatly increase the detection of anomalous sources, especially if the rest of the data set is scored low throughout.

2.3 Isolation Forest

Isolation Forest (iForest) is an outlier detection unsupervised machine learning algorithm, [74]. It is the main algorithm used throughout the thesis and is covered in detail

in this section. iForest is based on the Decision Tree algorithm, see Figure 2.1. A decision tree takes some input, applies a condition to it at a node and produces an output. As can be see in Figure 2.1, the first node splits the values into two, and the subsequent nodes split the values further until a class or type is determined. This is an *example* of a classification decision tree. The iForest algorithm works in a similar way, but with the goal of isolating points instead of labelling or classifying them. At each node for the iForest algorithm, a random feature is selected to be the node and a random value for that feature is selected by the algorithm. For anomaly detection, the *path* length, how long the path is before the source is isolated, determines its anomaly score. Section 2.3.1 details the working of the iForest algorithm in more detail.

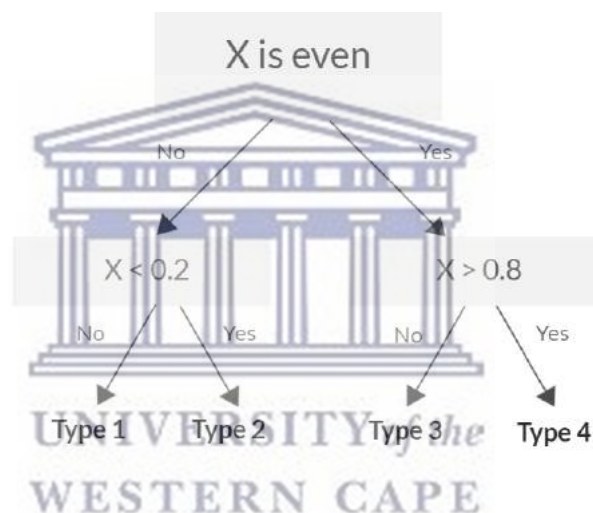


FIGURE 2.1: Example of a simple decision tree. At each node, an input is given and multiple outcomes depending on the output of the node can be achieved. In this example, there are two outcomes per node. The data is input at the top node, whereby a decision is made that splits the data depending on the condition in the node. Each outcome is passed on to another node until a classification or value for regression is obtained.

As the name suggests, the *Isolation* Forest technique uses isolation as the means to detect anomalies rather than the typical distance and density measures used in most outlier detection algorithms. It is based on the principle that outliers are rare, making them easier to identify since they will separate into isolated branches quicker. iForest utilises an ensemble of *Isolation Trees*, that forms the *Forest*, to identify outliers.

While most outlier detection algorithms attempt to model or define the *norm* of the data set first and then proceed to identify the outliers based on the norm, iForest does not. It does not define the norm or common points of the data set, instead it focuses on isolating data points directly. This reduces both the computational time required as

well as the memory requirements to run the algorithm as it does not have to go into the same depth levels as other algorithms. The depth of the relevant isolated point defines the score given to it; the faster the point can be isolated, the higher the score it will receive.

iForest achieves this isolation by splitting the data set based on random feature values. Random partitioning done in this way will create shorter paths for isolated or outlier points. A more detailed explanation of this procedure is done in subsection 2.3.1.

The running time of the iForest algorithm scales as $O(N)$, where N is the number of data points in the data set. This is a linear relation, making iForest capable of handling larger data sets more easily than other algorithms. Due to the random feature selection it does, iForest is also capable of handling high dimensional data; data that has a large number of features.

2.3.1 Algorithm

The procedure of iForest is illustrated in Figure 2.2, with the steps detailed as follows:

1. The first step is to obtain all of the data points with their corresponding features. It is important to note that the features supplied to the machine learning algorithm are used here.
2. A subset of these features are selected randomly by the algorithm. This is done to reduce overcrowding of the feature space.
3. From this subset, a random dimension is chosen. A dimension corresponds to a feature, so the number of different features that is used corresponds to the number of dimensions there are. It is easier to illustrate the concept in a two dimensional way since it is difficult to visualise higher dimensions. In the chosen dimension, a random value is selected, indicated by the cross in Figure 2.2. This produces the top node of the tree.
4. A line is drawn through the data point, which splits the data into two separate sections. These sections form the branches of the tree. It should be noted that this *line* exists in higher dimensions if the data is high dimensional. The figure illustrates a two dimensional case.
5. Another data point is randomly selected on one of the branches to form the next node. A line is drawn again to separate the data points. This line must be perpendicular to the first line since it must make contact with it at some point.

6. This is repeated until all sources have been isolated or until the preset depth limit has been reached.
7. The forest is built up by repeating the tree building process for all of the subsets of the data.

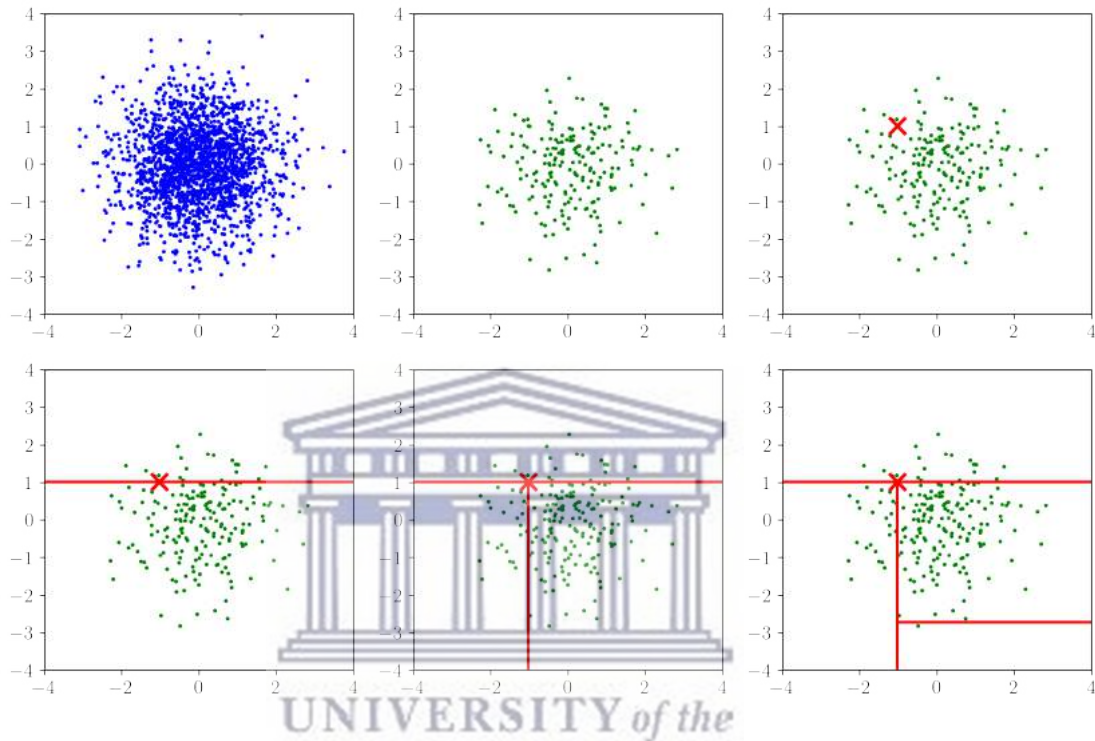


FIGURE 2.2: The plots shown here illustrate the procedural steps that iForest follows. The first plot shows the features of the data that are supplied to the iForest algorithm. The second shows a subset that is selected randomly. The third plot illustrates the starting point, where the algorithm selects a random dimension followed by a random data point in that dimension that forms the first node of the tree. Plot four illustrates the two regions, each forming a branch of the tree. Plots five and six shows some further branches and nodes being made on one side of the tree.

The goal of the iForest algorithm is to isolate each data point. The easier it is to isolate a data point using the above procedure, the more of an outlier it is. These sources can be viewed as being anomalous since they are more isolated from other sources than the rest of the sources.

Decision trees consist of nodes and branches and in the case of iForest. The nodes represent a feature that is selected and the two branches that stem from each node define the regions created that partition the data points. Given a data set that contains N points, X_N , a subset, X' , of the data is created: $X' \subset X_N$. For each node T that exists in the tree, T is either an internal node that contains two branches leading to further nodes, or an external node that has no branches or child nodes. Since each node

consists of a feature, it is that feature of the data points that determines which way the data point will proceed. For example, if the feature at that node has a minimum and maximum value of 0 and 10 respectively within the data set, then a random value, x , is chosen as the splitting point. If a data entry travels along the path and reaches that node, the data point's value of that specific feature will be compared to the random value x . If it is higher than x , it will go further on one branch, if it is lower, it goes to the other branch.

The reason why subsets are selected initially is because they work better when it comes to isolating sources. If the full data set is used simultaneously, there will be points closer to the outliers, which might cause some of them to be missed. Utilising subsections of the data at a time is referred to as sub-sampling.

2.3.2 Anomaly Score

Every machine learning algorithm produces output of some type. For unsupervised learning, specifically outlier detection algorithms, the output is typically some score or ranking given to the data points. The scores given to the sources typically fall within a predefined range so as to be able to compare them to each other. In iForest, outliers are typically isolated more easily, meaning that they tend to be located closer to the root of the tree and have a shorter path length. This forms the basis of the scoring system used by the algorithm. The scoring method of iForest explained here follows the original outline given by Liu et al. [74].

The path length $h(x)$ of a point x is defined to be the number of edges, or branches, that the point travels from the root to its location within the tree. The tree grows by an order of n in maximum possible height, resulting in the average height increasing by $\log(n)$. Normalising the scoring function is thus complicated, However, there are some factors that can be considered to simplify matters. The first is that iForest trees contain only two different types of nodes; one with no branches/children and the other with exactly two branches/children.

This means that the trees are binary and that the same principles for binary trees can be applied to the iForest trees. A node without branches in iForest is the equivalent of an unsuccessful outcome for the standard binary tree search (BST). In turn, the average path length, $h(x)$ to the empty node is the same as the average path length for an unsuccessful outcome in a BST which is given by:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (2.2)$$

where $H(i)$ is the harmonic number, estimated by $\ln(i) + 0.5772156649$, with the latter being Euler's constant, γ . For a given n , $c(n)$ is the average of $h(x)$ and can be used to normalise $h(x)$.

The anomaly score s of a data point x is thus given by:

$$s(x, n) = 2^{-E(h(x))/c(n)} \quad (2.3)$$

where $E(h(x))$ is the average of $h(x)$ from several iForest trees. From this equation we see that s is monotonic with $h(x)$:

- when $E(h(x)) \rightarrow 0$, $s \rightarrow 1$
- when $E(h(x)) \rightarrow c(n)$, $s \rightarrow 0.5$
- when $E(h(x)) \rightarrow n - 1$, $s \rightarrow 0$

From this it is seen that the score s of a data point ranges from 0 to 1. If s is scored close to 1, then it is definitely an outlier within the data set. If s is lower than 0.5, it is typically regarded as being part of the norm of the data set. If all of the points within the data set return scores close to 0.5, then there are no outliers whatsoever.

2.4 Determining The Performance Of Machine Learning Algorithms

Evaluating the performance of any machine learning algorithm is crucial to determine whether the results are reliable. Performance measurements are easier to make for supervised machine learning algorithms as a set of known solutions are available and the results can be compared directly. For unsupervised learning however, evaluating the performance becomes complex and more involved as there are no known solutions or labels to any of the data points. The performance is largely dependant upon the goal of the machine learning application.

Metrics are used to evaluate the performance of machine learning algorithms, but it is vital to use the correct or appropriate metric for the algorithm in question. Often, even using a single, albeit correct, metric is not sufficient to evaluate the performance of the model sufficiently. Metrics are inherently different from loss functions. Loss functions indicates a measure of the performance of the algorithm and can be used to train an algorithm. Metrics are used to measure and monitor the performance of the algorithm and is usually not used for training purposes.

Some methods of determining the performance of algorithms are detailed below. For the most part, these include classification algorithms as their performance is the easiest to determine.

Confusion Matrix

One of the most important concepts for any type of classification algorithm is that of the confusion matrix. It is a comparison of the predicted outcomes from the model and the actual labels or values.

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negatives (TN)	False Positives (FP) Type I error
	Positive +	False Negatives (FN) Type II error	True Positives (TP)

FIGURE 2.3: This shows the basic outline of a confusion matrix for a two dimensional classification algorithm. ¹

The following definitions are used to calculate some of the performance measures:

- True Positives - These are sources that are correctly labelled as being positive by the algorithm. It is the number of actual positives labelled correctly as being a positive.
- False Positives - These sources have been labelled by the algorithm as being positive but they are not. They are incorrectly labelled.

¹<https://www.nbshare.io/notebook/626706996/Learn-And-Code-Confusion-Matrix-With-Python/>

- True Negatives - Normal sources that are correctly labelled as non positives. They are correctly labelled.
- False Negatives - These are positives that are missed. They are labelled as being normal sources but are actually positives.

The evaluations and performance measures used are based on the above definitions:

Recall - Is a measure of how well the positives are identified. Also known as the Sensitivity or True Positive Rate.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2.4)$$

Recall is a measure of how many positives are correctly identified out of all of the actual positives within the data set.

False Positive Rate - Is a measure of the number of incorrect positive predictions made against the total number of negatives.

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} \quad (2.5)$$

The False Positive Rate is a measure of how many false positives are returned; how many values are incorrectly labelled as being positive out of all of the negative values.

Precision - In its simplest definition, the precision is the ratio between true positives and all positives:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2.6)$$

Precision is a measure of how many positives are correctly identified out of all of the sources returned as positives.

Accuracy - Is a measure of how many sources are correctly identified by the algorithm.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Predictions Made}} \quad (2.7)$$

Accuracy measures how many positives are correctly identified as well as how many normal sources are identified correctly.

ROC Curves

Common metrics that are used to evaluate the performance of classification algorithms include receiver operating characteristic curves (ROC-curves), [75, 76]. ROC curves illustrate the performance of classification algorithms as a function of the cut-off threshold. The true positive rate is plotted against the false positive rate for different cut-off threshold values. The area under the curve (AUC) value is used as the performance measure as it measures the performance of the classifier on all possible threshold values. The values lie between 0 and 1, where a higher value translates to a better performance.

2.4.1 Rank Weighted Score

The Rank Weighted Score (RWS) is a numerical formula that can be used to measure the performance of a machine learning algorithm, specifically an outlier detection algorithm where the rank of the output is important [77]. The RWS assigns a value to the location or *rank* of each anomaly. The higher up an anomaly is ranked, the larger impact it has on the RWS score. The RWS is given by:

$$\text{Rank Weighted Score} = \frac{1}{S_0} \sum_{i=1}^N w_i I_i \quad (2.8)$$

where

$$w_i = (N + 1 - i) \quad (2.9)$$

and

$$S_0 = N(N + 1)/2 \quad (2.10)$$

I_i is a Kronecker delta function that is equal to 1 when the source is an anomaly and 0 otherwise. N is the number of sources in total and i is the *rank* of the source. The value of the RWS ranges from zero (where no anomalies are found) to 1 (where the anomalies occupy all the top values). This provides a numerical representation of the performance of the algorithm. The RWS score is not to be confused with the individual scores of the data points themselves. Rather, it is a measure of how accurate the scoring system of the machine learning algorithm is.

2.5 Machine Learning Applications In Astronomy

Machine learning has become more popular in astronomy in recent years, with the large data sets typically found in astronomy being well suited for such applications. The wide range of properties usually available in astronomical surveys, such as the mass, luminosity, light curve, spectra and so on, provides ample features for machine learning algorithms. In most cases, the model is some form of classification, such as classifying spectra into stars or quasars [78, 79]. However, it is also possible to have a regression task, normally occurring where estimations such as redshifts are to be made [80]. Static data in the form of images based on the flux levels of the sources in question are utilised in this thesis with the goal of detecting outliers that correspond to astronomical anomalies. This section covers some literature reviews of similar machine learning applications on static astronomical data that are similar in nature to that used in this thesis.

Automatic Identification Of Outliers In Hubble Space Telescope Galaxy Images (Shamir 2021)

Galaxies are usually classified into morphological types based on their visual appearance. However, there are some that are considered *peculiar* and which can not be classified to belong to a specific morphological class. These galaxies can carry important information regarding galaxy evolution and are thus important to identify.

The data used within this paper consists of several HST fields that form the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) [81]. A total number of 176 808 sources were investigated, with the majority having redshifts below $z = 1$. Each source was converted into a set of numerical image content descriptors that describe the visual contents of the image in question. The similarity between each pair of galaxy images is determined by computing the Earth Mover's Distance (EMD) metric on the numerical image content descriptors. The EMD calculates the cost of moving from one object to another. Outliers are identified to be the sources with the greatest cost since they are furthest away from other sources [82].

The results indicate that the algorithm struggles to detect some anomalies. Only 2 of the 67 known gravitational lenses are detected when applying the algorithm. However, it is proven that a significant amount of *normal* galaxies can be identified and removed, thereby reducing the amount of data by two orders of magnitude. A total of 147 interesting and anomalous sources were detected that would otherwise have been difficult to identify, out of a total of 1 100 determined to be peculiar.

Unsupervised machine learning techniques are proven to be efficient in detecting outliers within large data sets. A large amount of false positives suggest that the algorithm struggles in detecting all of the outliers, but the algorithm can greatly reduce the initial data set, making manual investigations much easier [81].

Pushing Automated Morphological Classifications To Their Limits With The Dark Energy Survey (Vega-Ferrero et al. 2020)

Morphology is a key factor that is related to various physical properties of galaxies, such as the star formation rate and galaxy mass. Identifying the morphology of a galaxy thus enables the identification of certain properties without directly measuring them.

The data consisted of almost 27 million galaxies that form a part of the Dark Energy Survey science Data Release 1, which utilises the Dark Energy Camera and consists of the *griz*-bands. Convolutional neural networks, a type of deep learning algorithm, are used to morphologically classify the galaxy images. A training set consisting of fully classified galaxies that have been simulated to be at a higher redshift is used.

The results are an impressive 97% accuracy in classifying the morphology of the galaxies according to whether the galaxy is an early-, or late-type galaxy and whether the galaxy is edge-on or face-on. Five different models are trained using *k*-folding (where the data is split into *k* consecutive folds, each fold is used once as a validation set, while the remaining folds are used as training sets) to determine the uncertainty. Roughly 87% of the galaxies have secure classifications regarding whether they are early or late type galaxies, and 73% have secure classifications whether they are edge-on or not.

The faint images used are difficult to distinguish visibly. The work done here demonstrates that machine learning can be used to identify features that are visibly hidden due to the faintness of the sources. The method of creating a training set simulated at a higher redshift can be utilised in future surveys [83].

Anomaly Detection In Astronomical Images With Generative Adversarial Networks (Storey-Fisher et al. 2020)

Generative Adversarial Networks (GANs) are a type of deep neural networks that consist of two parts and are suited to identify outliers [85]. The first part is the generator: this models data/images based on the training set, performing better on common objects within the training set and poorer on rare or anomalous objects. The

second part is the discriminator; this distinguishes between the generated images and the real images and identifies the poorly modelled rare images generated in the first part. This is ideal when it comes to detecting outliers.

Data from the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) is used. The wide-field optical survey is imaged with the Subaru Telescope using 5 filters and contains over 430 million sources. A subset of 942 782 sources was selected by restricting the magnitude used to $20.0 < i < 20.5$, consisting of about 70% extended objects and 30% compact objects.

A Wasserstein GAN [86], similar to a GAN but with the discriminator replaced by a critic, is applied here to detect outliers which are then classified. The critic scores how real or fake an image is. The WGAN is used to recreate the images in the data set, which are then compared to the original images. The difference between the images forms the residual, which is used to determine if the source is anomalous or not. This works since WGANs recreate typical images much better and do not create anomalous sources. Where the residuals are high, the original source is deemed to be more anomalous.

A total of 9 648 sources were identified to be anomalous based on having a score greater than 3σ and were followed up on. A full catalogue is still being compiled for publication. The application of the WGAN shows promise for detecting anomalies within large data sets. Recreating images allow training sets to be created that would otherwise be difficult to obtain. It is also proven to be scalable to larger data sets and is easily reproducible [84].

Discovering New Strong Gravitational Lenses In The DESI Legacy Imaging Surveys (Huang et al. 2020)

Strong gravitational lensing systems provide useful information for astrophysics and cosmology; they can be used to study how dark matter is distributed in galaxies and galaxy clusters and are suited to study dark matter beyond the local universe. Strong lensing provides the only known way to study the morphology and internal structures of galaxies at sub-kpc scales at high redshifts that can extend to $z > 2$ and are thus important discoveries.

The DESI Legacy Imaging Survey's DR8 covers almost one third of the sky and is the source of the objects used in this study [58]. A training sample consisting of known

lensing systems as well as non-lenses in the Legacy Surveys is created for the machine learning algorithm. This training set contained about 21 000 non lens sources and 632 previously known gravitational lenses. Almost 10 million galaxies were investigated using the algorithm.

Deep convolutional neural networks (CNNs) and their variations have been shown to be highly effective in identifying instances of strong lenses within astronomical data sets. The results contain a total of 1210 new strong lens candidates that are identified. In addition, the efficiency of the neural network has been improved significantly. The results contain a significant amount of newly identified strong lensing candidates, a substantial amount compared to the ones currently known. This also indicates that the method used is successful in detecting instances of strong gravitational lensing [87, 88].

Practical Galaxy Morphology Tools from Deep Supervised Representation Learning (Walmsley et al. 2021)

Machine learning techniques have been applied to astronomical data in many different forms, from basic classifiers to more complicated deep neural networks. Often data is simplified, or features are extracted, to create a simpler representation of the original source. These representations, particularly those created for images, are important in astronomy.

The representations of these galaxies created by deep learning models can be useful for tasks outside of those for which they were created. These representations can thus be used to outperform existing methods for certain tasks when investigating large galaxy samples. The machine learning network used in the paper was trained on Galaxy Zoo data, which includes DECaLS data.

Throughout this paper, data from the Dark Energy Camera Legacy Survey DR5 was used [58, 89]. An r -band magnitude cut of $14.00 < r < 17.77$ was made to ensure that the fainter galaxies are within the bulk of the population with SDSS spectroscopy [90] and so that the brighter galaxies exclude those with unreliable radii measurements. The result was a catalogue of 305 657 galaxy images.

Combining these representations with the Astronomy framework, [91], resulted in 100% accurate identification of the most interesting 100 anomalies (as judged by Galaxy Zoo 2 volunteers). Additionally, with only a few extra labelled galaxies, these representations outperform models fine-tuned from terrestrial images or trained from

scratch when used to identify ring galaxies [89]. This very new approach represents a promising alternative to the techniques investigated in this thesis.

2.5.1 Advantages And Disadvantages of Machine Learning

Despite being a concept that has been around for nearly a century, machine learning has only recently expanded due to advancements in technology and the significant increase in volumes of data. New ideas and methods are constantly being developed and applied. Machine learning has already proven to be extremely valuable in astronomy, being used to detect rare phenomena such as gravitational lenses, galaxy mergers or even to perform more common tasks such as morphology classifications [92–96].

Large data sets in astronomy has also led to an increase in interest in the field. Machine learning applications to astronomical data has attracted data and computer scientists amongst others via the popular challenges hosted on Kaggle². Alternative solutions to handling large volumes of data include citizen science, where members of the public volunteer their time to complete various tasks [7, 97]. The outcome from these citizen science projects have produced valuable results, yet they also highlight the need for improved techniques. The Galaxy Zoo project for instance, has produced morphological classifications for numerous galaxies with the assistance of thousands of volunteers, but this is not enough on its own for upcoming surveys that will have much larger data sets. The results of the Galaxy Zoo projects, specifically the Galaxy Zoo DECaLS project, have been used to train machine learning algorithms in order to make further classifications [98].

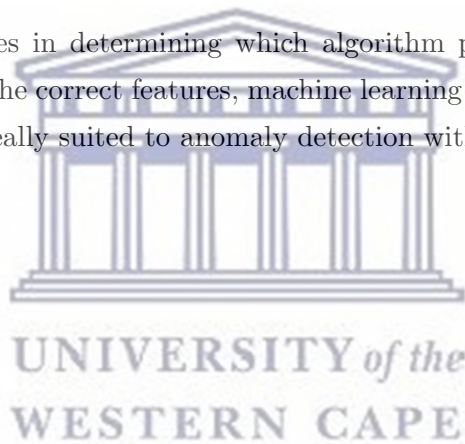
While machine learning automates and speeds up a significant amount of the procedures of handling the data, it can not replace manual labelling or human expertise and can not recreate all aspects of the process. By returning the more interesting results, or being able to identify large amounts of uninteresting data that can be removed, machine learning makes it much easier to utilise large data sets and to detect anomalies within them.

Other challenges also affect the performance of machine learning algorithms. Machine learning is almost entirely based upon the quality of data. If the data is not of good quality, or ideal to learn from, then the results will not be ideal. Studies have shown that data sets are the largest limiting factor for machine learning algorithms [99, 100]. The lack of enough labels for supervised learning is also an issue that must be dealt

²<https://www.kaggle.com/c/PLAsTiCC-2018>

with and can be sorted using mock data [101]. The next challenge that the algorithms face is that of the features extracted from the data. The features play a vital role as they represent the data, or the key aspects of the data, that the machine learning algorithm utilises. However, it is often the case that the wrong or less efficient features are chosen. In some instances of unsupervised and deep learning, the features are generated automatically by the algorithm [102]. This reduces human input error and bias but may not always be the most ideal feature set to use. Several algorithms also include random aspects, be it when selecting features or selecting random values of features or in some other way. This random part will almost always affect the results differently, with each application of the same parameters producing different results. Furthermore, studies have shown that outlier detection techniques, including the iForest algorithm used within the thesis, produce significantly different results when compared to each other [103].

Despite the challenges in determining which algorithm performs well, setting up the data and extracting the correct features, machine learning repeatedly outperforms other approaches and is ideally suited to anomaly detection within astronomical data sets.



Chapter 3

Methodology I: Applying Base Astronomy to DECaLS Data

3.1 Introduction

The DECaLS data set provides an ideal opportunity to test machine learning algorithms designed to find outliers. Not only is the data set massive in volume, the survey reaches depths not easily observed before and has a significant chance of containing previously undetected anomalous sources. Astronomy is a flexible framework for anomaly detection in astronomy [91]. It provides a complete pipeline with multiple options for each stage; from data access and processing to anomaly detection, output display and active learning. The details of the DECaLS data and of Astronomy will be discussed in this chapter.

Data selection is crucial for anomaly detection. Large volumes of data that are not fully explored are ideal for this purpose. When coupled with some of the deepest magnitude depths reached to date, it is difficult to find a better data set suited to the task than the one used in the thesis. Section 3.2 discusses the Dark Energy Spectroscopic Instrument Legacy Imaging Surveys, of which the Dark Energy Camera Legacy Survey data set used throughout the thesis is a part [58]. Technical details of the surveys are covered here, as well as details about the DECam Legacy Survey which contains the data used throughout the thesis. Some issues concerning large volumes of data are also covered briefly in this section. The format of the DECaLS data set is also reviewed briefly in this section.

This chapter also covers the basic procedure of applying Astronomy on a data set. An introduction and overview of Astronomy is covered in section 3.3. General steps are outlined in this section, highlighting the various applications of Astronomy as well

as detailing the overall modular design. The final section, 3.4, examines the processes used for the DECaLS data specifically. Since Astronomy has several functions to use during each processing step, those used specifically for the DECaLS data are covered here.

3.2 Brief Overview Of The Legacy Surveys

The DESI Legacy Imaging Surveys are three public imaging survey projects combined. The Legacy Surveys are designed to provide targets for the DESI survey. The main scientific goal of the DESI project is to study dark energy by measuring the cosmic distance scale using the baryon acoustic peak method. DESI will also investigate the growth of large scale structure using redshift-space distortions in the redshift range $0 < z < 3.5$ [104]. Other scientific goals will also be done, including cosmological constraint measurements and an in-depth survey of galaxies, clusters and quasars. The three Legacy Surveys along with their locations are as follows:

- **Dark Energy Camera Legacy Survey**¹ - Also referred to as the DECam Legacy Survey, utilises the Dark Energy Camera on the 4m Blanco telescope at Cerro Tololo.
- **Beijing-Arizona Sky Survey (BASS)**² - The Beijing-Arizona Sky Survey is hosted at the BOK telescope at Kitt Peak and uses the 90Prime instrument.
- **Mayall z-band Legacy Survey (MzLS)**³ - This is also hosted at Kitt Peak, but uses the MOSAIC-3 camera on the Mayall 4m telescope.

For the purpose of this thesis however, only part of one of the individual surveys within the Legacy Surveys will be used, namely the Dark Energy Camera Legacy Survey.

3.2.1 Sky Coverage Of The Legacy Surveys

The area of the sky covered by the various Legacy Surveys along with the bands used can be seen in Table 3.1. It should be noted that the DECaLS project made use of existing DECam data that is located within the DESI footprint. The largest of this comes from the Dark Energy Survey (DES), which is a 5000 deg² area located towards the South Galactic Cap. This existing data was incorporated directly into DECaLS as it met all requirements of the Legacy Surveys.

¹<https://www.legacysurvey.org/decamls/>

²<https://www.legacysurvey.org/bass/>

³<https://www.legacysurvey.org/mzls/>

TABLE 3.1: The table indicates the bands used by the individual Legacy Surveys. Also included in the table is the area of the sky covered by each individual survey.

Legacy Survey	Band	Area Covered
DECaLS	g,r,z	9500 deg ²
MzLS	z	5000 deg ²
BASS	g,r	5000 deg ²

The 8th public data release (DR8) of the Legacy Surveys contains the data used in this thesis and is the first to contain data over the entire footprint of the survey. It contains data from all three of the Legacy Surveys covered in Table 3.1. The sky coverage ranges from 19 437 deg² for single pass observations, to an area of 13 161 deg² with at least three passes. These regions contain observations using all three bands. Areas exist that contain fewer bands, but these are not used within the thesis. In addition to this, the DR8 also includes Wide-field Infrared Survey Explorer (WISE) flux values [105].

The DECaLS part of the DR8 will be the main focus to study as mentioned before. This is mainly due to the fact that it contains regions of the Southern Galactic Hemisphere, which is studied less than the Northern Galactic Hemisphere. This improves the chances of making new and interesting discoveries. Additionally, this can also improve the knowledge of the southern sky if new discoveries are made which would assist future projects that will be done.

3.2.2 Depths Of The Legacy Surveys

Figure 3.1 is a colour map that illustrates the z-band ($5\text{-}\sigma$) depths of the Legacy Surveys. It can be seen that the majority of the sources lie towards the fainter, higher magnitude range of the map as indicated by the darker blue regions, this is also illustrated in Figure 3.2.

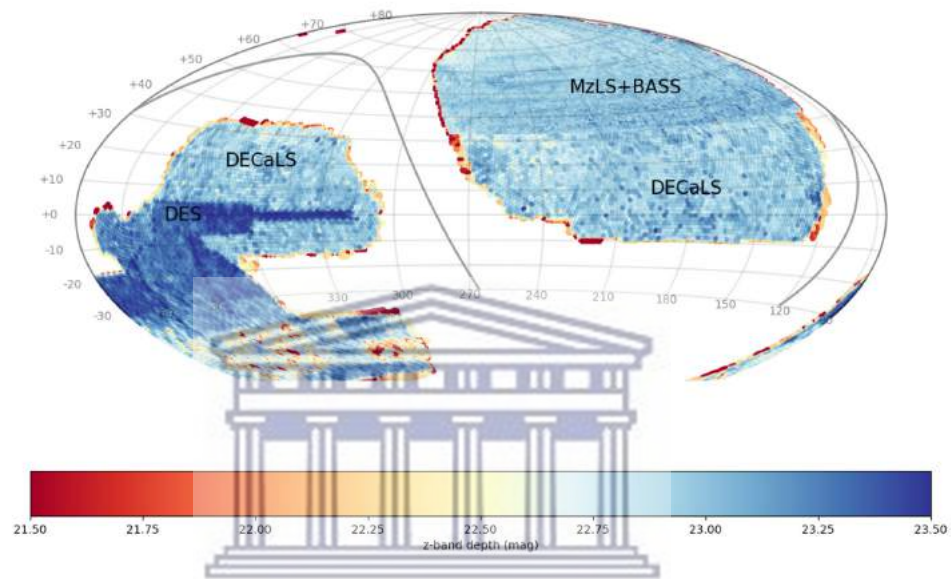


FIGURE 3.1: A colour map indicating the z-band depth in magnitudes for the Legacy Surveys. A redder colour indicates less depth, while a darker blue colour corresponds to a higher magnitude and thus fainter source. The map also illustrates the regions covered by the Legacy Surveys as well as the more in-depth Dark Energy Survey [106].

The outline of the Legacy Surveys can also be seen on the figure. Plots are also available for the g and r bands but are not included here. The depth varies due to a multitude of reasons, from filter quality to extinction levels, observing conditions and more.

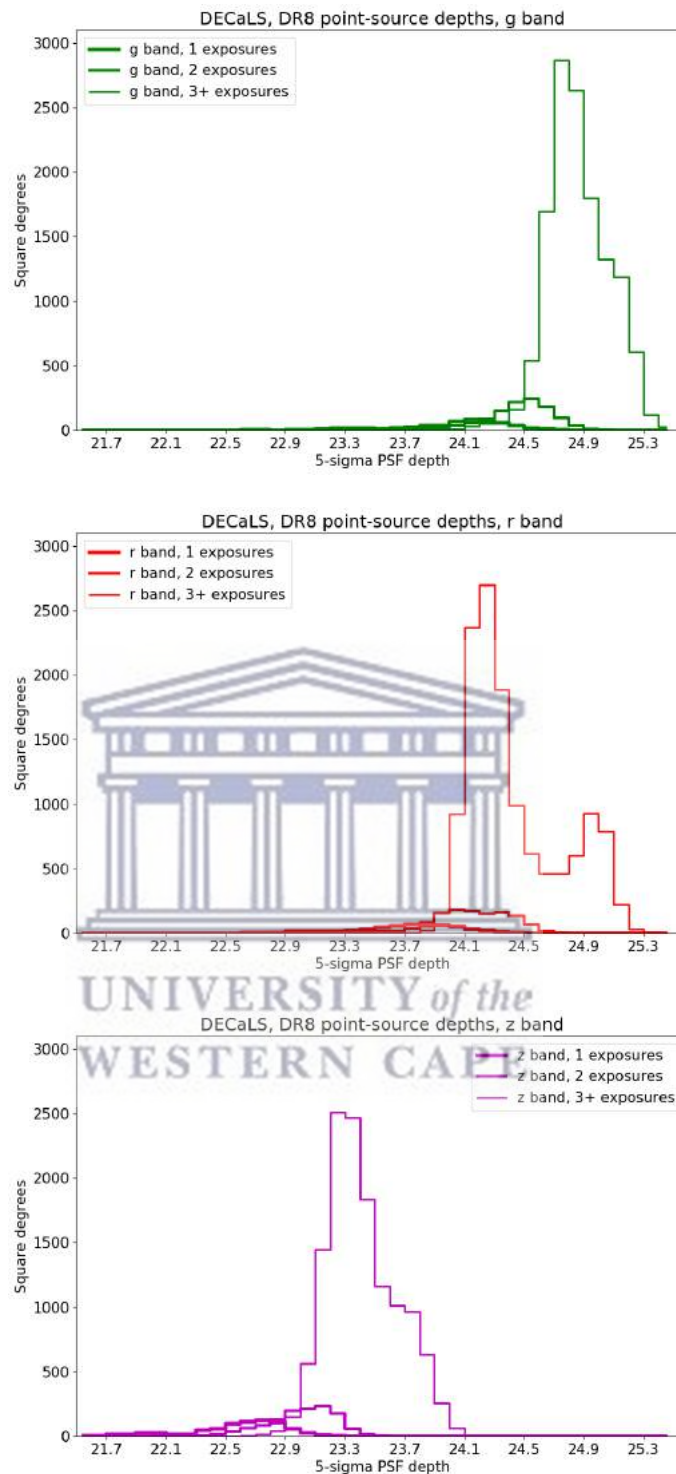


FIGURE 3.2: The plots illustrate the sky area covered against the number of exposures made over the region. Most of the sky covered by the survey has been observed using three or more passes in all of the bands [107].

Figure 3.2 contains histograms for each band used in the DECaLS project, comparing the area of the sky covered to the 5σ point source depths. In all three plots it can be seen that the vast majority of the sky mapped is done using multiple exposures.

Together, the Legacy Surveys provide remarkable depths and greatly increases the known footprint at fainter magnitudes compared to the SDSS. This makes it a potential treasure trove of anomalous sources. Imaging to the 5σ z-band depth of the Legacy Surveys is expected to increase the number of galaxies detected by a factor of > 15 for redshift $0.5 < z < 1.0$ galaxies and by a factor of > 200 for redshift $z > 1.0$ galaxies when compared to SDSS [58].

3.2.3 DECaLS Data

The total number of sources within the DR8 release of the DECaLS data is roughly 1.6 billion, of which about 870 million are resolved, non point sources. This is a substantial amount of data; both in the number of sources present as well as in the overall data size. Even with current technology, such large quantities of data presents several hurdles that must be overcome despite the data being readily available. The overall size of the data set is well over 10TB, severely limiting our ability to download and store the data and as such, smaller subsets had to be used.

Selecting this subset is dependent upon numerous factors. A random subset would be an ideal representation of the overall data set, but placing certain restrictions also make it easier to detect some anomalous sources. One such restriction made is to remove point sources since they are unresolved. No distinguishing physical features are visible for these sources, making it nearly impossible to identify them as being anomalous especially as the focus is on anomalous morphologies. For this reason all point sources are removed from the data first and foremost.

Several choices are made throughout the thesis to create these criteria for selecting the data. The majority of these criteria originate from complications encountered during the application of the various stages of Astronomy on the data. These complications are covered further in this chapter. Additional limitations on the amount of data that can be used are also encountered due to available computational storage and processing power.

3.2.4 DECaLS Data Format

The standard file type for astronomical data is the Flexible Image Transport System (FITS) type, designed specifically for astronomical data. Multiple channels are usually contained within fits files with each representing a different wavelength band. This

allows each band to be accessed and manipulated individually or all the bands to be processed together for the complete image. DECaLS data is obtained in fits files with three channels corresponding to the g-, r- and z-bands, but there are two distinct ways in which the data is formatted. The data is also accompanied by a catalogue that contains the details of the sources, for example: the flux levels of the different bands, the RA and Dec of the sources and so on.

3.2.4.1 Data Bricks

The first format that the DECaLS data can be found in is referred to as a brick. Bricks are fits files that encompass small, square regions of the sky and are coupled with a catalogue containing the details of the sources located within the brick. They can be thought of as pieces of a puzzle with all the bricks fitting together to encompass the entire survey. The number of sources within each brick varies, but each brick typically contains several thousand sources.

The brick names have a specific format to them that details the location of the sky that they cover. These names are formatted to indicate the central location of the brick. This is best explained using an example. For the brick 0267m062, the first section 0267 denotes the right ascension (RA) of the central coordinates multiplied by a factor of 10. Thus 0267 corresponds to an RA of 26.7° . The letter in the middle indicates plus or minus and is indicated by a *p* or *m*. Finally, the numbers at the end show the declination (Dec) of the brick center multiplied by a factor of 10. So *m062* corresponds to a declination of -6.2° .

The brick fits file consists of the three channels used, each one covering the entire region of the brick. Sources within the bricks are accessed by using the coordinates located within the corresponding catalogue. Individual sources are thus looked at by using their central coordinates and by specifying a surrounding region to display as well. Sources differ in angular size and would thus require different surrounding regions in order to view them in their entirety. This can cause some problems as selecting individual regions proves to be quite tricky.

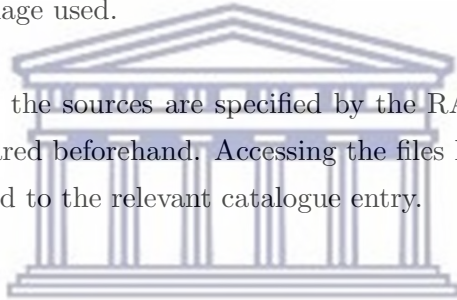
Bricks provide the benefit of downloading thousands of sources in one file with the corresponding catalogue readily available as well. The downside is that the individual sources have to be “extracted” to view them separately.

3.2.4.2 Data Cutouts

The other format that the DECaLS data can be found in also consists of fits files with three channels, except that they are much smaller regions of the sky. These are typically “cutouts” of the individual sources themselves. They are downloaded directly from the SkyViewer Server⁴. The files are labelled `cutout_xx.xxxx.yy.yyyy.fits` where `xx.xxxx` and `yy.yyyy` denote the relevant RA and Dec of the source respectively. A minus sign is included in front of the Dec if needed.

The benefits of using cutouts are that they are centred upon the source and the surrounding region size can be specified before downloading the data. This reduces the amount of data downloaded as the “empty” regions between sources are not downloaded and a minimal size of each source can be downloaded. Noise levels are reduced in this manner as well since there is a smaller region surrounding each source; less noise is located within the image used.

The downside is that the sources are specified by the RA and Dec and thus require a catalogue to be prepared beforehand. Accessing the files locally can also be a challenge as they must be linked to the relevant catalogue entry.



3.3 General Introduction to Astronomy

To efficiently explore the DECaLS data looking for anomalies, we use Astronomy, [108], a generalised framework for anomaly detection in astronomical data. It contains the whole process of applying anomaly detection machine learning algorithms on astronomical data with only minor data specific changes required. One of the main features of Astronomy is its versatility. It is highly modular in design and easily enables different functions and algorithms to be used in different stages by simply exchanging them. Astronomy is also applicable to various astronomical data types such as multi-channel images, time series data, light curves, spectra or even just general sets of features and contains a secondary machine learning algorithm in the form of active learning. The output is displayed in a user-friendly method.

⁴<https://www.legacysurvey.org/viewer>

3.3.1 General Steps Followed When Applying Astronomy

The individual stages of Astronomy are described here in the order in which they typically occur. The outlines of the stages are highlighted, concentrating on the purpose of the stage instead of its actual workings. While each stage has different functions that can be applied within it to yield similar results, certain functions are more effective for some types of data. For instance, some feature extraction methods only work on image data and not on time series data.

Data Pre-Processing

This is the first stage of Astronomy other than reading in and accessing the data. Depending on the data type, the pre-processing stage of Astronomy ranges from merely scaling the data, to complex combinations of functions applied to optimise the data for the next stage of the process. These functions can easily be implemented in any order simply by setting up the list of functions to use in Astronomy. One of the key functions of this stage is to reduce background noise within the data, usually by means of removing the noise directly by applying a sigma clipping function, see section 3.4.2.4. Some of the function include sigma-clipping, scaling using various methods, band weighting and grey-scaling for example.

Feature Extraction

Once the data has been processed, it is passed on to the feature extraction process. Feature extraction can be considered to be a form of dimensionality reduction, with the aim being to extract the most relevant information from the data. This extracted information represents the key characteristics of each data entry in a lower dimensionality space, reducing the amount of data passed on to the machine learning algorithm itself. Selecting the features to be used is critical as it is these features that will determine which anomalies are detected. Selecting which features to extract is thus also data dependent and goal orientated.

Feature extraction takes on many different forms, depending mainly on the type of data that is used. Astronomy includes a power spectrum feature extraction process that utilises the Discrete Fourier Transform (DFT), a wavelet feature extraction process and a shape feature extraction process amongst others. The feature extraction method required can easily be selected without the need to make vast changes in Astronomy.

Machine Learning: Anomaly Detection

With the features successfully extracted from the data, the next step is to apply a machine learning algorithm in order to detect the anomalies. Astronomy uses unsupervised machine learning since training sets are not easily available for anomaly detection. In most cases, training sets are not available at all. This is mainly due to the nature of the data used for anomaly detection. Due to the scarcity of anomalies, it is possible that even a subset of the data would not contain any anomaly.

The two machine learning algorithms for detecting outliers currently available in Astronomy are implemented with the *scikit-learn* software package [109]. They are the iForest [110], discussed in detail in section 2.3, and the LOF [68], discussed briefly in section 2.2.4.1. Additional machine learning algorithms can easily be implemented within Astronomy. The algorithm used can be swapped out simply by replacing one with another in the pipeline. However, a large variety of anomalies exist that are found in various different forms and according to the No Free Lunch Theorem [64, 111], it is not possible to create or select an algorithm that will be able to detect all types of anomalies. It is therefore necessary to explore various algorithms to find one that works best for a given data set.

Applying a machine learning algorithm on the extracted features results in a score being given to each source. These scores determine how anomalous a source is in relation to all of the other sources within the data set. The scores produced by the machine learning algorithms is re-scaled to range from 0 to 5 within Astronomy, with 5 being the most anomalous, solely for the active learning process that can be applied.

Active Learning And Output Visualisation

The front end of Astronomy allows the user to apply active learning in a user friendly manner. Active learning allows the user to incorporate personal experience and knowledge into the scoring system of the machine learning algorithms. Even though the sources are scored by the algorithm, it does not mean that the anomalies of interest are ranked the highest. This is often the case when using image data; an artefact or masked source might be included within the data and could be scored as being highly anomalous despite not being of much interest. Similarly, if a user is interested in a specific type of anomaly only then active learning will assist in scoring these sources higher.

Active learning allows the user to manually label the sources, thereby “training” the algorithm on a subset of the data based on the input from the user. A second machine learning algorithm, called the regressor, adjusts all of the scores in the entire data set based on these trained labels. The regressor predicts new values for all of the sources based on how similar they are to the ones that have been labelled. For example, for a source that is labelled as highly anomalous, the regressor will look for similar sources and will score them higher in turn. Once retrained, the sources are rearranged in the new order based on the users preferences. The regressor can be trained multiple times as new sources are labelled. This active learning allows the user to focus more on sources that they are interested in and reduces the time required to inspect large amount of data.

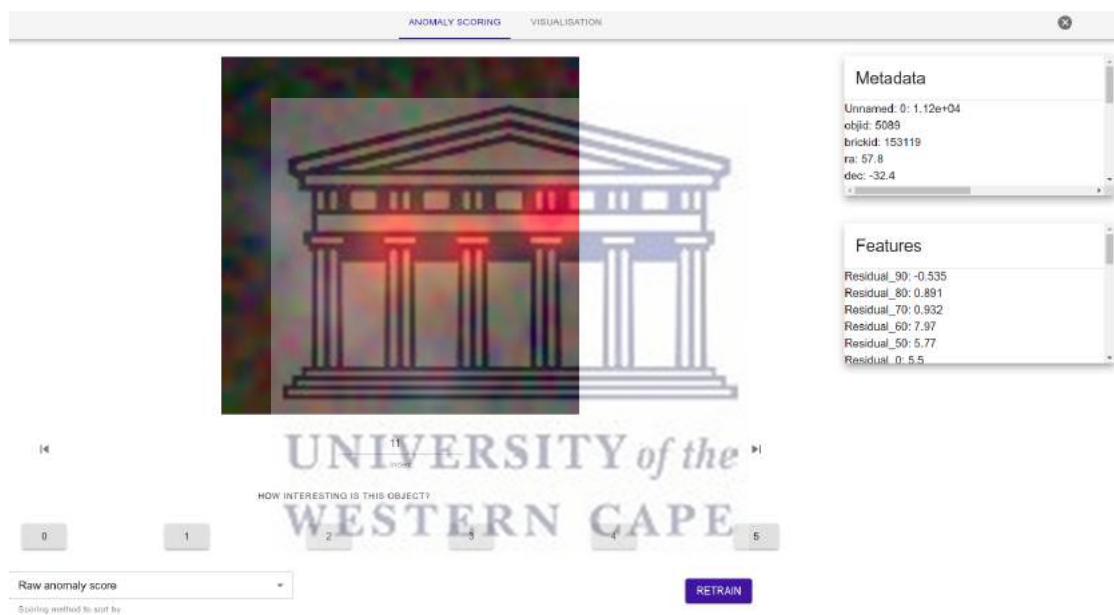


FIGURE 3.3: This is a screenshot of the front end of Astronomy. The images are displayed in order of the anomaly score given to them. At the top right we see the information from the catalogue displayed in the section labelled *Metadata*. This includes the flux values, coordinates, and name along with other details if available. Below this is the *Features* section, which displays the features of the source shown. Below the image is a series of numbers ranging from 0 to 5. These form part of the active learning and are used to manually label the sources based on how anomalous they are. To the bottom left is the option to change the order in which the sources are displayed and to the right of this is the button that retrains the scores based on the manual labels made.

Figure 3.3 is a screenshot of the front end of Astronomy when applied on image data. There are several options to order the output display data and a useful t-Distributed Stochastic Neighbour Embedding (t-SNE) plot can also be included for the data points [112]. The t-SNE plot is used to display high dimensional feature space in a lower, usually two or three, dimensional plot. Included within the front end are features and

information about the source currently displayed.

3.4 Procedure Used To Apply Astronomy On DECaLS

The typical application of Astronomy is in the form of a pipeline which joins together sections of Astronomy to obtain the desired outcome. Pipelines are thus user specific and vary depending on the data used as well as the functions required. The following sections illustrate the full pipeline used on the DECaLS data. They do not delve into the coding details of the pipeline, but rather detail the functions used within the pipeline.

The order follows the steps as they are done. A small section, 3.4.1, on reading in and accessing the data is included at the start and includes some complications that arise from using large data sets. This section also contains the parameters that must be set for the rest of the pipeline. The functions used for each step are explained in depth, except for the iForest machine learning algorithm as this is covered in section 2.3.

3.4.1 Accessing Data And Setting Up The Parameters To Use

This is a small part before the actual pipeline is created. It is the part where key parameters are set up. The data directory is specified here, along with the image and output directories. If required, the catalogue containing the information of the files is read in here as well. All of the parameters are set here, and both the image transform function and the display transform functions are defined here. Additional parameters specific to certain functions are also set here. This section essentially forms the foundation of the pipeline and any small tweak or change is typically done here.

3.4.2 Image Processing Techniques Applied To The DECaLS Data

The first section of the actual pipeline involves setting up the *image dataset* which is an adaptation of the catalogue set up to use in the rest of the pipeline. For example, if the DECaLS brick files are used, see section 3.2.4.1, the image data set would include the location of all of the sources as well the window size used which specifies the region around the sources to include.

This section of the pipeline also deals with data handling, specifically processing and manipulating the input files. Various functions that process the images are applied within this section. Both the image pre-processing function and the output display

transform function are created within this section. The focus in this part will be on the image pre-processing function.

This function is aptly named the *image transform function* and consists of various individual python functions applied onto an input image in a specific order. The following functions are applied to the DECaLS fits files in the order presented here.

3.4.2.1 Image Transform: Scale

Images are comprised of pixels, which can be numerically represented as an array with a value assigned to each pixel. For the fits files this translates to three arrays, one for each band, with a numerical value assigned representing the flux level within that particular pixel of the image.

This function normalises the values of the arrays so that they range from 0 to 1. This is useful for deep learning purposes as it reduces the distribution of the values.

The function determines the highest, I_{max} , and lowest, I_{min} , values of the image and scales each pixel, or value in the array, using the equation

$$\text{New pixel value} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (3.1)$$

where I is the pixel value in question.

This function is also applied at the end of the image transform function, after the sigma clipping in this case, to ensure that all of the values are normalised before they are passed on to the machine learning algorithm.

3.4.2.2 Image Transform: Axis Shift

The DECaLS cutout fits files are the main type of data format used throughout the thesis and are read into a three dimensional array with the shape $(3,x,y)$, where x and y are the cutout width and height respectively. The first number, 3, indicates the number of channels or bands there are within the fits file. It is unusual that the DECaLS data is ordered in this way since the conventional way of producing fits files is to have the channels located at the end such as $(x,y,3)$, although there is no set standard. Since some of the other functions within Astronomy rely on the standard form and so this small function shifts the order of the three dimensions to fit in.

3.4.2.3 Image Transform: Greyscale

This function combines the three channels into a single channel. This is needed for the sigma clipping function, see section 3.4.2.4, so that the image is not clipped on a single band only, but on all of the bands. If the bands are not combined, there might be noise left over that is not clipped out by the sigma clipping which would result in errors further along in the pipeline. This is explained in more detail in the next chapter.

OpenCV's function, *BGR2GRAY*, is used to combine the bands together into one channel [113]. The reason for using this function instead of simply adding the bands numerically is that the function uses a known weighting system. A weighting system scales each band by a specific factor, adding more value to some bands and less to others. The weightings used by OpenCV are the standard weightings used to create grey images which is an ideal method to use when combining bands together. More detail pertaining to why this function is used, as well how how it works, is given in the next chapter.

3.4.2.4 Image Transform: Sigma Clipping

Sigma clipping is used to deal with unwanted background noise surrounding the source. Noise up to three sigma are clipped away to reduce errors. This sigma value is calculated for each image based on the flux values within the image itself. Sigma clipping determines the standard deviation from a centre value of the data and removes the parts that are outside of this standard deviation. After the sigma clipping is applied, contours are fit where the source is expected to be; where the flux remains after the sigma clipping has removed the areas of low fluxes. In instances where there is a nearby bright source, the contour shifts to the edge because the fainter source has been clipped out and the only contour left is the bright source, causing the feature extraction to fail. For the DECaLS data, the sigma clipping is set to 3σ , which is 3 standard deviations away from the central value. The clipping is done a maximum of 5 times or until convergence is reached when there is no more data to clip. An example of sigma clipping is shown in Figure 3.4. It should be noted that in some cases where the source might contain a tidal tail, the sigma clipping function might affect the tail itself.

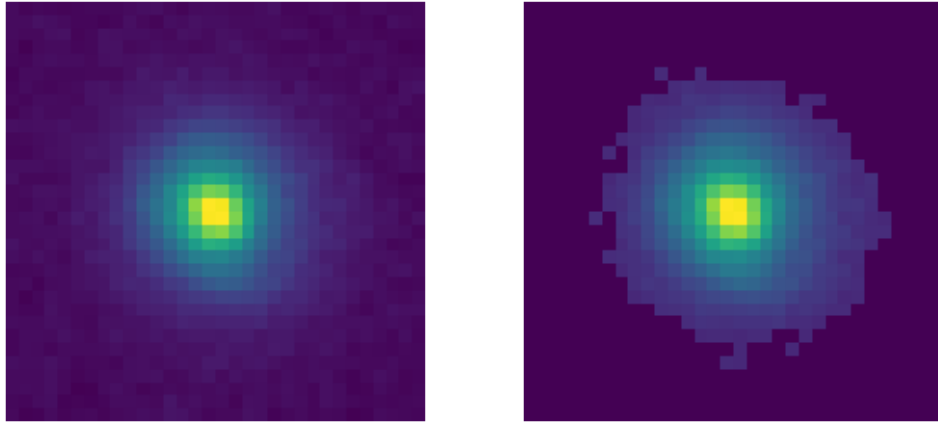


FIGURE 3.4: The images shown here display the before and after effect of sigma clipping applied to a sample of the DECaLS data. On the left we see the image after the previous transformation function have been applied but before sigma clipping has been applied. On the right we see the same image after sigma clipping. The noise surrounding the source has been clipped away.

3.4.3 Ellipse Fitting Feature Extraction Method For Optical Data

One of the key aspects of any machine learning algorithm is the feature set used. Features are used to uniquely define a source or object using a reduced number of dimensions. The ellipse fitting feature extraction method was designed mainly for images of galaxies. Astronomical images are two dimensional representations of three dimensional sources.

Optically, whether they are elliptical or spiral, galaxies tend to be spheroidal in nature to a degree with a few exceptions only. This is due to the formation and rotation of galaxies in general. A two dimensional representation of a spheroidal, or rotational ellipsoid, is an ellipse. Even a spiral galaxy viewed side-on with a large bulge will appear to be mostly elliptical. Some irregular galaxies might be an exception, but their overall outline, the two-dimensional representation of the galaxy as seen in an image, would still be mostly ellipsoidal in shape. Those that do differ significantly tend to do so due to other factors which can lead to such galaxies being classified as being anomalous. For example, they might have a distorted shape that has been caused by past interactions with other galaxies. In the case of the DECaLS optical data, where the data consists of images, the feature extraction method used is the *Ellipse Fitting* feature extraction method.

The ellipse fitting method fits several ellipses to the source. If they are all similar in shape and alignment and fit the contour well, then it is indicative of a “standard” galaxy without any external interference visible. If there are significant deviations between the ellipses then it indicates an unusual shape and the source could possibly be an anomaly. After the image pre-processing has been done and the input images have all been through the transformations applied, they appear similar to the example shown in Figure 3.4.

The points of the source are numerically valued between 0 and 1 due to the scaling transform applied. Six flux percentiles are chosen for the DECaLS data namely; the 90th, 80th, 70th, 60th, 50th and 0th percentiles. For each percentile, the values or pixels that fall within that percentile are located. This forms regions of adjacent points that have similar brightness levels. A contour is drawn around the perimeter of this region and a best fit ellipse is created and fit to this contour. Both the contour and ellipse are created and fit using OpenCV functions and an example can be seen in Figure 3.5.



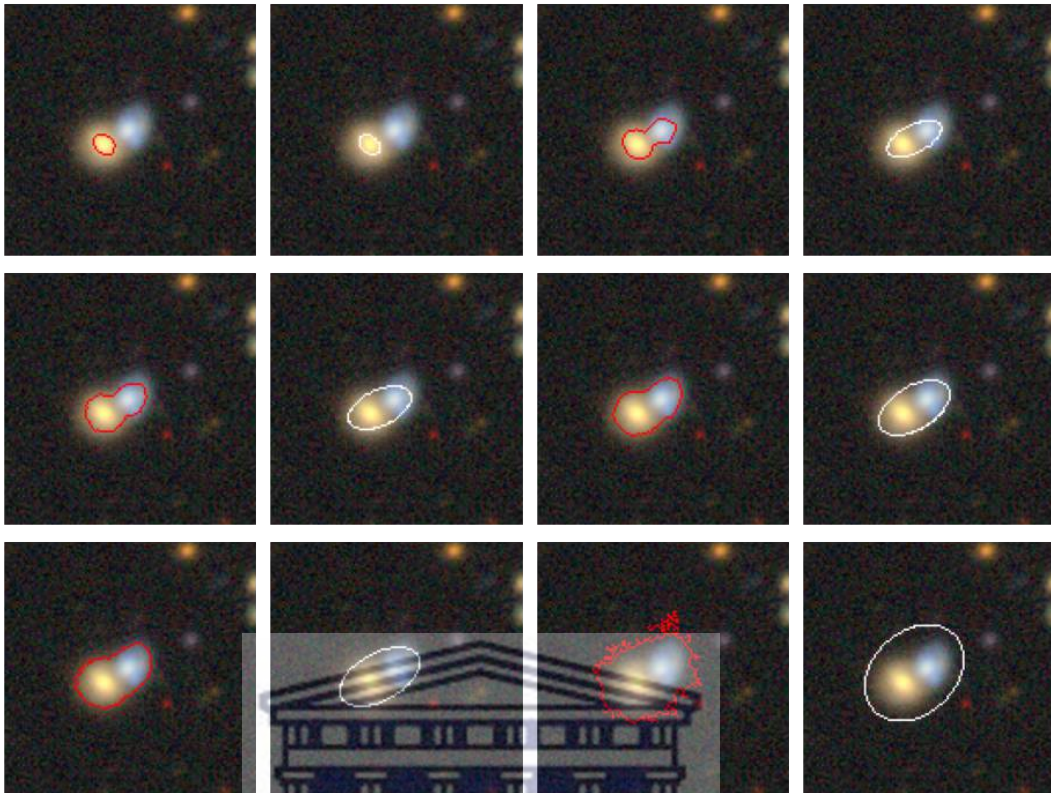


FIGURE 3.5: The six contours and their best fit ellipses are shown here for a random sample. The contours are drawn in red, the ellipses next to it in right. The outermost contour varies the most since it is the boundary between the source and the region cut by the sigma clipping. This indicates how important the sigma clipping step is within the process. Despite being two sources located close to each other, or in line with each other along the line of sight, the ellipses are all quite similar in shape and orientation.

Only the outermost ellipse appears to be rotated with respect to all the others.

An ellipse is fit for each percentile and the parameters of the ellipse forms the features that are extracted from the image and used by the machine learning algorithm. These parameters describing the shape of the ellipses, as well as their relation to each other are as follows:

- **Residual:** This is the difference between the ellipse and the contour. The differences are summed together to create the residual.
- **Offset:** This is the distance between the centre of the ellipse in question and the centre of the 90th percentile ellipse.
- **Aspect:** This is the aspect ratio of the ellipse; the ratio of the major axis to the minor axis of the ellipse. The ratio is then divided by the aspect ratio of the 90th percentile ellipse.

- **Theta:** The rotational angle of the ellipse after the rotational angle of the 90th percentile ellipse has been subtracted. The absolute value is taken since it is with respect to the 90th percentile ellipse.

These four parameters are determined for each of the six percentiles and form the features that are extracted for each image. Since the features are relative to the 90th percentile, only its residual is used as a feature for it [91]. This provides a total of 21 features for each image.

3.4.4 Machine Learning

With the features extracted from the data, the next stage of the pipeline is the machine learning aspect of Astronomy. For the DECaLS data, the machine learning algorithm applied is the iForest algorithm. This can be a time consuming section of the pipeline and plays a role regarding the computational capabilities of Astronomy. In section 4.5.1 the limits of the iForest algorithm are tested.

The sources are all scored by the iForest algorithm and ranked according to this *anomaly rating*. This determines the output order that is displayed by the frontend of Astronomy. Active learning, see section 3.3.1, can be considered to be a part of the machine learning section of Astronomy. The manual labelling and training is done within the frontend, but the labels are passed to a secondary machine learning algorithm that uses these labels to revalue all of the scores.

3.4.5 Astronomy: Frontend

After the sources have been ranked according to their anomaly scores, the output is produced by way of an interactive webpage that is run locally by Astronomy. An example of this can be seen in Figure 3.3.

The sources are displayed here after they have been transformed using the *display transform function*. The transform requires the input fits files from DECaLS to be adjusted using the axis shift function explained in section 3.4.2.2. Once the order is corrected, an *adjusted display function* is applied to better display the files. It is similar to the greyscale function in that it weights each band separately before adding them together, but it uses different weightings. The transform function ends off with the same scaling function as discussed in section 3.4.2.1.

3.5 How And Where The Feature Extraction Process Fails

Early attempts to apply Astronomy to DECaLS data showed very quickly the limitations of the basic algorithm. This is due to the failure of the feature extraction method on realistic and noisy data. There are several different instances where the feature extraction process fails. The failures can be separated into two distinct categories, those due to the source within the images themselves, and those that are caused by the image itself.

The latter often requires a different image to be downloaded and used altogether, whilst the former requires adjustments to be made within the functions applied to the data. In some instances, these can not be dealt with sufficiently or there are no clear methods to deal with the failures that arise.

3.5.1 Feature Extraction Failures Due To Sources

The first type consists of failures caused by the source itself and is largely independent of the image. The various different failure methods are briefly outlined below and the adaptations made to reduce these failures are discussed in detail in the next chapter.

Bright Source Close To The Object

For the DECaLS data, a common reason for the ellipse fitting to fail is due to there being a much brighter source close to the source in question. This is typically a foreground star that lies along the line of sight to the source being looked at. Often the bright source is so bright and large that the algorithm detects it as being a part of the original source and places the ellipses towards the edge where it extends beyond the image itself and fails. In some instances, some of the inner ellipses and contours extend to the bright source and then extends beyond the image itself as well. In this case, the outermost ellipse or ellipses will fail to be fit. The figure below, Figure 3.6, illustrates the source when another bright object is close to the line of sight.

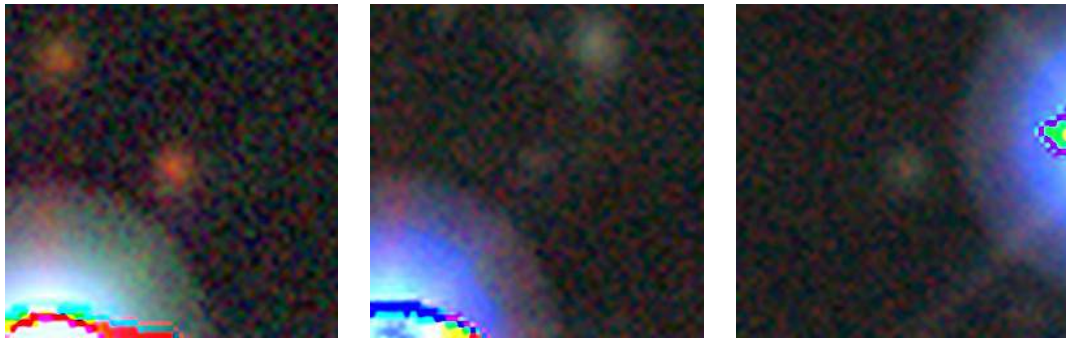


FIGURE 3.6: The three images shown here all fail the ellipse fitting feature extraction method due to there being a much brighter object located along the line of sight to the source in question. These images all contain masked sources, where an attempt was made to mask the brighter sources.

Masked Sources Creating Challenges

Similar to the bright objects close to the source, masked sources can also cause problems. In a sense they are more challenging than having a nearby bright source as the ellipse fitting process does not always fail when a masked source is present. Instead, the ellipses are fit successfully, but incorrectly and have to be dealt with. Figure 3.7 shows some images containing masks. In some instances, the source itself is masked and is typically of a stellar nature, but in others there is a masked source close to the line of sight which can affect the ellipses fit. Masks arise from various phenomena, from solar flares to satellites to stellar sources.

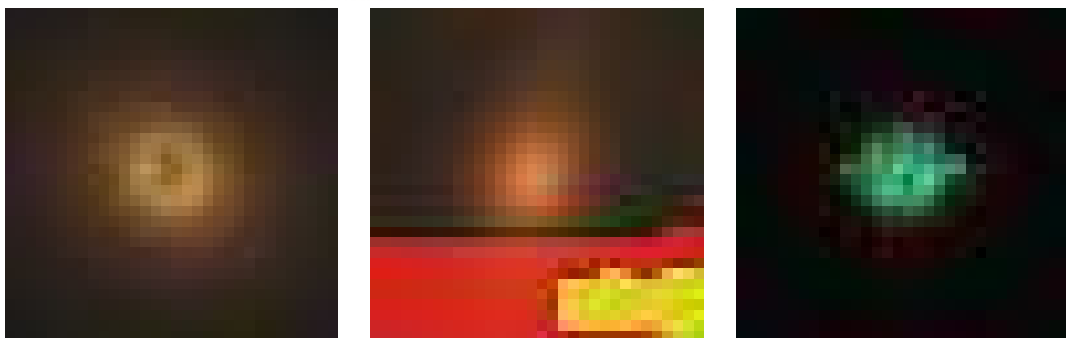


FIGURE 3.7: The images shown here contain masked sources. They are more challenging to deal with than nearby bright sources as they cause incorrect ellipses to be fit rather than failures within the process. Even though they are not desired, they can be scored quite highly by the machine learning algorithms due to their unusual shapes.

Source Too Faint To Fit Ellipses

A significant amount of the sources within the entire DECaLS data set consist of sources that are either too faint to fit ellipses properly, or the signal to noise ratio is too low to distinguish the source from the background noise. For the faint sources, the ellipses can not all be fit properly and there is not much that can be done about it. For the low signal to noise ratio sources, the noise can be clipped using sigma clipping, but care must be taken as the ellipse shape can be affected greatly if this is done incorrectly. Figure 3.8 shows a few faint sources that have low signal to noise ratios.



FIGURE 3.8: These images fail the ellipse fitting process because they are either too small and faint for the ellipses to be fit correctly, or the signal to noise ratio is too low for the source to be identified correctly.

Band Weightings

The three bands are joined together into a single band during the pre-processing stage of Astronomy. Often the bands are weighted differently, meaning that each band has a specific scalar multiple applied to it in order to achieve a certain goal. For instance, certain weightings used will create a grey scale image, while others will highlight certain features more. Some weightings can cause the ellipse fitting process to fail by up-weighting the background noise too much so that ellipses are not fit, or down-weighting certain values causing the source to become “blended” and not having the required distinguishable brightness ranges.

3.5.2 Feature Extraction Failures Due To Image Problems

The second group of failures is caused by the actual image themselves and does not depend on the source within the image. This ranges from incorrect image sizes for the source in question, to incomplete band passes for the sources. Most of the time these

failures can be solved by using different sized images of the source. For some of the instances there is no method to avoid the failure and these sources have to be removed beforehand or the failure is just accepted as part of the process.

Band Passes Not All Available

For some sources, especially those lying towards the boundary regions of the surveys, not all three bands are available. This could be due to various reasons such as faulty equipment or time constraints not allowing all filters to be used. These sources fail the ellipse fitting process due to the lack of a band and is not dependent on the source at all in any way. While there are methods to deal with this failure, these sources are preferably left out to avoid inconsistencies.

Feature Extraction Failures Due To Incorrect Cutout Sizes

Determining the correct cutout size that corresponds to the source is difficult to do. If the cutout size chosen is too small, then the entire source would not fit into the image. This would result in an *open ellipse* that is not a failure in and of itself, but rather an incorrect fit since the entire source is not visible. This results in incorrect features being extracted for the cutout in question, which could affect the overall scoring of the entire data set used. Figure 3.9 shows sources that are too big for the cutout size used.



FIGURE 3.9: Open ellipses, such as those that would result from these images, are not failures since the ellipses are often fit successfully although they would be incorrect.

It is also possible for the cutout size to be too big for the actual source in question. This often results in problems arising from the sigma clipping function during the image pre-processing stage. Figure 3.10 illustrates the issue. Other sources are present throughout the cutout, some of which might be bigger and brighter than the source in question, which can cause the sigma clipping function to clip away parts of the source

that is to be inspected. This would result in incorrect ellipses being fit.

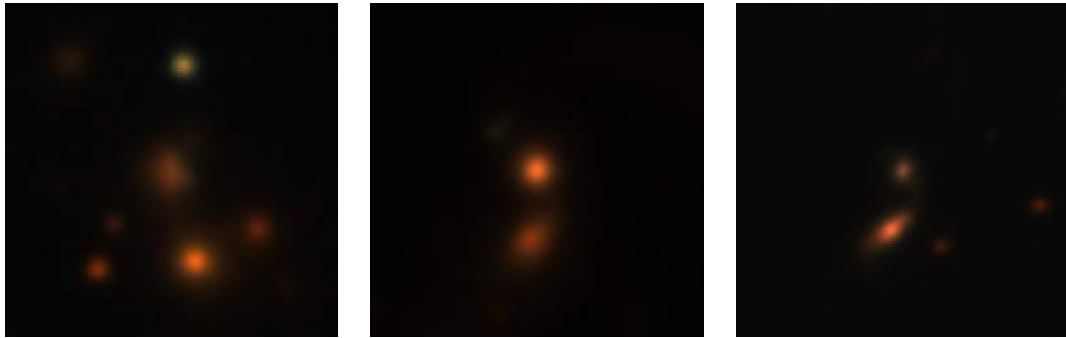


FIGURE 3.10: Other sources within the image will affect the image pre-processing stage, especially the image scaling and the sigma clipping functions. These images all show cases where the image is too large for the source in question.

The biggest issue regarding the image sizes however, is that a significant proportion of the sources are all small and faint in nature. A common image size for these faint sources within the DECaLS data set is 32 by 32 pixels, which is so small that at this resolution, the source itself only occupies a handful of pixels. A large amount of feature extraction failures arise from a lack of data points to fit the innermost ellipse successfully. Five pixels are required to fit the ellipse successfully, but there is not always a sufficient amount of points available for this.

In the following chapter, the changes and adaptations made to Astronomy dealing with these failures and challenges are discussed in detail. The performance increase made from the changes is also discussed. Unless otherwise stated, the changes and adaptations, along with the selection criteria and other data cuts made, are implemented in the data sets used.

Chapter 4

Methodology II: Extending Astronomy

4.1 Introduction

This chapter details the additions and adaptations made to Astronomy to improve its performance and solve the other difficulties encountered during the application of the base version of Astronomy to DECaLS data. The additional changes made to Astronomy relating specifically to the DECaLS data is explained in detail in this chapter. The chapter starts in section 4.2 by detailing the data specific changes made, along with computational complications that arise when using large data sets. A significant change made to Astronomy for the DECaLS data is due to the unusual ordering of the channels found therein. The changes incorporated revolve around increasing the number of sources that have ellipses fit successfully. Sources without ellipses fit are discarded even though they could contain anomalies. Therefore it is crucial to look at as many sources as possible and to understand why the sources are failing.

The feature extraction process of Astronomy used on the DECaLS optical data consists of the ellipse fitting procedure, which is covered in detail in section 3.4.3. One of the main reasons why this feature extraction method fails is due to the source and image size relation; that is, how much of the actual image is covered by the source itself. Sources vary in shape and angular size, both physically and in their appearance to us in images. The angular diameter of the source in question is directly related to how many pixels it covers within an image, regardless of the size of the image itself. It is these pixels that contain the information of the source and it is also these pixels that are used to fit the ellipses on. To avoid unnecessary background noise or interference and to ensure that the entire source is displayed within the image, it is best to use images which conform as much as possible to the angular size of the source.

However, this is not always easy to implement. For instance, if the data is contained within a brick, then a set cutout size surrounding each source is used as the image for all of the sources. Once specified, the image size remains the same for all of the sources located within the specific brick. If the data consists of individual cutout files, then each source can have its own image size specified, but this must be determined beforehand for each source. Section 3.2.4 contains the full explanation of the DECaLS data bricks and the individual cutout files.

The majority of the changes made revolve around the goal of achieving the best fit between the source in question and the corresponding image in order to reduce the failures encountered. Figure 4.1 illustrates the problems that arise when the image size used does not fit well with the angular size of the source in question. All three images displayed are the same size, but the second and third sources require different image sizes for the sources within them. Changes made to reduce the failures resulting from these issues are discussed in section 4.3 and section 4.4. These sections also illustrate the improvements in the feature extraction rate resulting from the changes made.



FIGURE 4.1: The images shown here illustrate the issue with using a single image size (128 by 128 pixels) for sources that have different angular diameters. The first image shows a source that fits in the image well; no other sources are located within the image and the source does not extend beyond the image. The central image shows multiple sources throughout the image, which will affect the sigma clipping process. The last image shows a part of a source that extends beyond the image size used. Ellipses for this source will also extend beyond the image and do not represent the true features of the source.

Section 4.5 focuses on testing the limitations of some of the machine learning algorithms available within Astronomy. Most importantly, the computational limitations of the iForest algorithm are tested. While the iForest algorithm is the sole algorithm used in detecting the anomalies within the DECaLS data set, the LOF algorithm is also tested to provide a comparison in the event that it proves to be a better option

computationally. These tests are very important as they provide limitation estimates regarding the data set size that can be investigated using available computational processing capabilities.

The chapter finishes off with section 4.6, which is a short section summarising the limitations and data selection criteria found within the rest of the chapter. The improvements gained from the adaptations made is also quickly illustrated within this section on a subset of the DECaLS data set.

4.2 Data Format Adjustments And Selection Cuts

In this section, the adjustments made in order to handle the DECaLS data files are explained. As mentioned in section 3.2.4, the DECaLS data files are available in two different formats, bricks and cutout files. The differences between the two different formats is also explained in section 3.2.4. The first choice made regarding the data is deciding which of these two format types to utilise since each has distinct advantages and disadvantages.

Given the large size of the survey, as well as the significant range in angular sizes of the individual sources throughout, the cutout format is better suited for a large scale investigation. This is based on the contents of the bricks, where a significant number of the sources contained therein are typically faint sources with low flux values, as well as the difficulty in selecting the appropriate window sizes for the sources when using bricks. The downside to using cutouts however, is that large amounts of files are needed. This can present storage problems depending on the storage set up used.

A significant change made to Astronomy specifically for the DECaLS data is based on the order that the channels are within in the fits files themselves. Astronomical convention for fits files¹ follows that the two dimensional data be given first followed by the channels, but the opposite is given by the DECaLS data where the channels are given first. This is more of an inconvenience than an issue since Astronomy is designed to follow the norm. A function is thus introduced so as to adjust the order of the fits file where needed. Without this *channel reordering function*, all of the other functions present within Astronomy would fail by default. The data itself is not affected in any way by this function, merely shifted around.

¹<https://www.loc.gov/preservation/digital/formats/fdd/fdd000317.shtml>

The DECaLS data set also contains a significant amount of point sources, mainly stellar in nature that are not investigated in any way in this thesis. This forms the first selection cut made in selecting a subset of the overall data set to investigate. A significant restriction also placed on the DECaLS data is to investigate sources located within the Southern Hemisphere only. This is done mainly on the basis that the Southern Hemisphere is less explored than the Northern Hemisphere and it stands to reason that there is thus a better chance of finding anomalous sources that have not been previously identified. Additional criteria, or cuts, made are discussed in more detail throughout the following sections, and include criteria such as flux levels and bands available amongst others.

4.3 Changes Made For Image Based Failures

This section covers the changes made to Astronomy to reduce the number of feature extraction failures that are caused by the images themselves and not the sources within the images. In section 3.5.2, some of these failures are highlighted. This section covers the changes made in response to these failures as well as the improvements gained from the changes made.

The failures can be broken down into the causes behind them. For images that are too small, there are two related failures. One which is not a true failure, but rather an incorrect fitting caused by the images being too small for the angular diameter of the source. In this case, the outermost ellipse actually extends beyond the image and is fit incorrectly since the entire source is not taken into consideration. This is referred to as an *open ellipse* and is not a true representation of the source in question. The other is due to a lack of pixels within the images to fit the innermost ellipse. This latter issue is a direct result of the actual number of pixels available in the corresponding image and is not related to the area of the sky covered by the image itself.

For images that are too large, other sources are often located within the image, which can cause ellipses to be fit incorrectly as it becomes more difficult to distinguish between sources that are close to each other. More noise will also exist within these images, which will affect the sigma clipping process as well. Another failure is based on the band passes available for the source in question as discussed very briefly in the subsection 3.5.2. A lack of band passes can have detrimental implications on the machine learning algorithm as not all data points are equal. Unfortunately no fix exists for this as it is a case of data being incomplete with respect to the rest of the data set.

4.3.1 Ellipse Fitting Errors Due To Small Image Sizes

This is one of the failures that occur when the image size is too small. These failures occur when the image used is too small for the angular diameter of the source in question. They are challenging in that they do not always cause the ellipse fitting to fail, but often cause incorrect ellipses to be fit. This is because the angular diameter of the source within the image is too large for the image size used, and parts of it exist outside of the image. The ellipse fitting process fits only based on the parts that are within the image, so the outer parts of the sources are not taken into account. This is referred to as an *open ellipse* since the ellipse that is fit actually extends beyond the image, but can not be drawn to be beyond the image. The result is an ellipse that it cut off at the boundary of the image and incorrect ellipse parameters are returned.

Adjustments made to the image itself do not solve this problem since it is a case of the data surrounding the image being missing. The only method to fix this failure is to increase the image size for the corresponding source by including surrounding regions as well. A random subset of 15 000 cutouts consisting of non point sources was used to test various image sizes. Initially all cutouts were downloaded with the image size of 32 by 32 pixels. Out of the 15 000 sources, a total of 909 sources failed the feature extraction process due to various reasons. It should be noted that not all failures are due to incorrect image sizes. These 909 sources were then replaced with images consisting of 64 by 64 pixels around the same sources. The feature extraction process was then applied to the 15 000 sources and it was found that 440 sources failed. These were in turn replaced by images consisting of 128 by 128 pixels, which then resulted in 240 sources failing the feature extraction process. It is evident that the image sizes used play an important role in the feature extraction process, but it is also evident that there exists a range of sizes to be used. Fortunately, using the cutout format allows the sizes to be selected individually for each source. A two-fold approach has thus been incorporated to determine the ideal image size to use for each source.

First, the existing catalogues of the DECaLS DR8 data set have been inspected and the initial cutout size is based on model predictions already made. These are models fit to the sources themselves and contain a semblance of the diameter of the source. This value is used as the base for the cutout size to be downloaded and is increased by 20% to compensate for the endpoints of the ellipses that tend to extend slightly beyond the sources. This means that the cutouts input into Astronomy *initially* are not all necessarily the same size or dimensions, although no negative effects stem from this. This reduces the number of sources that initially fail the feature extraction process due to this image size issue quite significantly. Not only does this improve accuracy, it also

decreases computational times needed since less cutouts need to be replaced by larger ones. Performance is also improved since the sources have ellipses that fit better and more sources are fit successfully. Unfortunately though, it is not able to compensate for all of the sources and some still fail.

The additional step is introduced within the pipeline of Astronomy to assist with this failure. Sources with an open ellipse are flagged within the pipeline by a function that determines whether or not the ellipse fits completely within the image. A new window size is then calculated based on the same ellipse that extends beyond the original cutout size, along with an additional increase to compensate for any changes to the ellipse size that might occur when drawn for the larger cutout. These adjusted window sizes are then used to download larger cutouts which then replace the initial cutouts.

Applying these two steps for the same 15 000 cutouts used reveal that the first step, using the catalogues to determine an initial image size, produces a total of 206 failures. The second step of basing a new image size on the ellipse fit to the initial image, reduces this even further to a total of 188 failures.

4.3.2 Implementing Adaptive Image Scaling In Astronomy

This section covers the other failure type associated with having too small an image. Unlike the previous type, where incorrect ellipses were fit, this failure results in actual failures to fit ellipses. While the previous type was caused by the outermost ellipse, this one is due to the innermost ellipse and arises due to a lack of points to fit an ellipse uniquely. A brief overview of ellipses is given to gain insight into what this means. Ellipses are given by the equation:

$$ax^2 + by^2 + cxy + dx + ey + f = 0 \quad (4.1)$$

Any two ellipses can thus intersect at up to four points, and so four points are insufficient to uniquely determine an ellipse. A minimum of five points are needed as can be seen by Equation 4.1. For small and faint sources, the smallest image size typically used is composed of 32 by 32 pixels. The number of data points available to plot ellipses is limited, more so if the source has a small angular diameter. The innermost ellipse fails since it is only fit on the four central pixels, which is not enough to fit a unique ellipse.

A solution to this problem is to upscale the small images, increasing the number of pixels available and thus providing the additional data points to fit the ellipse without losing any information about the image. This upscaling is done using OpenCV's *resizing function*, allowing the image to be upscaled to a higher number of pixels without losing information [113]. It should be noted that this does not increase the area covered in the image to include extra pixels, but rather increases the number of pixels within the same area, effectively increasing the pixel density of the image. Both the source and the image remain unaffected, other than the fact that the image is now composed of more pixels.

This adaptive scaling was tested upon some bricks of the DECaLS data format. A compilation of ten adjacent bricks were joined together into a data set containing 41414 not point sources. Additionally, a subset comprising of the 500 brightest non point sources of another data brick was also created to determine what effect the scaling will have on bright sources. The results from implementing the scaling is shown in Table 4.1. The results from both data sets show a significant increase in the number of sources that pass the feature extraction process. Quantifying exactly how many of the sources that fail in this way is difficult, as there will often be failures caused by other reasons. This is the reason why the top 500 brightest sources of a data brick were used. From this particular data set it can be seen that 15.4% of the sources failed the feature extraction process before applying the adaptive scaling, but only 1.4% failed after it has been applied. The large amount of failures still present for the compilation data set is thus due to other problems, with the majority due to the significant amount of faint sources present in this data set.

TABLE 4.1: This table represents the results obtained from implementing the adaptive scaling technique. For both data sets tested, it is clear that a minor increase in the number of data points (pixels) available produces the best results.

Percentage Scaled	Brightest Sources	Sources From Compilation
101	423	20642
102	423	20642
103	423	20642
104	493	27113
105	493	27113
106	493	27113
107	486	26363
108	486	26363
109	486	26363
110	475	26415
115	477	26856
120	468	26327
125	473	26260
150	429	24279

Oddly though, it is seen that a small percentage increase in the number of pixels performs better than a larger percentage increase. Further investigations show that this is dependent upon the initial size of the image. For small images, a smaller percentage increase provides the single additional data point needed but a larger percentage increase tends to cause complications with the resizing function. The function ends up expanding the small central part where the innermost ellipse is located, but if the scaling is too much the ellipse ends up back in the central part and still missing the data point needed.

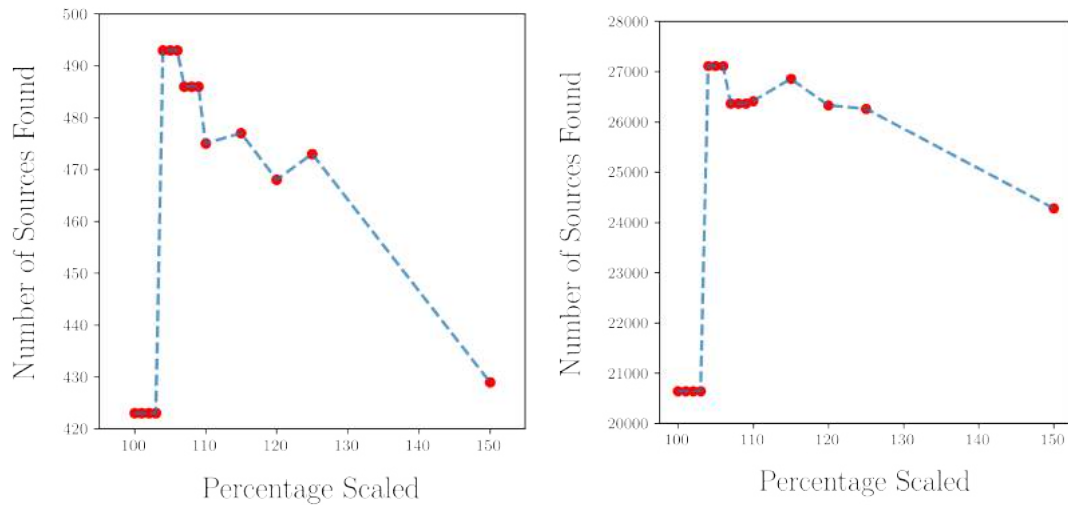


FIGURE 4.2: These two plots illustrate the results shown in Table 4.1. For both plots, a significant increase in the number of sources that pass the feature extraction process is found with a minimal upscaling factor. It is also seen that the number of sources with features extracted successfully drops off as the image is upscaled to higher percentages.

Figure 4.2 illustrates the results found from the adaptive scaling applied to the two data sets. A common trend is seen with the best results obtained from the small percentage increase in the number of pixels only, along with a steady decline as the percentage of pixels is increased to higher values. It is seen that the adaptive scaling provides a significant improvement to both data sets, with the benefit of returning the best results from the smallest changes to the images themselves. This reduces the risk of the adaptive scaling function affecting the actual ellipses that are fit during the feature extraction process as the images are as close as possible to the original images as can be.

4.3.3 Ellipse Fitting Errors Due To Large Image Sizes

Similar to the failures found in section 4.3.3, the failures due to images that are too large for the source in question is not a failure, but is also a case of the incorrect ellipse being fit. The solution to these failures is the same as the first adaptation that is implemented for images that are too small, namely using the ideal cutout size for the source in question.

The difference however, lies in the fact that there are no open ellipses or a similar type of flag that can be used when the images are too large for the source. Detecting these instances is thus extremely difficult, but can be easily remedied once identified.

Implementing the initial cutout size selection procedure reduces these failures, although the actual improvement is uncertain. These images are dealt with on a case by case basis when detected within the output of Astronomy.

4.4 Changes Made For Source Based Failures

This section addresses the challenges and failures of the feature extraction process as encountered in section 3.5.1, where the problems are due to the source located within the image itself. It is seen that in some instances, there is not much that can be done to correct the issues and that it is often best to remove these problematic sources from the data set during the data selection process.

In other cases, the failures stem from the method by which the images are displayed. Although this is related to the image, it is also dependent upon the source within. Some image transform methods cause feature extraction failures for certain sources where others do not. The transform methods also affect the detection rates of anomalous sources throughout the data set and are thus vital to optimise.

This section starts off with the short subsection 4.4.1, in which can be seen that some of the failures are actually caused by other objects within the image that lie along the line of sight to the source in question. Following this is subsection 4.4.2, wherein the failures due to faint sources is discussed. This section finishes off in subsection 4.4.3, where the investigation into the band weightings used is done.

4.4.1 Effect Of Nearby Bright Sources And Masked Sources

Section 3.5.1 showed that the feature extraction process failed when there are nearby bright sources along the line of sight to the object. This goes hand in hand with the cases presented in section 3.5.1, where it was seen that masked sources are included in the data set. Both of these are actually related to each other, since they are both masked sources. The difference lies in whether the masked source is the focus of the image, or whether it lies close to the source in question within the image.

The failures from these are mixed; in some instances the ellipse fitting fails completely but in others an ellipse is produced for the source, yet these ellipses will be affected by the masked sources. Sometimes the masked object itself is the source in question in the image and even passes the feature extraction process, but these sources are undesirable due to the irregular shapes that can occur. More often than not, masked sources are

stellar in nature and are thus not important for the work done in this thesis. As such, they can be discarded from the data set used. Fortunately, the majority of masked sources can be identified and ignored initially from existing flags set up in the catalogues.

It is worth noting that artefacts are also common within the DECaLS data set. Artefacts are caused by various phenomena, ranging from satellites passing through the observation to solar flares occurring at the time of viewing. They too will affect the ellipses fit and care has to be taken when they are encountered. Artefacts do not have a pre-existing flag within the catalogues to identify them, so encountering them is unavoidable. There are two options for dealing with artefacts, they can either be removed once identified which would require the outlier detection algorithm to be reapplied on the remaining features, or the active learning will have to be relied upon to handle the scoring of the artefacts.

4.4.2 Faint Sources Causing Failures

In section 3.5.1 it was seen that faint sources can cause feature extraction failures. The low signal to noise ratio is the main cause of this as the sigma clipping process used within Astronomy will not be able to identify the difference between the source and the surrounding noise easily. Even if the sigma clipping is done successfully, the remaining data often includes some of the surrounding noise and would thus result in incorrect ellipses being fit. This can be improved by fine tuning the sigma clipping process, but the variations are too vast for there to be a single set of ideal parameters for all of the sources.

From this arises an important aspect regarding the data; namely that of flux limitations. The majority of the DECaLS data is comprised of faint, low flux sources that cause the feature extraction process to fail. It is therefore important to identify what the flux limitations are for the DECaLS data set that can be investigated successfully. To identify what these limitations are, the compilation of DECaLS data bricks used earlier is investigated. In section 4.3.2 it was seen that 20 642 out of the 41 414 sources had features extracted successfully before implementing the adaptive scaling feature. The 41 414 sources were thus visually inspected, one by one, after being ordered in decreasing flux levels, to determine at what point the sources can not be identified *visually*. This complete inspection was also done to gain a better understanding of the DECaLS data itself.

The flux values differ for each band, so the inspection was carried out in two different ways. The first is to average the flux values across the bands and ordering the sources by this value and then inspecting them. The second is to use the maximum of the

three flux values and ordering the sources by this value. After ordering the sets by the flux values in descending order, the sources were inspected one by one until they became indistinguishable from the background. The outcome from this visual inspection indicates that when the average flux value reached 7 nanomaggies, features and structures within the sources became difficult to identify.

A nanomaggy is 10^{-9} times a maggy, which is the flux f of a given source relative to the standard, zero point source, f_0 . For a source with a given nanomaggy value f , it will have an apparent magnitude given by²:

$$m = 22.5 \text{ mag} - 2.5 \log_{10} f \quad (4.2)$$

Below 1 nanomaggy the sources are nearly impossible to distinguish from the background. For the second method of ordering, the same results were seen, but at levels of 10 nanomaggies and 1 nanomaggy respectively. As such, a requirement of having at least 10 nanomaggies in a given band can be placed on the data that can be investigated. For the compilation of bricks this corresponds to nearly 40% of the data that can be reduced.

4.4.3 The Impact of Band Weightings On Images

Images are composed of pixels, which can be numerically represented as an array with a value assigned to each pixel. For the DECaLS data files this corresponds to three arrays, one for each band, with a numerical value assigned to represent the flux level for each pixel of the image. Each of the three bands, g, r, and z, has its own array and when reproduced digitally, each band is assigned a different colour corresponding to the wavelength of the band. For the bands used, the colours are blue, green and red respectively. These *coloured* bands are then combined together to produce the output image displayed. They are often assigned a different scalar weighting that adjusts how much impact each colour or band will have on the displayed image.

These are called the *band weightings* and they affect the image directly. The weighting used for a specific band determines how much of that *colour* is displayed within the image. This is the visual interpretation of the band weightings, which can be seen in the images displayed in the front end of Astronomy. During the output process, the band weightings only affect the images visually. The affect on the output would affect the labelling process for active learning as the displayed images appear differently

²<http://www.sdss3.org/dr8/algorithms/magnitudes.php#nmgy>

depending on the weightings used.

Additionally, during the pre-processing stage of Astronomy, the image bands are not treated equally when sigma clipping is applied. By default, only the first band in the image is used, i.e., the g-band for the DECaLS data. A region in the image might contain high flux values in one band, but low values in the g-band that is used. The result from using the g-band only is thus a low SNR image that does not represent all of the available information. To prevent this, the bands should be stacked together before applying the sigma clipping function, or the sigma clipping function should be applied to all three bands simultaneously. The former is easier to implement and it is more common to stack the bands to create a single banded image.

However, it is not merely a case of directly combining the bands and displaying them. It turns out that certain band weightings are more beneficial to the feature extraction process while others make visual inspection easier. For example, there are standardised weightings used to create greyscale images or to replicate colours corresponding to what the human eye would see. Different weights produce different images which would produce varying results from the feature extraction process.

Band weightings are essentially scalar multiples applied to each band. Let \mathbf{G} , \mathbf{R} and \mathbf{Z} , represent the g-, r- and z-bands respectively. The greyscale and display weighted functions used in this section are represented by the following two equations respectively:

$$\text{Greyscale Function} = 0.1140 \mathbf{G} + 0.5870 \mathbf{R} + 0.2290 \mathbf{Z} \quad (4.3)$$

$$\text{Display Weighted Function} = 0.5357 \mathbf{G} + 0.2679 \mathbf{R} + 0.1964 \mathbf{Z} \quad (4.4)$$

The default method used by Astronomy stacks the images directly and scales the values from 0 to 1, 0 corresponding to the point with the lowest value and 1 to the brightest point [114]. From Equations 4.3 and 4.4 we see that the bands are favoured in different ways by the relevant weighting functions. To determine which function would be best to use, each one is inspected, first by inspecting the visual effect from each weighting function and then the effect that they have on the feature extraction process.

The reason for this is that measuring the impact of the band weightings is difficult to do. The only quantifiable measurement that can be made is the number of sources that have ellipses fit successfully; sources for which the feature extraction process is successful. This is thus used to measure the *performance* of the different band weighting

functions. Using this as a performance measure is ideal since we want the most number of sources returned, meaning that fewer sources are “lost” during the feature extraction process. The impact of stacking the bands together is investigated first. This is followed by a comparison of the different weighting functions. Then it is investigated whether an optimal function can be created that would perform better than the other weighting functions.

4.4.3.1 Visual Inspections Of Different Band Weightings

The output display of Astronomy is the starting point in order to determine what the ideal weightings should be. Changes within the weightings used produce a visual difference that is easy to see. The three different band weightings mentioned above are illustrated in Figure 4.3. The first image is the display as seen through the SkyViewer³, which corresponds to Equation 4.4. The second image is the default output from Astronomy, the image produced by matplotlib, while the last is the greyscale image given by Equation 4.3, produced by OpenCV [113].

It is immediately clear that the output from Astronomy differs significantly from the images of the same sources when viewed in the SkyViewer. For the gravitational lens in Figure 4.3, the arcs are difficult to see in the default output image and such important aspects can easily be overlooked. In the greyscale function, the arcs are more visible, but not easily identifiable due to the lack of colour information. Colour in itself assists in visually identifying features within an image and as mentioned previously, the colour of these images are produced by the weightings used for each band.

³<https://www.legacysurvey.org/viewer>



FIGURE 4.3: These images display the same source, but with different band weightings used. The first is the cutout image as seen directly within the SkyViewer, the middle image is the default display from Astronomy and the image on the right is the greyscale image produced using OpenCV's greyscale function. A significant difference is observed between the images. It is seen that the reproduced colours are important to visually distinguish between features.

From Figure 4.3 it can be seen that the weightings used by the SkyViewer produces an image that is clearer and in which aspects, such as the arcs, are more easily identifiable. The output from the SkyViewer is reproduced in Astronomy by the *display weighted function* as shown by Equation 4.4. This new function is used for the DECaLS data during the output process as it provides clearer images. However, measuring the impact of this display weighted function is challenging. The sources would all still be scored the same as they have the same features extracted. Only the output display function has changed. The only noticeable change from this would arise from the manual labelling for active learning, which is subjective in itself.

Since the various band weightings produce such different results within the output displayed images, the affect that these weightings would have during the pre-processing state is also investigated. The first change that arises is that the functions combine the bands into a single band, where as the base version of Astronomy focuses on the first band only. As mentioned before, not combining the bands is an issue that can cause information to be lost. Applying either the greyscale or the display weighted functions will provide a solution to this. The second change stems from the actual weighting values used. Different weights affect the images in different ways, which would affect the feature extraction process. This is investigated to determine what the ideal weightings are for the DECaLS data.

4.4.3.2 Single Channel Benefits For Feature Extraction

The greyscale weighting function is applied to determine the effects of stacking the bands before the sigma clipping function is applied. The results are compared to applying sigma clipping on a single band only. Only the greyscale function is used here

for the purposes of determining the effect of stacking the bands since the weighting functions are compared to each other in the next section.

Testing the greyscale function on the previously used data set of 500 brightest sources from a DECaLS data brick shows an increase in the number of sources that have ellipses fit successfully. Using the default method of Astronomy, utilising a single band only, 77 of the 500 sources fail the ellipse fitting process. This is reduced to 20 sources by stacking the bands using the greyscale function. When repeated on a larger and unconstrained data set however, the improvement is present but not to such an extent. For the brick compilation data set used earlier, a total of 20642 out of 41414 sources have ellipses fit initially. With the stacking applied this increases to 23174 sources only. Investigations into the sources that still fail the feature extraction process reveal that these sources are the ones that are very faint and small, indicating that the greyscale function performs better on brighter sources, but still provides an overall improvement. In both instances it is evident that stacking the bands together provides better returns from the features extraction process.

4.4.3.3 How Different Weightings Affect Feature Extraction

In the previous section, it is clear that stacking the bands together instead of using a single band provides a performance increase. However, the band weighting methods introduced both stack the bands, so investigations must be done to determine which, if any, provides quantitative results. In order to do this, a data set is set up specifically for this by including known anomalies. This allows us to see the effect that the weightings have on the anomalies and on the number of sources that pass the feature extraction process. It should be noted that this procedure was adopted so as to incorporate the anomaly detection aspect that the band weightings might have in addition to the number of sources that pass the feature extraction method.

This data set contains 15 000 random sources subject to a minimum flux level of 10 nanomaggies within each band to reduce the number of faint sources as per the results found previously. Additionally, the set contains 342 gravitational lenses; 60 high confidence lenses, 106 fairly confident lenses and 176 suspected lenses recently found by Huang et al., [87]. The lenses added are easily identifiable since they have different names. Astronomy is not a classifier or lens finder, but having known anomalies within the data set is needed to make the actual performance measurements and to determine whether any set of weightings provide an advantage for the anomalies.

Astronomy was first run on the data set using the greyscale function and then the display weighted function. For each function, the top 50 sources were investigated and the ranks of the lenses within the 50 sources was noted. The number of sources looked at was increased in increments of 50 until 250 sources were looked at. The ranks of the lenses are noted each time. This compares the number of anomalies that each function detects. In addition to this, the procedure was repeated several times with active learning applied, each time with a different amount of sources labelled. This compares the two functions with active learning applied.

The results are shown in Figure 4.4. For little to no active learning, the greyscale function tends to return more gravitational lenses than the display weighted function. When more training is applied the display weighted function tends to perform better. Looking at the top 250 sources, it is seen that the greyscale function performs better for all increments of active learning other than the last increment where 10% of the data has been labelled. This is the most important result since more than 250 sources will most likely be looked at for any data set. Labelling 10% of the data can also be time consuming for large data sets and is unlikely to occur.

In addition to these results, the greyscale function resulted in successful feature extraction for a total of 15012 sources while the display weighted function returned 14595 sources. Since the ranks of the lenses are known in the output, the Rank Weighted Score (RWS) can be calculated using Equation 2.8. It is found that the greyscale function has a RWS of 0.03828 and the display weighted function has a RWS of 0.03715 for this data set. The greyscale function has a higher RWS value, returns more sources and contains more anomalies within the top sources. For these reasons, the greyscale function is chosen as the preferred function to use during the pre-processing stage, with the display weighted function used for displaying the output.

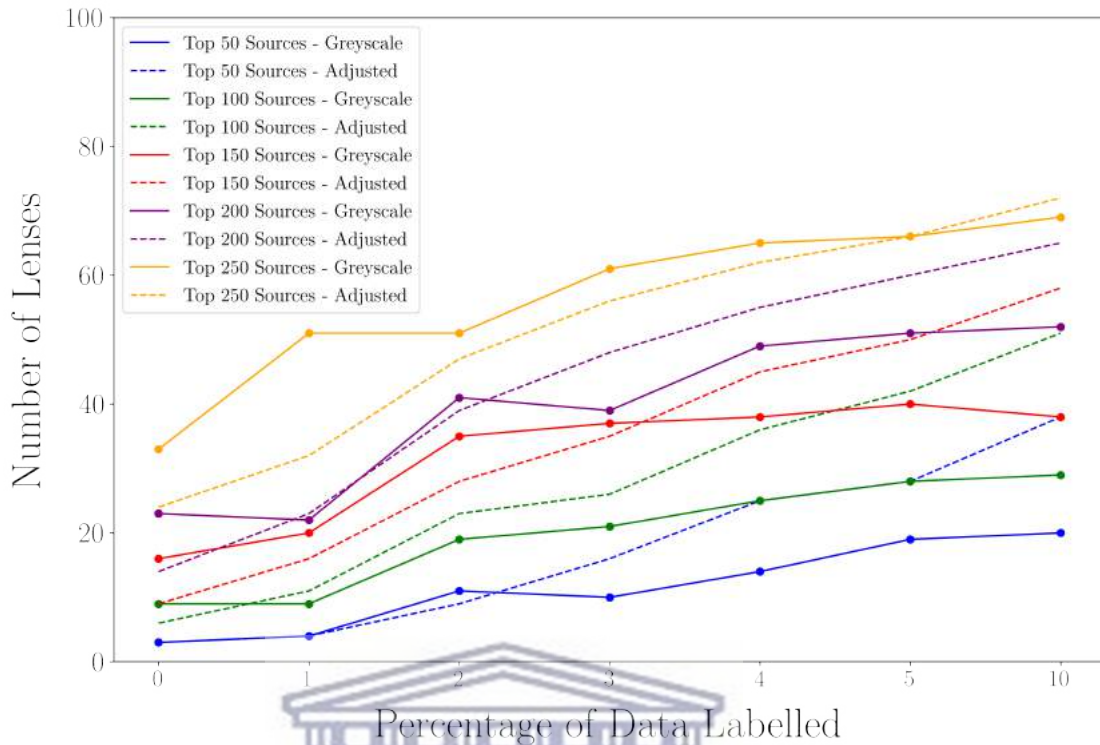


FIGURE 4.4: The plot illustrates the recall of the gravitational lenses for the two functions. The solid lines indicate the greyscale function, the dashed lines the display weighted function. Each line represents the number of sources looked at, i.e the Top 100 is determined by looking at the top 100 sources and counting how many lenses are located there. The results are cumulative with the previous one; the Top 100 Sources contains the same lenses as those in the Top 50 Sources. From the plot it is seen that the greyscale function outperforms the display weighted function when no active learning is applied. When 10% of the data is labelled, the display weighted function outperforms the greyscale function on all occasions.

4.4.3.4 Finding An Optimal Function

In addition to the greyscale and display weighted functions, a third “optimal function”, was created specifically for the gravitational lenses. This function was created by applying an optimisation function on 100 runs of Astronomy on the same data set of 15 000 random sources and the 342 gravitational lenses. The optimisation goal was set to maximise the RWS value returned. The parameters that were adjusted between each run were the weightings themselves. The optimisation function starts with a random value for each weighting used during the pre-processing stage. The RWS score is then calculated for that run. The weightings are then adjusted for the next run. This process is repeated in order to determine the best RWS value from all of the different weightings tested. For each of these 100 runs, the number of sources that pass the feature extraction process is also noted.

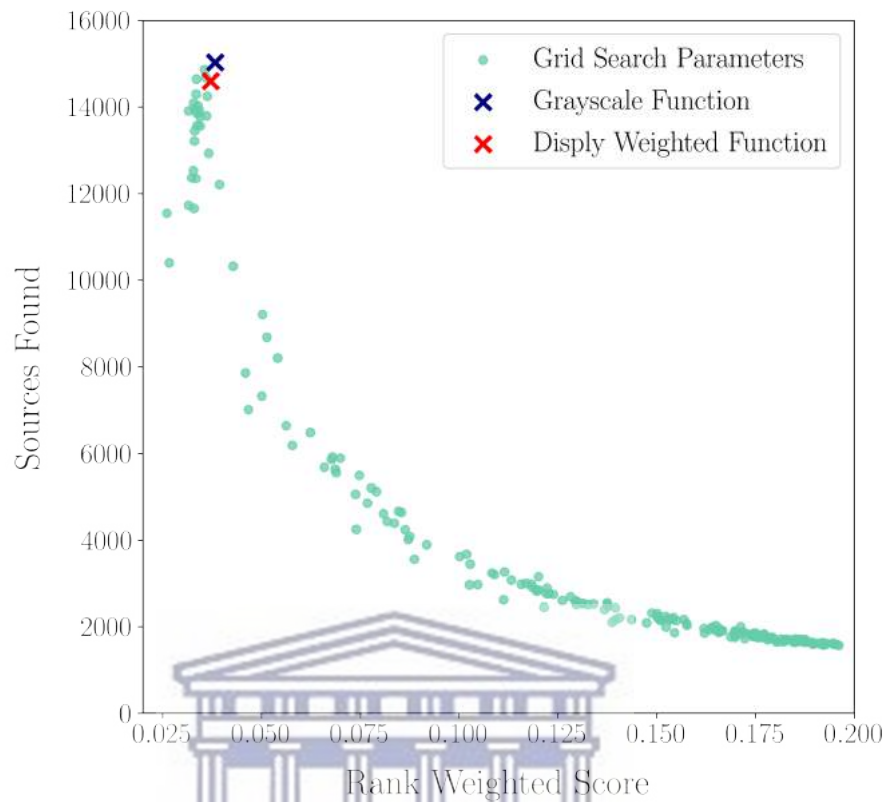


FIGURE 4.5: The figure shows the results from a grid search done to find the optimal weighting parameters. Both the greyscale function and the display weighted function have their corresponding values plotted as a comparison. The higher the Rank Weighted Score, the better the weightings perform in detecting the anomalous sources. However, there is a clear trade-off between the RWS value and the number of sources that pass the feature extraction process. This must be taken into account when considering which function to apply.

Figure 4.5 shows all 100 of these searches. It is seen that the majority of the runs produce higher RWS values than both the greyscale and the display weighted functions. However, the number of sources returned drops off exponentially as the RWS value increases. This “loss” is too significant to accept as even a small improvement in RWS value causes a large decrease in the number of sources returned. Out of the 342 lenses present within the data set, 81 failed the feature extraction process for the run with the highest RWS value. This is nearly a 25% loss in the number of gravitational lenses detected. Such a significant loss in detection rates is not acceptable and so this optimised weighting system is not used.

This *optimal function* is found to have the following band weightings:

$$\text{Optimal Function} = 0.0255 \mathbf{G} + 0.5720 \mathbf{R} + 0.4025 \mathbf{Z} \quad (4.5)$$

Inspection indicates that the optimal function has a strong z-band weighting which corresponds to red and a weak g-band weighting, corresponding to blue. This increases the fainter regions within an image and decreases the actual source, up to the point that the noise is blended with the source. Sigma clipping would then clip away most of the data within the images and there would not be enough points remaining for the ellipses to be fit successfully and so the feature extraction process fails. This also explains why the RWS are higher since these lensed sources tend to have less blue in them; the flux values in the g-band are seen to be lower throughout, due to the higher redshifts of the lenses in question.

4.4.3.5 Discussion On The Band Weightings Used

It is seen that the display weighted function provides the clearest images when used during the output process of Astronomy. It is thus used during the output process when applying Astronomy on DECaLS data. Stacking the bands together before applying sigma clipping provides an increase in the number of sources that have their features extracted successfully. The improvement amount varies, with the best performance increase seen for brighter sources.

The actual features that have been extracted would also differ depending on the weighting function applied. The band weighting functions would alter the images in such a way that different values would be obtained for the features extracted. The outcome, or variations in the features, have not been investigated directly, but it stands to reason that the features for all of the sources would be affected in a similar manner. As such, investigating the final results provides an indication of the performance of each band weighting function since it is dependent upon the features extracted. Comparing the greyscale function to the display weighted function shows that the former returns more lenses when there is no active learning applied, or when little active learning is applied. The display weighted function returns more lenses when a larger number of sources are labelled. It is possible to label a higher number of sources, but it becomes impractical when large data sets are used. For instance, the data set used here requires 1534 sources to be labelled if we were to label 10%. This takes a significant amount of time to do, and it would take much longer for larger data sets. For smaller data sets it might be preferable to implement the display weighted function during the pre-processing stage, but for larger data sets the greyscale function would be better.

An optimal function was found within the 100 Astronomy runs. However, the number of sources that failed the feature extraction process is too high to implement this optimal function. Since it is based on gravitational lenses, this optimal function might serve a better purpose as a classifier for lenses, but this requires further investigation as a significant number of lenses failed the feature extraction process as well. Other anomalous sources might also fail when this optimal function is used. For this reason, this function is not applied in any way further in this thesis.

4.5 Limitations Of Outlier Detection Algorithms

Astronomy has only been applied to tens of thousands of sources. While the feature extraction is trivial to parallelise, the algorithms usually need to consider all the data at once. Thus the capabilities of these algorithms was tested before attempting to implement them on large data sets.

Mock data consisting of random Gaussian distributions are used to simulate the features that would normally be obtained from the data. These mock features can easily be controlled by adding additional dimensions to each data point and outliers or anomalies can be recreated by positioning them far enough away from the norm. Four groups of Gaussian distributions are created, the “norm” consisting of 99% of the sources and three outlier groups that combined consists of 1% of the total number of sources, representing the anomalous sources. Figure 4.6 illustrates the groupings for a two dimensional case.

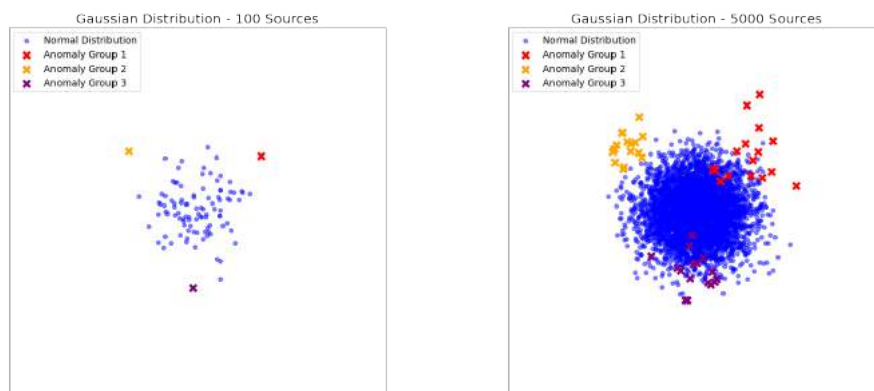


FIGURE 4.6: These plots show the way in which the groupings are made for a two dimensional case. The first shows the distributions for 100 sources. In it there are three anomalous sources but there are other sources that form part of the norm which can be mistaken to be outliers. The second shows the groupings for 5 000 sources and in it we can see that there exists an overlap between the anomalous groups and the norm group which can make detecting the outliers difficult.

Two different tests are done for each of the machine learning algorithms and are covered in the following sections. The first test for each algorithm is to determine the limit of the number of sources that can be investigated, while the second test focuses on the dimensionality of the data. The first test for each uses two-dimensional data, but increases the number of sources until such a point that the limits of either the algorithm itself or the processing power available is reached. The second test uses a fixed number of sources, 50 000, but increases the amount of dimensions that each point has until the limits are reached.

4.5.1 iForest

The key limitations to be tested are the memory usage, which would directly affect the number of sources that can be investigated, and the runtime, which could play a limiting role as well. Along with these limitations, the accuracy, recall and precision was determined for the algorithm as well.

Amount Of Sources Test

The results from the first test, for the amount of sources, can be seen in Figures 4.7 and 4.8. The number of dimensions was kept constant for this test, only the number of sources was increased.

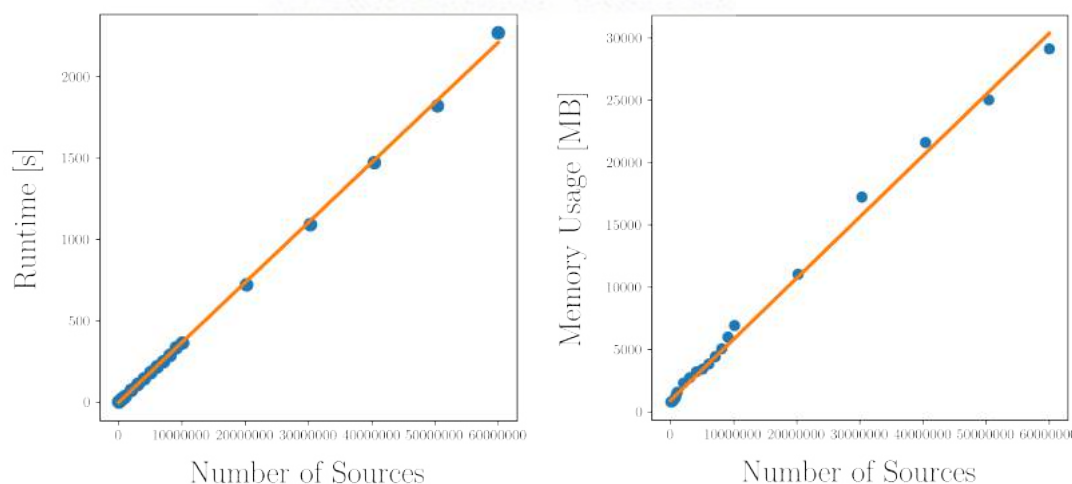


FIGURE 4.7: Both the runtime and the memory usage plots shown here indicate that the iForest algorithm scales linearly with an increasing number of sources. This is an important result as it allows easy yet accurate estimates to be made to determine what data size can be run and how long it would take.

The straight lines for the memory usage and runtime plots are best fit using the least squares method. The data points are not evenly spaced due to the way that the test was run; the sources were not increased in equal steps throughout the test. This does not affect the results since the same data format and algorithm is used each time. From the plots it can be seen that the main limiting factor for the iForest algorithm is the amount of memory used. For the amount locally available, this corresponds to a bit more than 60 million sources, with a corresponding runtime of about 40 minutes.

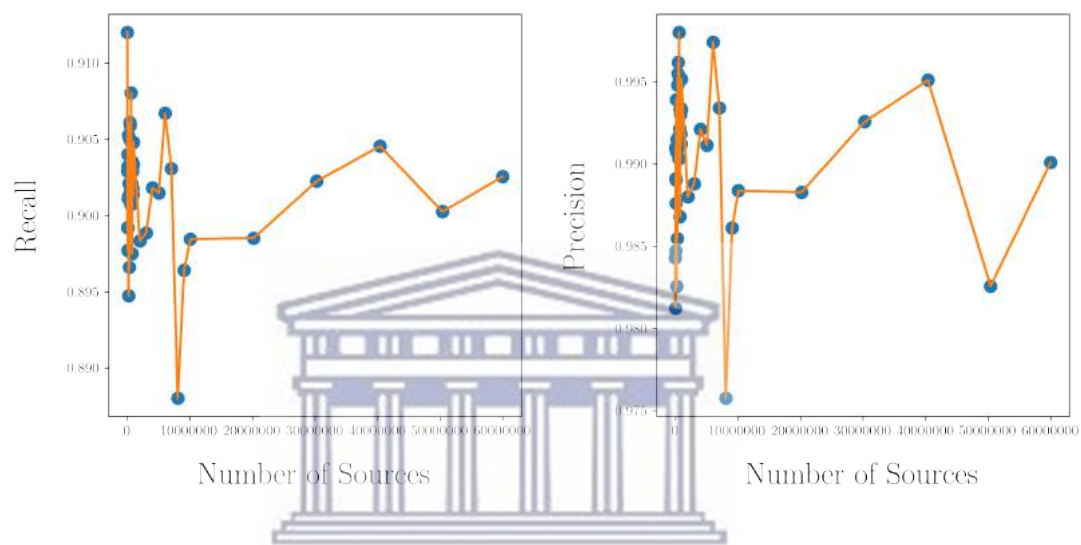


FIGURE 4.8: All three plots appear to indicate good results from the test. However, it is actually a case of overfitting by the machine learning algorithm that creates these results. This is indicated clearly by the lack of trade-off between the precision and recall that is usually seen by a machine learning algorithm.

Figure 4.8 indicates that there were issues in the set up of the data. The variations seen throughout is a strong indication of a simulation that contains significant overlap between the anomalies and the normal sources.

Number Of Dimensions Test

The same limitations were tested, but on a data set with a constant number of sources (50 000). The dimensions were increased and the results can be seen in Figures 4.9 and 4.10. The dimensions were increased in steps of one dimension for each run.

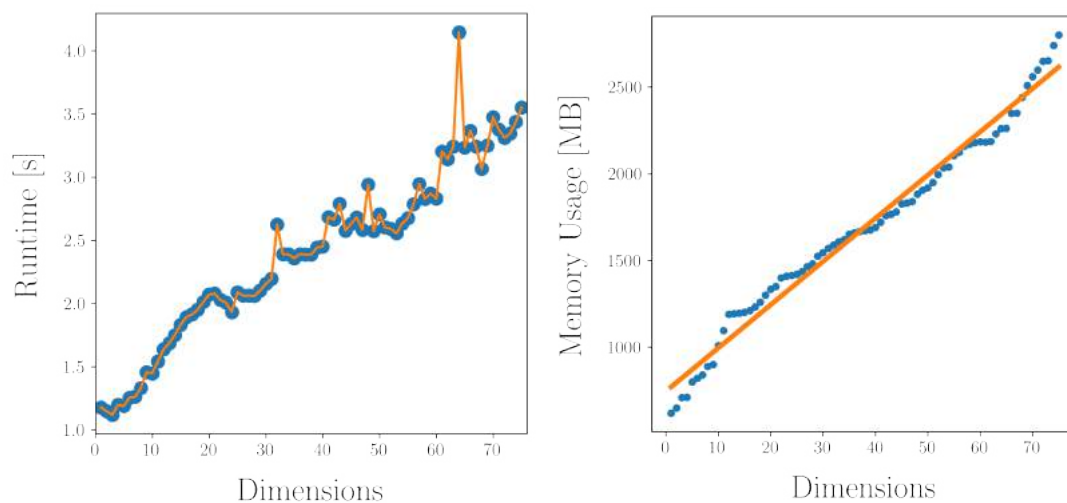


FIGURE 4.9: A clear linear relation is seen in both plots. The first plot for the runtime has a few spikes throughout but closer inspection shows that the runtime itself is very short and that these spikes are less than a second long. They are most likely due to some external factor. For the most part the memory usage appears linear, although there is a sharper increase towards the end of the plot. The ellipse fitting feature extraction method of Astronomy used 24 features, which is within lower part of the plot.

Both plots indicate mostly linear relations between the runtimes and the number of dimensions, as well as with the memory usage and the number of dimensions. The plot comparing the runtime to dimensions has a lot more variance in it, but the overall trend is still linear. The runtime values are so low that the variances are less than a second in length. Once again it is found that the limiting factor is the amount of memory available for use.

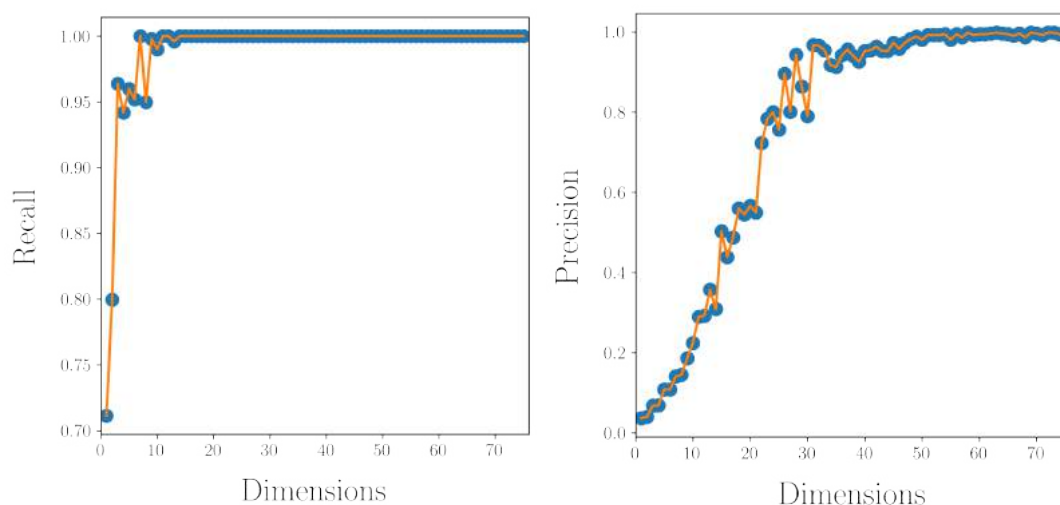


FIGURE 4.10: At a high enough dimension, all three plots reach perfect scores. This is an unlikely result and is indicative of a model that is overfitting the data.

It is seen that the model over-fits, due to the high dimensionality making it easier to detect the anomalies, and so the results can not be treated as being completely reliable. This is a clear result of the curse of dimensionality, where each additional dimension separates the sources more and more, making it easier and easier to identify the anomalies.

4.5.2 Local Outlier Factor

Another machine learning algorithm present within Astronomy that can also be used to detect outliers or anomalies is the LOF algorithm. The limitations of this algorithm is tested in the same manner that the iForest algorithm was tested.

Amount Of Sources Test

Testing the sources and dimensions was carried out in the same way and using the same data as used previously. The results can be found in Figures 4.11 and 4.12.

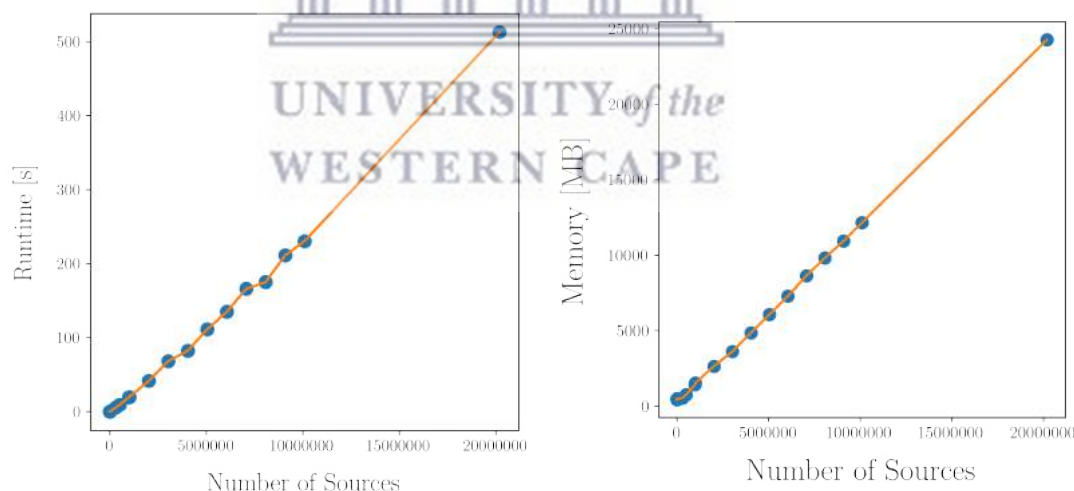


FIGURE 4.11: Both of the plots have a linear relation for the number of sources used. The runtime is shorter than that of the iForest, but closer inspection shows that the algorithm only managed to complete up to 20 million sources before reaching the computational limits.

The LOF algorithm uses almost three times as much memory as the iForest algorithm, severely constraining the number of sources that it can be run on locally. Similarly to the iForest algorithm, the LOF algorithm overfits on the data. This too suggests that it is the data itself that causes the issues to occur, not the algorithms used. In Figure

4.12, it is seen that the recall drops to zero when the number of sources is increased significantly. The cause of this is the set up of the data set; the “anomalies” start to overlap with the “normal” sources as the number of sources in total is increased. Fewer anomalous sources are detected by the algorithm as all sources tend to become similar.

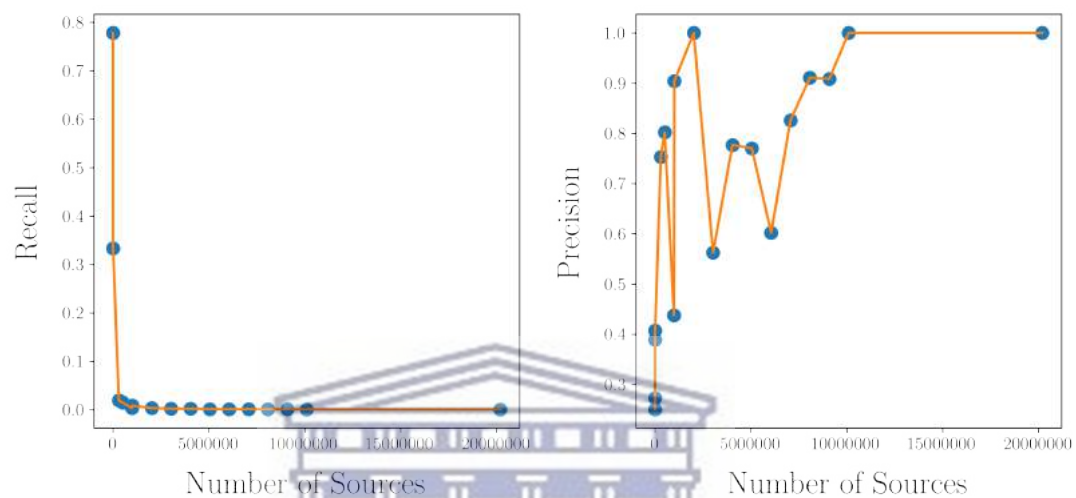


FIGURE 4.12: It is seen that LOF fails for a large number of sources, most likely a result of not optimising the k parameter.

Number Of Dimensions Test

The dimensions were tested in the same way as the iForest test and the results can be seen in Figures 4.13 and 4.14. These are the first nonlinear results seen for both the runtime and the memory usage. The runtime increases exponentially and then drops and flattens out. At the same number of dimensions the memory usage increases significantly and flattens out as well. The cause of this peculiarity is unknown and warrants further investigation, although it appears to be related to the number of neighbours looked at.

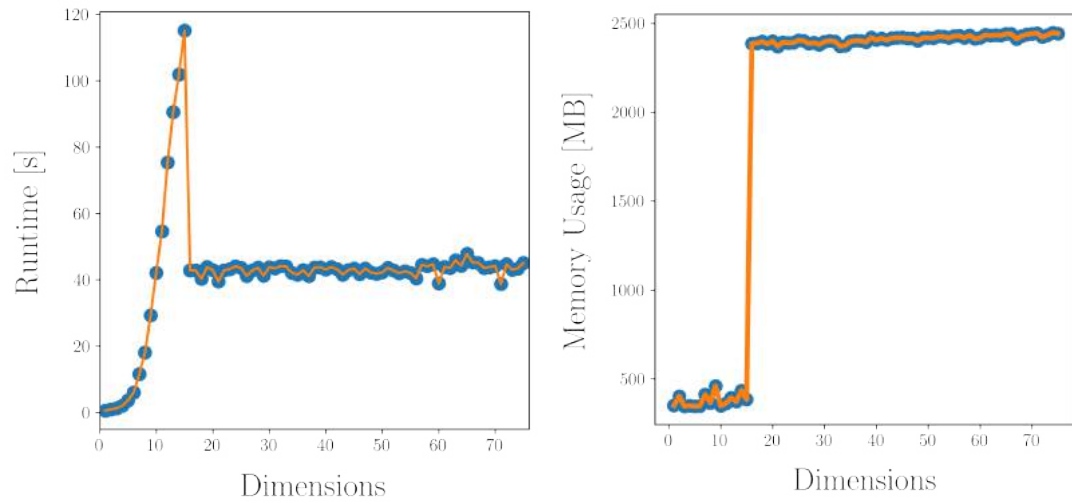


FIGURE 4.13: Plots of the runtime and memory usage against the number of dimensions. A peculiar drop in runtime and a corresponding jump in memory usage is seen within the plots. Besides this interesting feature, the memory usage appears to increase linearly by a small amount only while the runtime initially increases exponentially.

The LOF algorithm can be run on significant numbers locally, but the dimensionality of the data might cause some issues.

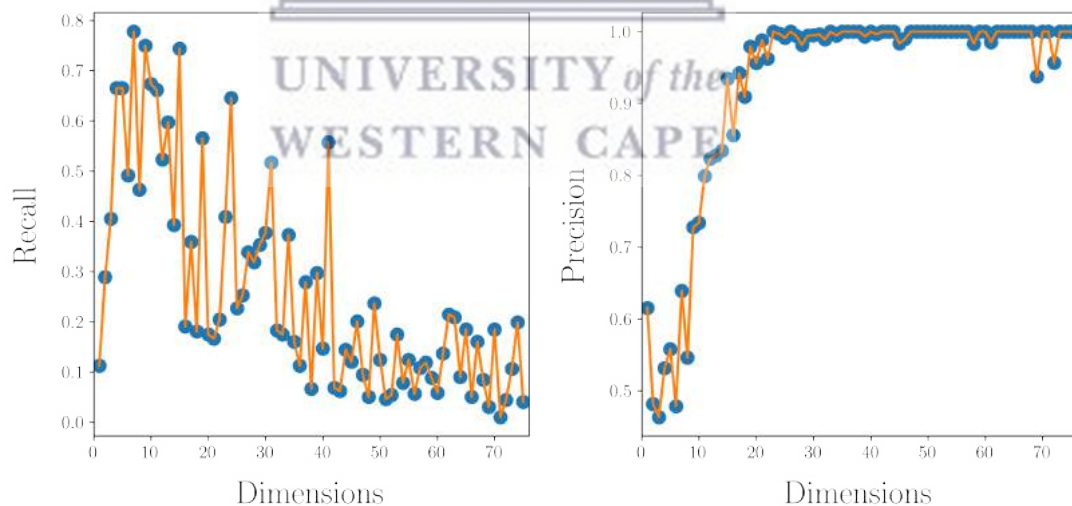


FIGURE 4.14: Recall and precision plots for the LOF algorithm against the number of dimensions used with a constant number of sources (50 000). Unlike the runtime or memory usage plots in Figure 4.14, there are no clear peculiarities and no sign of any change at the dimension where the peculiarity occurred in Figure 4.14. When compared to the iForest algorithm in Figure 4.10, it is seen that the recall drops off at higher dimensions for the LOF algorithm. This is a result of the reliance of the LOF algorithm on distances, causing it to struggle with high dimensional data.

4.5.3 RWS - Variations Between Runs

Another key feature tested is the consistency of the iForest algorithm itself. iForest creates its branches based on random values selected from the features themselves, which could result in a different anomaly score for a given source each time iForest is applied to the data set.

A good baseline to test this with is the data set contained within the Astronomy paper [115], for which the anomaly scores obtained for the sources are available. As detailed in the paper, the data set used consists of all objects with a Class 6.1 score greater than 0.9, meaning that 90% or more of the volunteers of the Galaxy Zoo project labelled the galaxy as odd. The result is 924 anomalous sources out of a total of 61578 galaxies. It should be noted that these sources are not expertly labelled and identified and that there could be other anomalies located within the data set as well that have just not been labelled in such a way as to fall within the selection criteria.

The performance of the algorithm can be determined by calculating the RWS, 2.8, based on these anomalies. This is not typical for unsupervised machine learning since the data sets are usually unknown, but it is useful as it allows the performance of the algorithm to be evaluated for different parameters and runs.

Testing the consistency of iForest involves applying it to the same data set, using the same parameters, and collecting and comparing the results. This is done on the data set described above. Astronomy is applied ten different times, each time using the same pipeline as the one used in the paper. The scores are collected and compared to each other and to the scores that were obtained within the paper as well. A part of the results can be seen in Table 4.2, which illustrates the scores given to five random sources for the different runs of Astronomy. It is clear that there are variances with each run that arise from the inherent randomness within the iForest algorithm.

TABLE 4.2: This table contains the scores given to the first five sources by the ten different runs of iForest on the data. Each run produces a different score for each source, but the overall variations appear consistent between the sources. The ones ranked higher are consistently higher.

Index	100008	100023	100053	100078	100090
Run 1	0.134979783	0.262751067	0.209853262	1.858548684	0.302624109
Run 2	0.119456987	0.392547851	0.257869102	1.830837854	0.29262809
Run 3	0.082700135	0.395604346	0.21600787	2.30294151	0.302855328
Run 4	0.081143936	0.332077173	0.227049578	1.953016112	0.128202146
Run 5	0.080605902	0.394849011	0.223891892	2.228423805	0.210845819
Run 6	0.107863126	0.308342155	0.205533798	1.957506749	0.193213866
Run 7	0.057889005	0.441916366	0.281334987	2.303204822	0.159698612
Run 8	0.089964938	0.461069973	0.224150421	2.204988806	0.195663398
Run 9	0.141601234	0.351550062	0.188050934	2.056108123	0.353090826
Run 10	0.117164409	0.407346612	0.212804181	2.080952645	0.268128672

iForest contains numerous parameters that can be adjusted to tweak the performance. The ten runs were done for some of the parameters to determine if they also produce varying scores. The parameters used are the bootstrap parameter; meant to hone in on the results more accurately by re-sampling the data, the verbose parameter; which controls how strictly the tree is built each run, and the random state parameter; meant to reproduce the same results each run when set to a specific value. Each parameter was set individually and ten runs were performed on it. The scores varied for each parameter set and for each run for each parameter. A quick summary can be seen in Table 4.3, which shows the scores given to a source for each of the ten runs made with different parameters.

TABLE 4.3: Scores given by iForest to the source with index value 100008 for each of the ten runs done for each parameter setting. All of the values differ somewhat, even the random state parameter which is meant to prevent this randomness from occurring.

	Default	Bootstrapping	Verbose	Random State
Run 1	0.134979783	0.112538964	0.089096976	0.056243771
Run 2	0.119456987	0.056026746	0.116359405	0.130954996
Run 3	0.082700135	0.076762931	0.10848139	0.199757334
Run 4	0.081143936	0.135557938	0.198903614	0.087315227
Run 5	0.080605902	0.106083643	0.153036701	0.041529418
Run 6	0.107863126	0.134913541	0.076614419	0.06816127
Run 7	0.057889005	0.178257503	0.143586869	0.107148105
Run 8	0.089964938	0.138754332	0.159743844	0.123728335
Run 9	0.141601234	0.176248253	0.069903531	0.096207762
Run 10	0.117164409	0.100195132	0.084830371	0.085212271

All of the scores differ from each other although the general trend from one source to the next seems to remain the same. Even if all of the scores differ each time, it is still possible for the results to be the same if the sources are returned in the same order consistently. To fully determine whether or not there are any differences in the results, the RWS is calculated from the known anomaly list for various parameters of the iForest algorithm.

This can be seen in Figure 4.15 which shows that the RWS for each parameter differs when compared to the RWS obtained in the Astronomy paper.

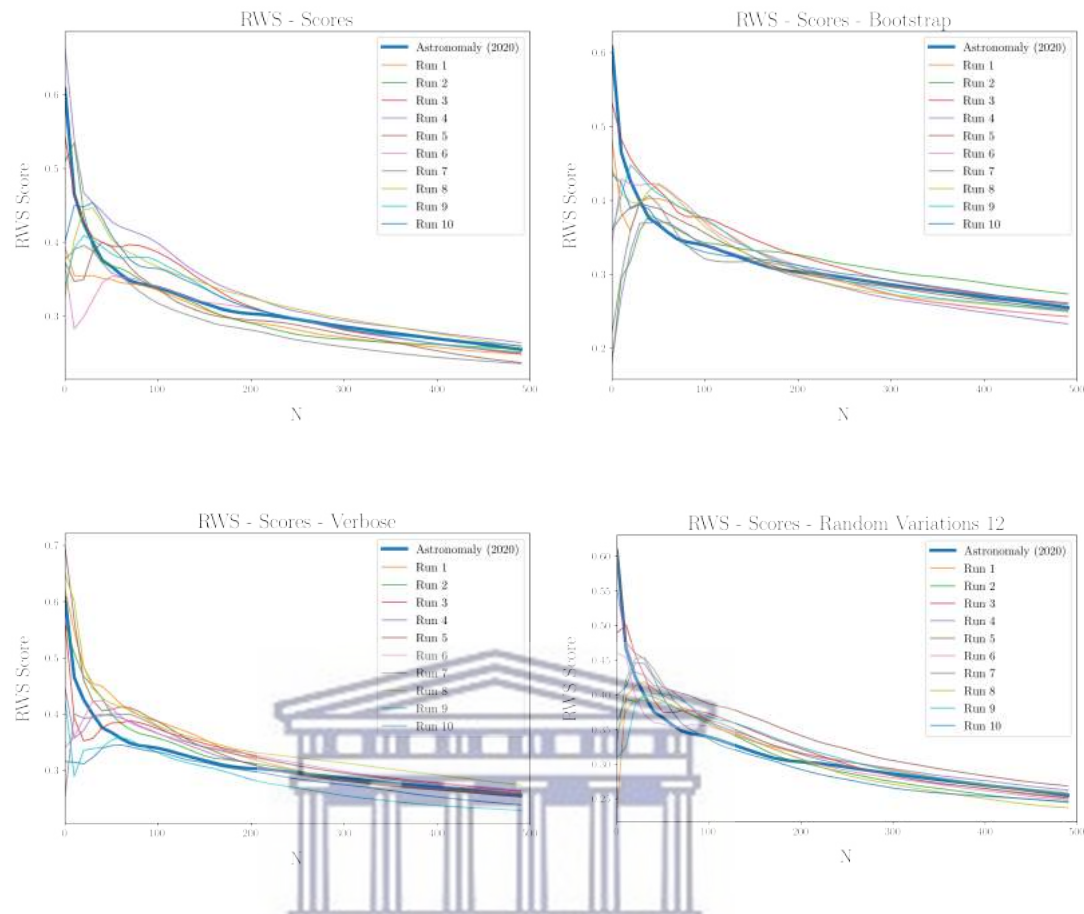


FIGURE 4.15: The four plots shown here are the RWS values obtained for each of the parameters used. The thicker blue line is the RWS from the original Astronomy paper [115]. It is seen that the RWS values differ slightly for each run for all of the different parameters used. It appears to differ greatly, especially when N is low, but this is only due to the way that the plot is made; the plot is more sensitive to variations when N is small. The variations are minor as N increases, suggesting that the individual runs are similar in nature and do not produce significantly different results. However, the most surprising result from these plots is that of the random-state parameter. This is meant to replicate the cuts made each time and should thus result in the same scores for each run but it does not.

A more important aspect of the differing runs would be whether or not the number of anomalies within a certain number of sources, N , differs between the runs. Since all of the N sources would be looked at, minor positioning differences would not be that important compared to a different number of anomalies within the N sources. The number of anomalies within the top N sources is thus calculated for each run to determine whether the different scores result in a different number of anomalies seen. The result of this is seen in Figure 4.16.

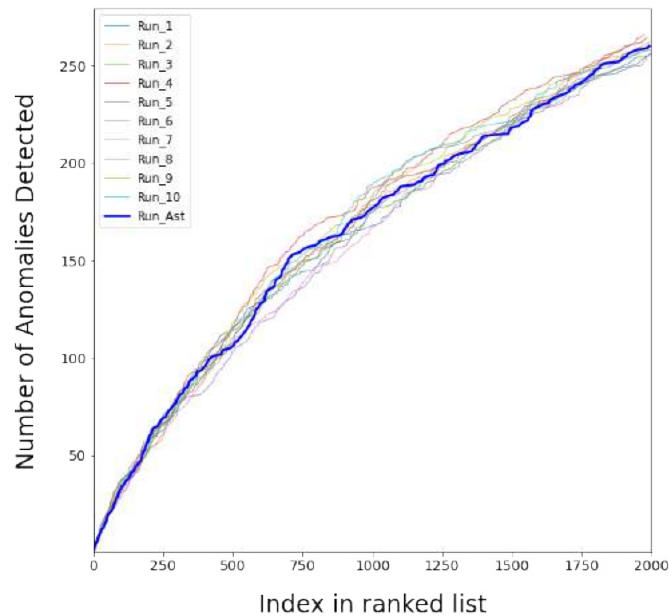


FIGURE 4.16: This shows the number of anomalies found within the top N sources for each run of the default parameter settings. The thicker blue line indicates the original Astronomy papers result.

There is a small difference for the number of anomalies detected between the runs, but the variations are minor compared to the overall number of anomalies detected by the algorithm. Averaging multiple runs would be ideal to compensate for this and could be used to estimate the error margin, but this would be time consuming so the cost must be considered beforehand. Active learning still needs to be applied on these runs and could make up for the minor variations seen.

The other important aspect to note is that Astronomy is designed to make detecting *more* sources more easily, not to perfectly predict all of the anomalies within each data set. Astronomy is unsupervised and is often applied to unknown data where the actual anomalies are not known. In such instances it would not be possible to determine how many anomalous sources are *not* within N sources for each run without looking at all of the sources in the data set.

The precision and recall values for the ten default runs were also determined to illustrate the differences and can be seen in Figure 4.17. Again, it would be difficult to calculate these values for unsupervised learning when the actual labels are not known.

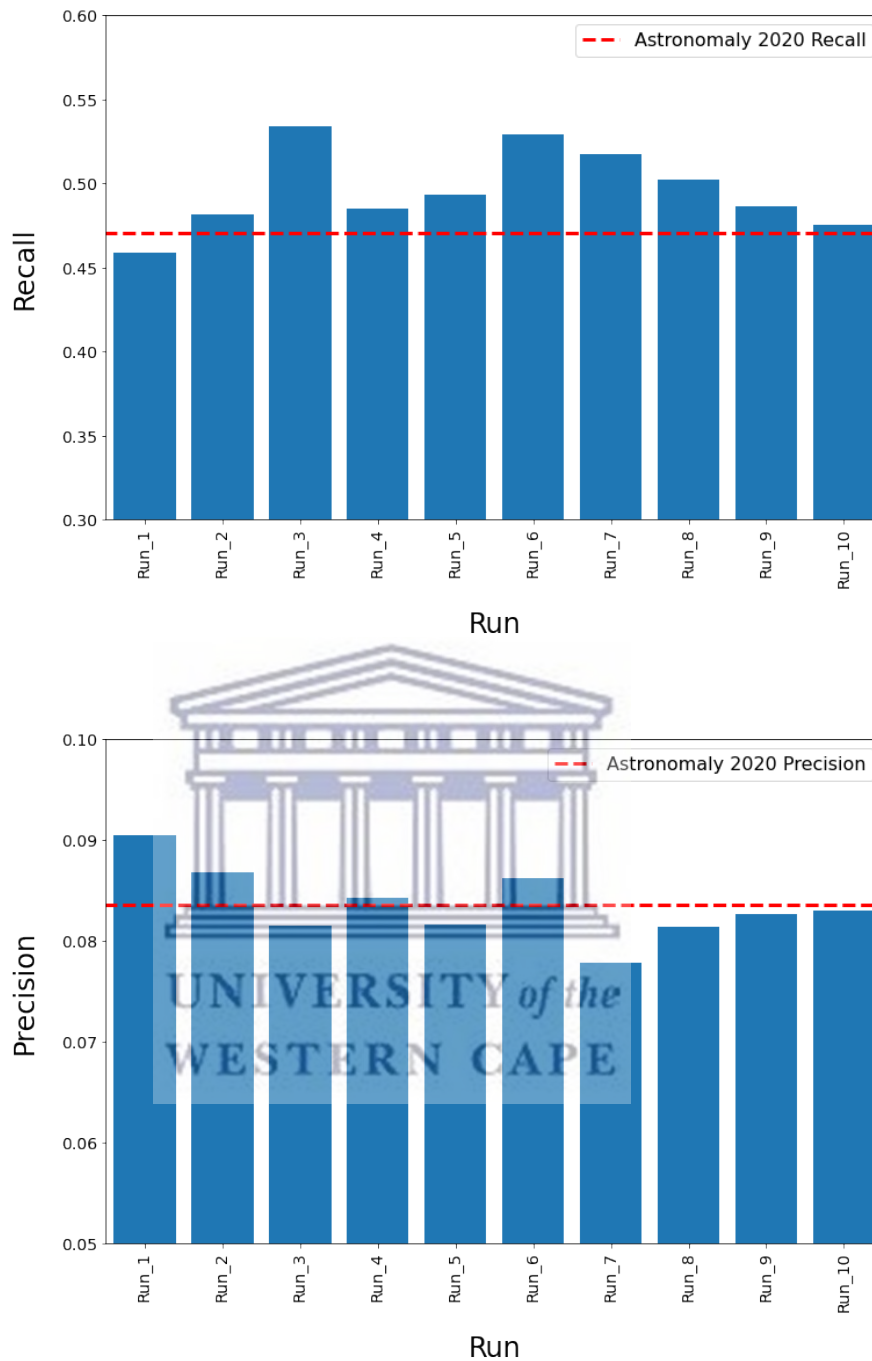


FIGURE 4.17: The recall and precision values for each run of the default parameters are shown here compared to the recall and precision of the original Astronomy paper, shown by the dotted line. Both plots indicate different values for each run.

4.6 Improvements Due To The Changes Made

This small section finishes off the chapter and provides a quick summary of the changes made to Astronomy and also illustrates the improvements gained from the changes and adaptations made.

Data selection is restricted before even obtaining the data. Point sources and masked sources are removed based on the flags available within the DECaLS catalogues. A minimum flux level of 10 nanomaggies is also adapted since this will reduce the amount of feature extraction failures. These criteria are all applied when selecting the data to be used within a subset.

Image pre-processing is key to producing the optimal input for the feature extraction process. It was found that joining the separate image bands together, and using the greyscale weighting values when doing so, produced the best results for the rank weighted score and returned the most sources with features extracted successfully.

During the feature extraction stage, adaptive scaling is implemented to increase the amount of data points available for a specific image. This reduces the amount of feature extraction failures.

After the feature extraction process, a list of sources is created that indicates a need for larger image sizes for these sources. The sources are thus downloaded again with the recommended image size and the feature extraction process is applied on all of the sources again.

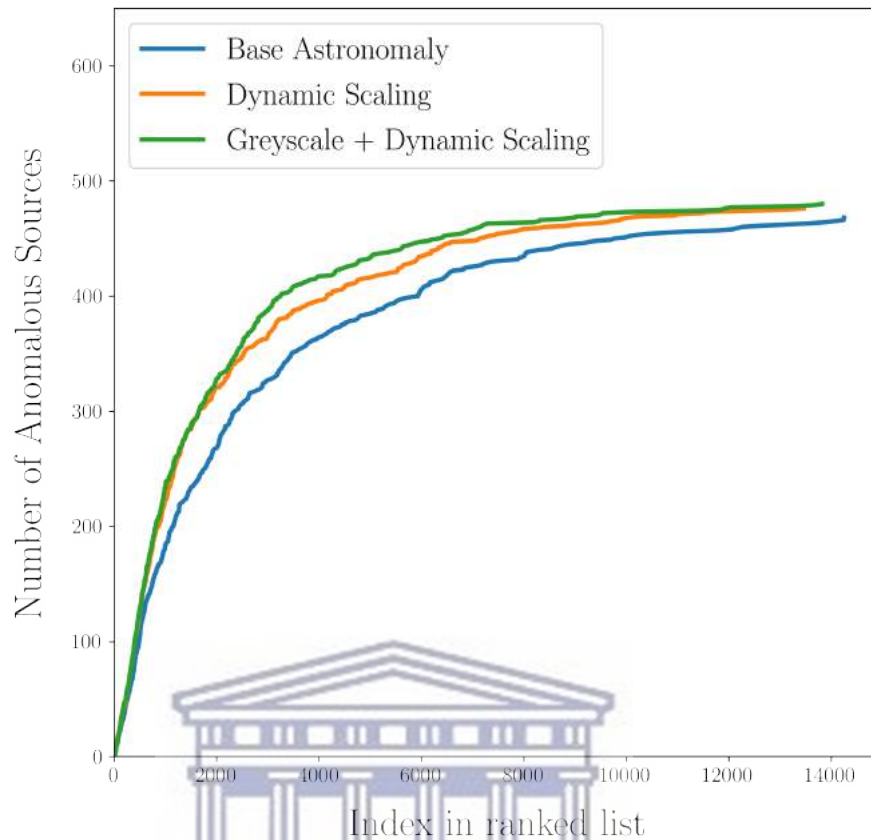


FIGURE 4.18: This figure shows the improvements gained from applying the two major changes to Astronomy. The increase in the number of sources detected in substantial in both instances.

Figure 4.18 illustrates the gain obtained from implementing the adaptive scaling and the greyscale weighting system on the known anomalous sources within the lens data set, see section 5.2 for a description of the data set. The adaptive (dynamic) scaling was implemented first and then the greyscale function was introduced in addition to the scaling. It is clear that both of these adaptations create a substantial increase in the anomaly detection rate. From the figure it is also seen that a higher overall value is reached, corresponding to more sources with features extracted being returned.

Table 4.4 contains the number of sources that have ellipses fit successfully for each adaptation implemented in Astronomy. The results are mixed in the sense that the Greyscale function provides a better improvement over the Dynamic Scaling for the Lens set, but otherwise for the compilation of bricks, see section 4.3.2. This is most likely due to the composition of the two data sets, whereby the lens set is more “refined” and has the lower flux limitation of 10 nannomaggies implemented, but the compilation of bricks does not. This also explains why the Dynamic Scaling has a larger impact on the compilation of bricks as it affects fainter sources more.

TABLE 4.4: The table indicates the number of sources with features that have been extracted successfully for the changes made.

Changes Made	Compilation Of Bricks	Lens Set
Base Astronomy	20642	14512
Greyscale Only	23174	14935
Dynamic Scaling Only	27113	14717
Greyscale and Dynamic Scaling	30401	15012
Total Number Of Sources	41414	15342

The dynamic scaling implemented is quite specific to the failures resulting from the Ellipse Fitting feature extraction technique. It improves results, but only because the failures arise as a combination of the feature extraction method and the nature of the data used.



Chapter 5

Results of Applying Improved Astronomy On DECaLS Data

5.1 Introduction

The previous chapter detailed the changes made to Astronomy and ended with a section illustrating the effects that these changes have individually. This chapter focuses on the results found when applying Astronomy with these changes applied to a few subsets of the DECaLS data set. In section 5.2, an overview of the subsets is given. Some of these have been used in previous chapters, but have not yet been described in detail as they are here. Section 5.3 details the procedures used for testing. The various tests done are described here, along with the goals of each test and what will be used to determine the performance of the algorithm. The results from these tests are illustrated and discussed in section 5.4, where the changes made to Astronomy are tested against the base version, and section 5.5, where the active learning of Astronomy is tested. The chapter ends off with section 5.6 where some of the more interesting sources found within the subsets are discussed. A comparison to known galaxy mergers is also made here.

5.2 Data Sets Investigated

The subsets used within the thesis are all derived from the DECaLS data set. Some restrictions determined in previous chapters are applied to all of these data sets, but some subsets have specific selection criteria. Point sources and masked sources have been removed before the sets were created based on the flags available within the catalogues themselves. This successfully removes all point sources before selection is made, but the flags do not incorporate all of the masked sources within the data

set. This results in entries that exist within the different subsets that are still masked sources and some artefacts are also still present. Although methods exist that can remove most images of sources that contain traces of artefacts, this has the risk of also removing potentially interesting anomalies. Instead, we use the active learning part of Astronomy to handle artefacts. All three subsets used here are obtained from the Southern Hemisphere within the DECaLS data set since this side is typically less explored.

Subset 1: Brightest Subset

This subset has been created by selecting the 10 000 brightest sources within the Southern Hemisphere of the DECaLS data set. The highest flux value out of all three bands for each source was used as the selection criteria. Unfortunately this has resulted in some masked sources being included as they typically have high flux values. The reason why the brightest sources were chosen was based on the idea that they would be the most visible and thus be the easiest to identify. Most of the anomalous sources within this data set turn out to be ordinary spiral galaxies that have bright bulges. Some of them turn out to be galaxy mergers, or show signs of interactions with other galaxies, but it is seen that there is a fundamental issue throughout the data set; the full sources are not displayed properly within the cutout sizes used. This data set is thus not reliable and its shortcomings will be discussed in the conclusions section, with this also being the reason why some of the results presented later on do not include data from this subset.

Subset 2 : Random Subset

The second subset consists of a random selection of 10 000 sources located within the Southern Hemisphere. Point sources and masked sources have been removed as mentioned earlier. The only other selection criteria made on this subset is to restrict the lower flux levels so that they have to be at least 10 nanomaggies in the bands. Previous testing has shown that this reduces the amount of feature extraction failures significantly and is key to visually identify the sources in question. It is expected that the majority of the sources will still be on the fainter side, although the random selection should indicate an even representation of all flux levels within the DECaLS data set.

Subset 3: Lens Subset

The last subset created is similar to subset 2, except that it contains a total of 15 000 randomly selected sources and includes an additional 342 known and suspected gravitational lenses. This is done due to the rarity of such sources; testing whether the algorithm works on lenses would be extremely difficult if only one or two are available. It is also useful to have a fairly large number of known anomalous sources as this allows the performance to be determined more easily. Even though this ratio of gravitational lenses to other sources is unusually high and very unlikely to be found naturally, this set does provide a better means to test the anomaly detection rate. It is worth noting that the aim is to find all anomalous sources, not just gravitational lenses, although they do provide a useful means to test the algorithm.

5.3 Testing Procedures And Performance Measurements

The sources within all of the subsets have been inspected manually beforehand in order to create a list of interesting and anomalous sources for each set. This is required in order to determine the performance of the algorithm. Without knowing which entries are interesting, there is no consistent method to determine the performance of the algorithms and the Rank Weighted Score for each data set could also not be calculated without this.

Additionally, Astronomy is not a classifier and the *order* that the sources are returned in is the backbone of the performance of the algorithm. It is not meant to identify sources but rather to return the more anomalous sources in an ordered list so that it is easier and quicker to identify the majority of them. The locations of the interesting sources are thus used as the main performance measure of the algorithm. Along with this, the number of sources with features successfully extracted is also used as a performance measure.

The first set of results are aimed at illustrating the improvements obtained from the adaptations and changes made to Astronomy as described in chapter 4. For each data set, the base version of Astronomy is applied and the results are noted. This is repeated for all data sets with the adaptations and changes included. The improvements in the number of sources that have ellipses fit successfully are noted, along with the improvement in the detection rate of the anomalous sources.

The next set of results focus on the active learning function that is incorporated in Astronomy. Both the amount of active learning required and the method of labelling is tested. Labelling is done in increments of 1%, retraining the scores after each

increment, up to 5% and this is compared to a single labelling stint of 5%. These results are compared to those from a random sample from each data set as well as to the outcome from applying no active learning at all.

The actual interesting and anomalous sources are then compared to a known galaxy merger catalogue [116]. Sources present in both are noted and the sources within the DECaLS data set that are similar in nature to those in the merger catalogue are also noted as they could be possible unknown merger events.

5.4 Impact Of Adjustments Made To Astronomy

The first test for the adaptations made to Astronomy is to test the number of failures in the feature extraction process. It was seen in chapter 4 that there are numerous causes for failures regarding the ellipse fitting procedure. Some of these are dealt with before obtaining the data; using the flags within the catalogues to reduce masked source, or to limit the lower flux levels. These changes have been implemented before creating the subsets described earlier in this chapter. This allows the remaining challenges to be tested and dealt with directly.

The adaptations are tested as a whole since it was seen in section 4.6 that the improvements from the changes stack successfully. Figure 5.1 illustrates the improvements made in the number of sources that have their features extracted successfully for the random and the lens subsets.

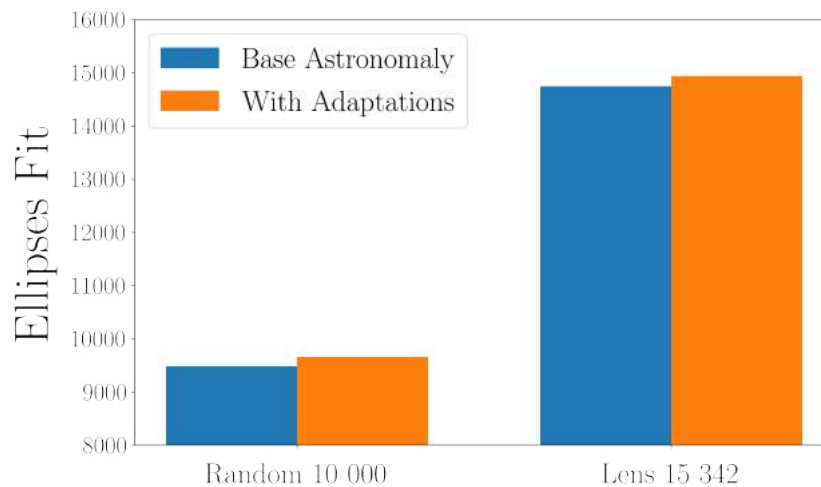


FIGURE 5.1: This histogram details the number of sources that have features extracted successfully for the random and the lens subsets. It is clear that both subsets show an increase in the number of sources that pass the feature extraction process.

The number of additional sources varies up to roughly 10%. This may not seem that significant at first, but on a much larger data set this could be thousands of sources added.

The next aspect of the adaptations tested is the location in the list of sources ranked by anomaly score. As mentioned earlier, Astronomy is designed to identify the more anomalous sources more easily. The results should indicate that there are a higher number of anomalous sources located higher up in the returned order of sources if there is an improvement.

Figure 5.2 contains the plots that show the improvements gained from the adaptations made. Both the random subset and the lens subset show an improvement from the start to the end of the plot. Both plots reflect the improvement in the number of sources detected when using the adapted version. This is clear from the higher values reached within each plot.

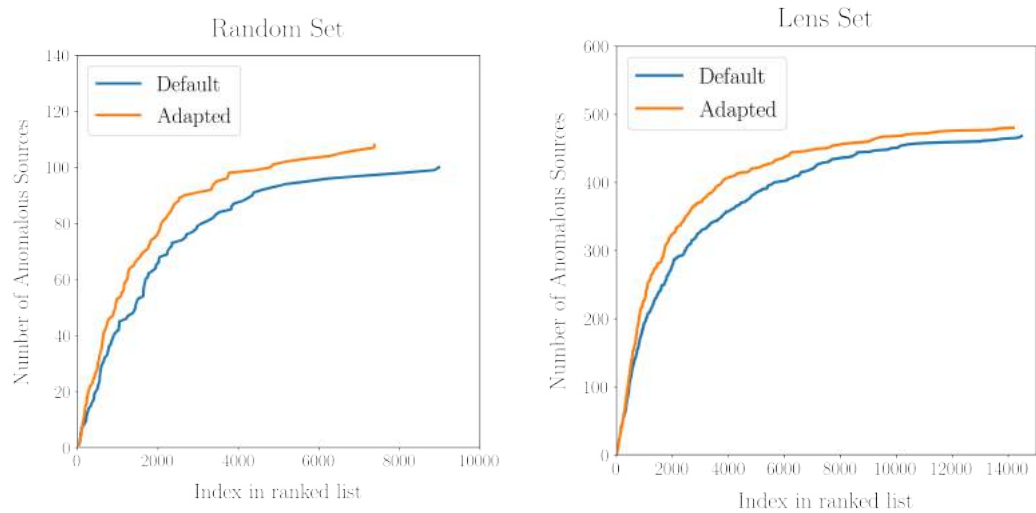


FIGURE 5.2: These plots show the comparison between the base version of Astronomy and the version with the adaptations made. Overall it is clear that there is a substantial improvement in the number of sources detected. There is also a clear shift in the locations of these sources within the returned order.

These results can be somewhat misleading though as the largest difference between the base and adapted versions tend to lie towards the lower end of the returned order. While there is a clear improvement in this region, this somewhat defeats the goal of applying Astronomy; if all of the sources are looked at, then the order that they are returned in has a much smaller significance. As such, the actual increase in performance of Astronomy lies towards the start of the plots, where the top sources are returned.

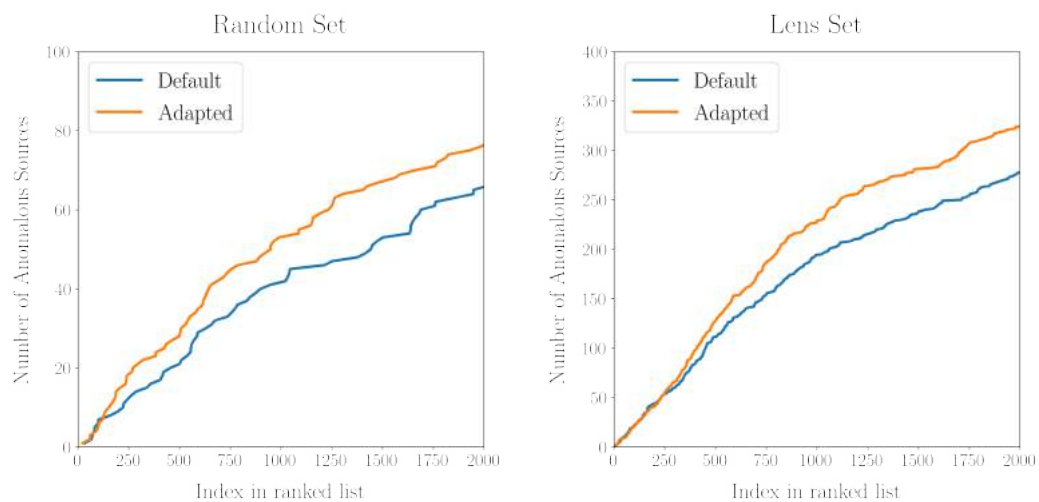


FIGURE 5.3: These plots are the same as those in Figure 5.2, but are limited to the top 2000 sources. Again a clear improvement is seen in the number of interesting sources located higher up in the list.

Figure 5.3 contains the same plots as those in Figure 5.2, except that the plots are limited to the first 2000 sources returned. This represents a reasonable number of sources that will be looked at when inspecting a data set. In both plots it is clearly seen that the adaptations made to Astronomy provide a significant improvement in the order of the interesting sources.

5.5 Improved Performance From Active Learning

Active learning is a key aspect of Astronomy, where the user is able to manually label sources based on how interesting they are. Not only does this reduce false positives, but it also allows specific sources to be weighted higher according to the needs of the user. This section is based on the improvements gained from applying active learning.

As seen in the previous section, it is the top end of the returned order of sources that matters most in determining the performance of Astronomy. As such, the outcome of the subsets will mostly be restricted to the top sources only.

Active learning is based on the manual labelling of sources by a user. In this case, the sources that appear interesting in any way are labelled higher and the false positives and uninteresting sources are labelled low. See Figure 3.3 and section 3.3.1 for an overview of the labelling process.

How many sources to label and how they should be labelled are important to know. If the data set is small, labelling a substantial amount of sources is relatively quick and easy and will provide good results, but for a large data set, even 5% of the data could be several thousand sources to label. This leads to Figures 5.4 and 5.5, where the active learning is tested in increments and this is compared to a single labelling set.

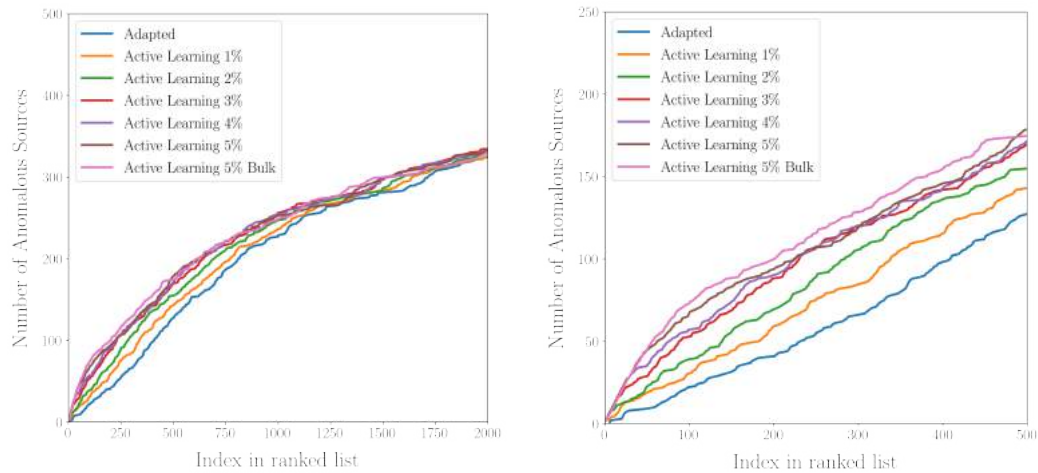


FIGURE 5.4: The plots shown here illustrate the results from applying active learning in 1% increments, up to 5% on the Lens subset. This incremental 5% is compared to a single, bulk 5% training set done in order to determine if there is any difference between the two methods.

Figure 5.4 shows that the various percentages of active learning applied appears to converge when looking at more sources. They still differ at higher amounts, but to a lesser degree. At the more important lower end, it is seen that each percentage increment labelled and trained provides an increase in the results compared to the previous percentage. This is somewhat given, since more sources labelled would provide a better result. The significance of this however, is that labelling 2% provides a noticeable gain over labelling 1%, and the same for the higher percentages, although the difference tends to become smaller and smaller. It is also clear that the bulk training method provides a better performance than the 5% incremental method. This is due to the larger variety of sources labelled when doing bulk labelling; incremental training returns similar sources after each training and so they are more reinforced, but a smaller variety of sources are labelled overall.

However, 5% of a data set can be a large amount of sources to label. There is a clear difference between labelling 1% and labelling 2%, but is it worth labelling more if the time required is taken into consideration? Figure 5.5 indicates that the RWS values flatten out quickly after labelling 3% in increments. The method of labelling used is the reason for this. After labelling a certain number of sources, the algorithm is retrained and returns sources with higher scores. When this process is repeated a few times, sources that are similar in nature tend to be found, and labelling these won't have as much of an impact on the rest of the data set. While better scores are obtained when labelling more sources, the extra time required might not be worth the small gain in performance. The amount to label is thus dependent upon the data set in question;

for large data sets it might not be possible to label large amounts and 1-2% might be sufficient, but for smaller data sets it is best to label as much as possible within reason.

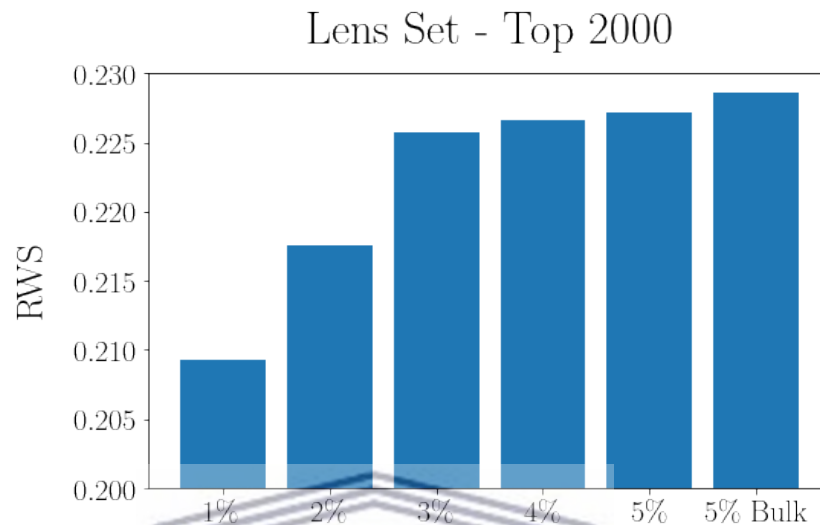


FIGURE 5.5: This histogram shows the RWS values for the top 2000 sources from the Lens set when applying incremental labelling for active learning. Also included in the histogram is the 5% bulk labelling RWS value, which returns the best results.

It was seen in Figure 5.3 that the adapted version of Astronomy provides improved results. The 5% bulk active learning is thus compared to the adapted version to see what performance gains are achieved. Figure 5.6 displays the results for the random and lens subsets when using the iForest algorithm. In these plots, a random selection was also made from each set and the number of sources that it contained was also plotted. This is to give an idea of the performance gain in detecting anomalies when applying machine learning on a data set. It is evident that machine learning provides a significant improvement over a random sample and that active learning provides an even better result. From both plots it is seen that more anomalous sources are detected within the top 2000 sources of each data set. More importantly though, these anomalous sources are also ranked higher in the top 2000 as can be seen by the shift to the left of the plot when compared to machine learning only.

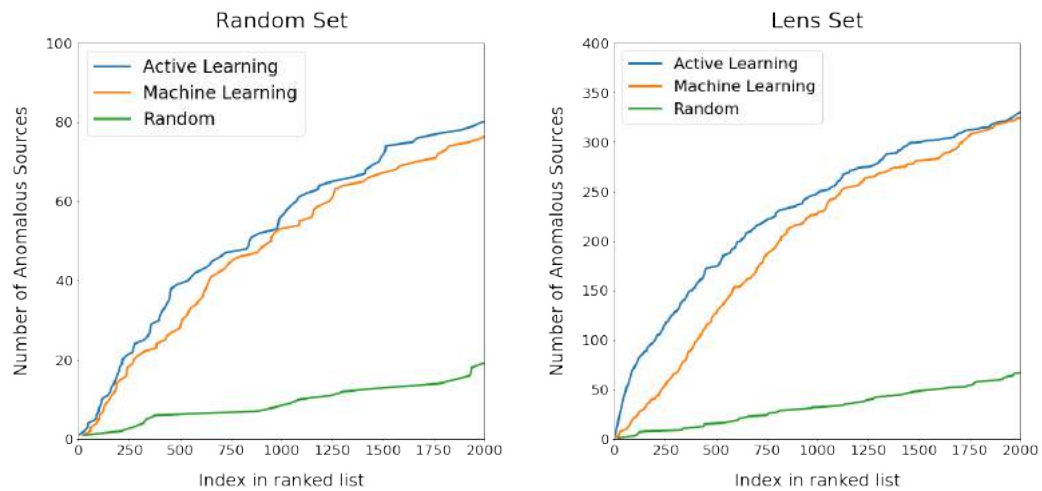


FIGURE 5.6: These plots show the performance of active learning compared to the adapted version of Astronomy along with a random selection of sources. The random selection is included to further highlight the gains from applying machine learning (iForest) on the data sets.

Figure 5.7 shows the RWS scores obtained from the different methods for the random and lens subsets. It should be noted that these are all based on the entire data sets and not just on the top 2000 sources. This is the reason why there are certain discrepancies in these results, where the 5% bulk labelling does not necessarily perform the best in all instances. This result should be taken with caution, as it is highly unlikely that the entire data set will be looked at, and if it is then the application of machine learning is not required. The key result from this figure is that any form of machine learning, with or without any application of active learning, significantly outperforms a random selection of the data.

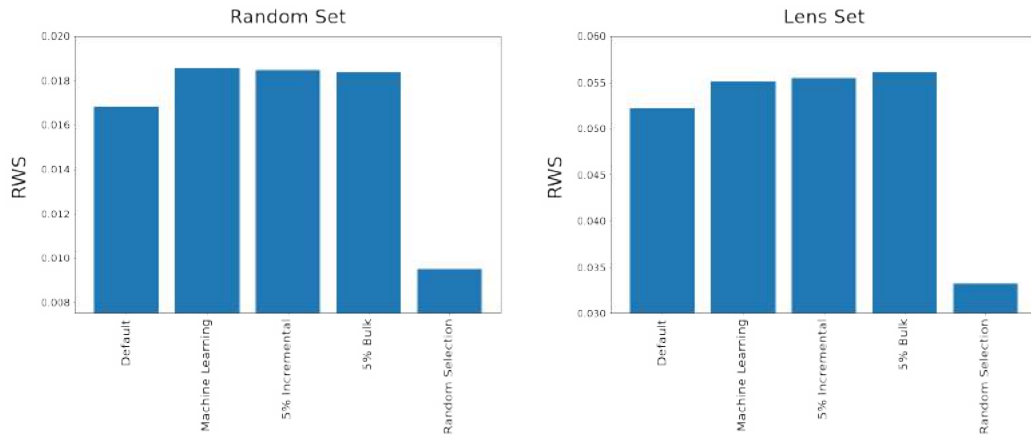


FIGURE 5.7: These histograms illustrate the RWS values obtained from the various methods applied to the data sets. It is clear that any one of the machine learning applications provides an improvement over a random sample of the data.

5.6 Interesting Sources Identified

Interpreting the actual results found is more of a challenge. The scale and depths of the Legacy Surveys are difficult to compare to with other surveys and finding catalogues of anomalous sources is extremely difficult due to the scarcity of such events. The Catalog of Merging Galaxies, [116], is used as a comparison for some of the results. There are several challenges when it comes to making cross comparisons; faint sources can be hard to identify and match accurately, the coordinates make it difficult to match the exact source in some instances, the difference between the quality of the observations made also make it difficult to match sources. Another significant challenge is finding catalogues that cover the same region of the sky. This is also why the Southern Hemisphere was chosen to be investigated, since the chances of finding something that no one has seen before is greater. Nonetheless, the Catalog of Merging Galaxies contains regions that overlap with that of DECaLS and the quality of the observations are good enough to make some accurate matches.

Table 5.1 contains all of the sources located within the three subsets used above that are also identified within the merging catalogue.

TABLE 5.1: This table represents the images located within the galaxy merger catalogue that have also been located within the various subsets of DECaLS explored. The merging type is as follows: M=mergers, CP=close pairs, CM=close multiples.

Subset Located In	Identification HC2009	RA J2000	DEC J2000	Field Name	Merging Type
Random	1942	01h26m19.5s	-02h24m07.8s	RCS0133	M
Random	3489	01h44m16.9s	-02h24m06.9s	RCS0133	M
Random	6801	11h03m14.2s	-05h09m57.4s	RCS1111	M
Random	6868	11h04m17.6s	-03h39m27.4s	RCS1111	M
Random	12714	21h57m37.1s	+02h08m40.3s	RCS2143	M
Random	13657	23h27m25.6s	-12h12m13.2s	RCS2338	M
Lens	2744	01h37m33.3s	-00h28m32.9s	RCS0133	M



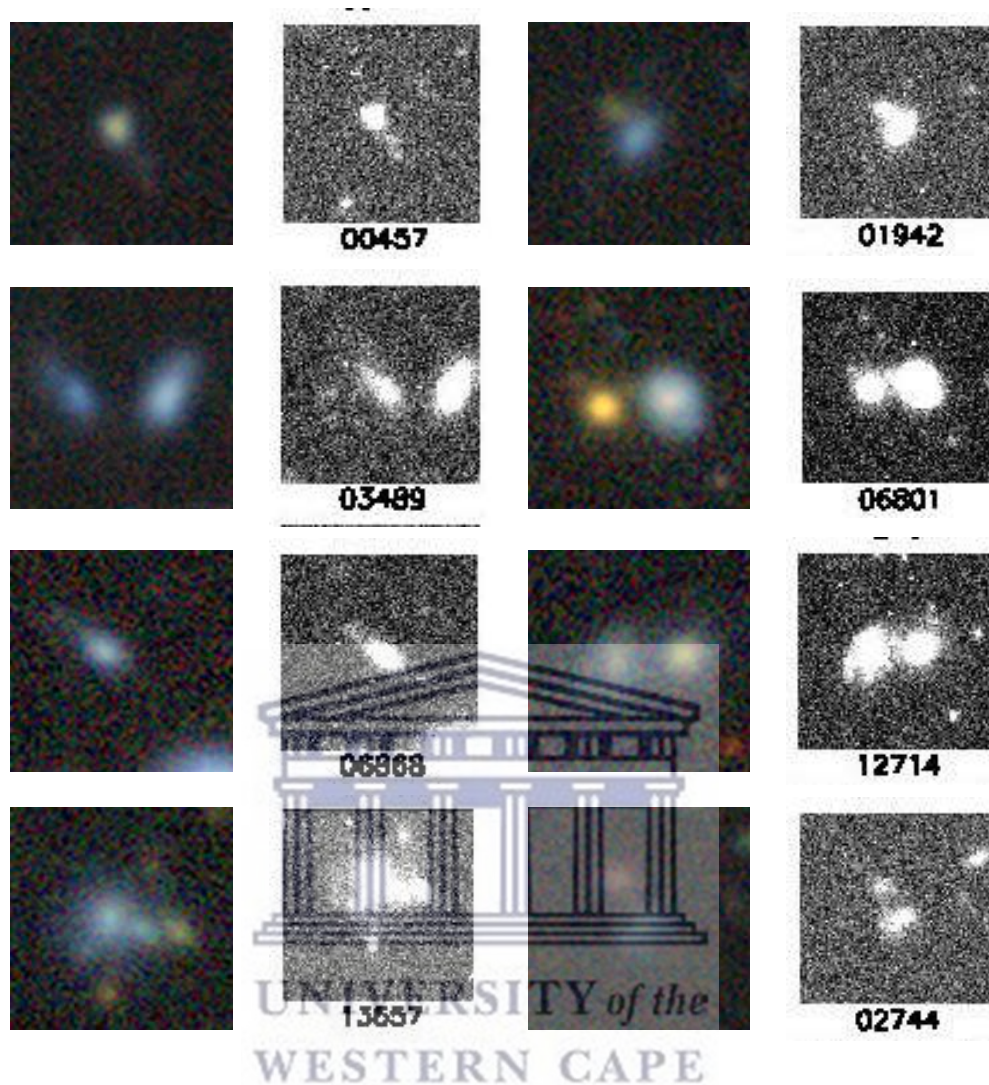


FIGURE 5.8: These images are the ones contained within both the DECaLS subsets used in the thesis and in the Catalog of Galaxy Mergers. The first and third columns are the sources obtained by DECaLS [58], the second and fourth columns are the ones from the Catalog of Galaxy Mergers [116]. It is evident that making accurate cross matches is challenging and is largely dependent upon the quality of the observations.

Figure 5.8 contains the images of these sources from both the DECaLS data set and the Catalog of Merging Galaxies. It is clear that the majority of these sources are faint and difficult to detect, yet they have all been identified as merger events.

Table 5.2 contains similar images that have been identified within the subsets created earlier. These sources are similar in nature, but are not identified within the merger catalogue. This is due to the sky coverage such that these sources are not within the same field as that of the merging catalogue. The Catalog of Galaxy Mergers covers an area of 422 deg^2 , compared to the $19\,437 \text{ deg}^2$ of the DECaLS data set. It stands to reason that randomly selecting sources within the DECaLS data set has a small chance

of being in the same region as the Catalog of Galaxy Mergers despite the two areas overlapping.

TABLE 5.2: This table contains the information of some of the sources that are similar in nature to those found within the galaxy merger catalogue. The corresponding images of these sources are located in Figure 5.9. It should be noted that there are no sources from the brightest set, mainly due to the fact that the galaxy merger catalogue is focused on faint sources. The order is the same as that in which the sources were ranked as being anomalous.

Subset	ObjID	RA (degrees)	DEC (degrees)	Peak Flux (nanomaggies)
Lens Set	11445	239.7461	30.73416	20.89395
Lens Set	13821	337.07996	-1.15186	69.58491
Lens Set	2962	27.63204	28.60366	25.38894
Lens Set	1947	13.2149	-25.10791	32.4932
Lens Set	3764	36.83676	-1.12947	26.64307
Lens Set	6906	146.0695	6.26327	46.95848
Lens Set	5735	112.03286	26.99792	244.46881
Lens Set	7179	147.83899	30.66448	31.50535
Lens Set	1545	10.64194	20.60109	55.33132
Lens Set	4431	52.42099	-0.7785	809.40283
Random Set	9051	334.3656	-9.1567135	20.320427
Random Set	6101	183.90918	10.311913	20.440084
Random Set	9305	345.55133	-12.7322445	19.318884
Random Set	2555	54.56285	-52.45696	31.580101
Random Set	8674	324.71393	9.353114	45.52135
Random Set	7800	255.18332	26.90739	10.209927
Random Set	99	2.6280801	-2.1222188	30.379688
Random Set	8402	318.0885	15.651253	26.046024
Random Set	9359	346.98325	11.854805	36.805374
Random Set	2799	63.09433	-1.3849617	31.563364

Figure 5.9 contains the images of the sources located within Table 5.2. All of the sources contains multiple individual sources that are either along the line of sight to each other or they are interacting. There are several similarities between these sources and those found in Figure 5.8, suggesting that these are possibly cases of galaxy mergers or galaxies that are interacting with other galaxies.

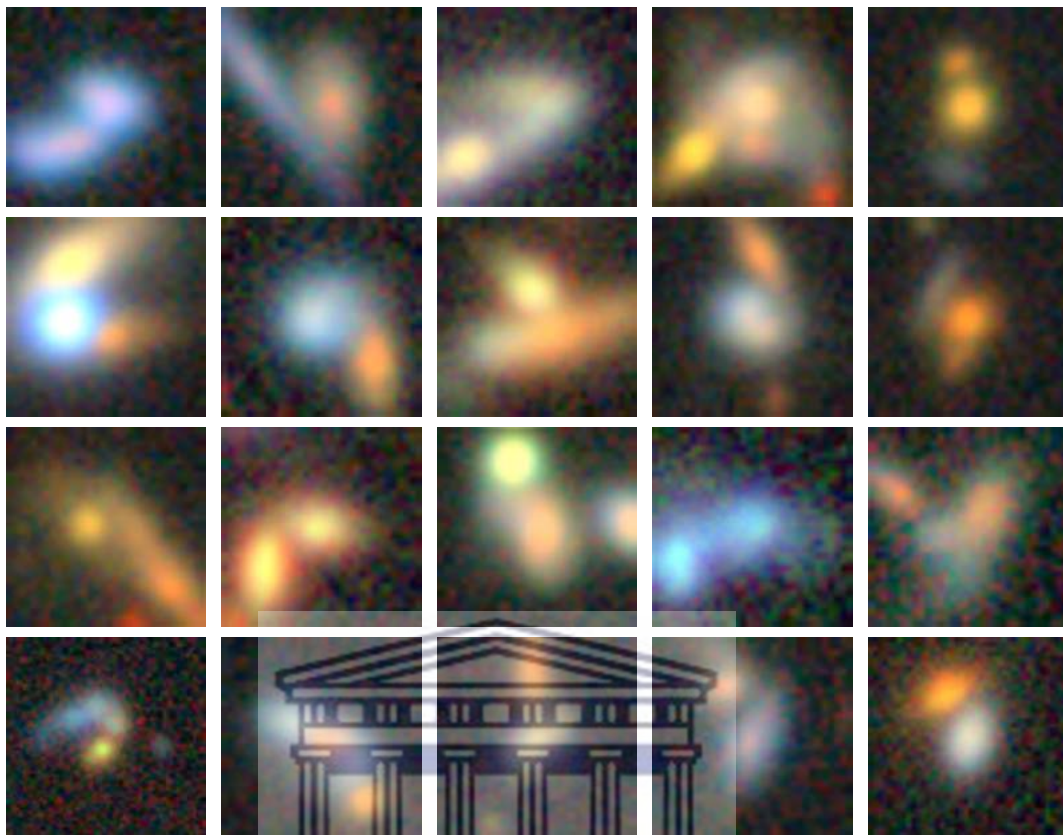


FIGURE 5.9: These images are the sources indicated in Table 5.2. The first ten sources are all located within the subset containing the gravitational lenses, while the last ten are all located within the random subset. They are all visually similar in nature to those found in the galaxy merger catalogue, except that they contain colour due to the band weightings applied. There is a possibility that some, or all of these sources are coincidental alignments, merely being along the line of sight to each other.

It is worth noting that none of these sources are located within the Bright subset. The reason behind this is that the sources located within the Bright subset are typically close by and have large angular diameters. This causes the issue that the entire source is not displayed within the cutout images used, even when basing the initial size on the suggested model as found in the DECaLS catalogues and applying the open ellipse function. Obtaining the correct cutout size for such large sources is thus a challenge that has not been successfully handled at this stage. The issue lies in the fact that these types of sources are not identified unless a follow up investigation is done. Since there are so many sources in the data set that has this problem, doing a follow up of each source is not an efficient way of dealing with it. Figure 5.10 shows some of the more interesting sources that suffer from this issue and which are not correctly identified within Astronomy due to the incorrect cutout size obtained. This brings into question the results pertaining to the Bright data set as a whole since not all sources are utilised equally.

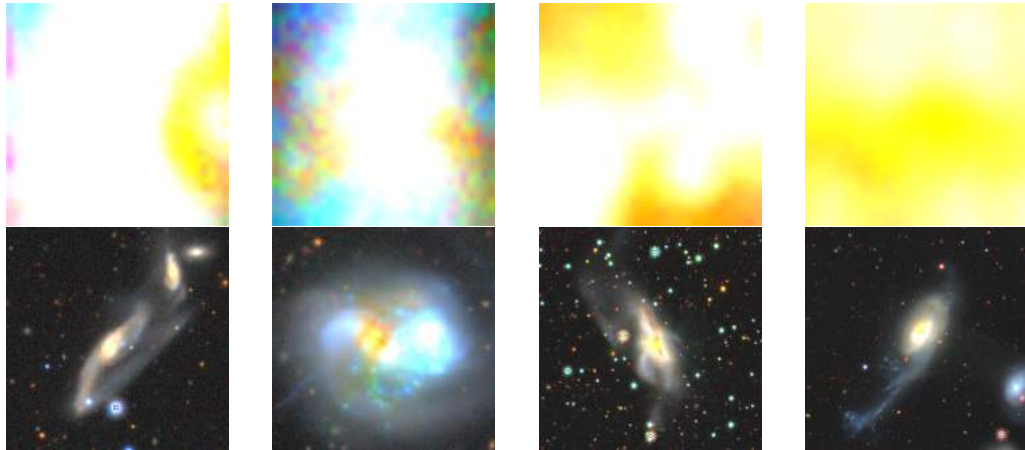


FIGURE 5.10: These images are some of the sources located within the Bright subset. The model fit within the DECaLS catalogue does not provide the correct image size and the open ellipse function implemented within Astronomy is not able to deal with such a large image difference. The top row shows the incorrect image size obtained and used within Astronomy. The bottom row shows the sources within the correct image size after a follow up investigation has been done.



Chapter 6

Discussion

Scientific discoveries are often made when observing new and rare sources, yet they are becoming more difficult to identify within the rapidly expanding volumes of data. Astronomical data sets have already reached the point where the sheer volume of data is beyond the typical means of inspection. Even crowd-sourcing projects such as Galaxy Zoo, which utilises hundreds of thousands of citizen volunteers, is likely not efficient enough to handle the large volumes of data available [7, 117]. Upcoming astronomical surveys will produce significantly more data to explore. Machine learning is becoming a more and more popular solution to tackling such large data volumes.

The goal of this study is to apply Astronomy, which incorporates machine learning techniques, on optical astronomical data to assist in detecting anomalous objects located therein. Astronomy is a framework incorporating numerous different techniques for each stage of the process; from pre-processing to feature extraction and the machine learning algorithm applied. As such, there are several feature extraction options that can be utilised. The ellipse fitting procedure is used throughout the thesis, which works by fitting ellipses to isophotes (contours of a given flux level) and using the parameters of the ellipses as features. Various machine learning algorithms are also available for use, from which the iForest algorithm is selected.

The results indicate that machine learning is not only capable of handling large volumes of data, but that outlier detection algorithms significantly improve the detection rates of anomalous sources. Additionally, active learning provides a further improvement to the detection rates and can be used selectively for specific targets if needed. It is seen that Astronomy provides a complete solution, from data pre-processing to a user friendly output display, that makes identifying anomalous sources much easier.

The optical data from the DECaLS data set utilised in the thesis contains three different bands, each corresponding to a different observational band used. Combining the bands together during the pre-processing stage allows the information from all three bands to be utilised, a feature that was not taken into account in the base version of Astronomy. The method used to join the bands together also plays an important role in both the pre-processing and the output display section of Astronomy. Some selection criteria for the DECaLS data was also required to ensure optimal performance; these include selecting sources with a minimum flux level, only using sources that have all three bands available and removing the point sources and the masked sources as options to be selected. Several tests were done that concluded that the sources should be acquired using the individual cutouts; the fits files containing single sources only instead of the DECaLS bricks. This allows differing image sizes to be obtained as well as increasing the randomness of the selections made.

6.1 Changes Made To Astronomy

During the study, several changes and additions had to be made to Astronomy to improve the algorithms performance on the DECaLS data. The changes range from the pre-processing stage to the output display. Most of them are aimed at improving the number of sources that have their features extracted successfully. This is because there was a substantial amount of feature extraction failures that were caused by various factors; from incorrect image sizes to a lack of data points to fit ellipses. Image pre-processing thus plays a vital role in the entire process as it directly affects the features that are extracted, which in turn affects the machine learning algorithms.

For example, *adaptive scaling* was introduced to increase the number of data points available within a given image. This was required to reduce feature extraction failures where the innermost ellipse could not be fit due to a lack of enough data points. In some instances, incorrect features were extracted where the source was too large for the image itself. For these occurrences, a method was implemented that flagged them, allowing a better fitting image to be acquired, in turn improving the accuracy of the extracted features. The other significant adaptation made was for the output display, where the individual channels of the images are weighted using specific weights to improve their visual appearance. This results in more efficient and accurate manual labelling as the sources are easier to identify. Familiarity with the data used is therefore important; knowing what effect the various channels may have, or how they are displayed, may significantly affect the results. It is thus important that each step of the process be studied to determine how the data is impacted and how it affects the

rest of the process.

These adaptations and changes, when applied to Astronomy, shows an increase of up to 10% in the number of sources that have features extracted successfully. This is a significant result, especially when applied to large volumes of data where there could be millions of sources. For instance, it is estimated in [88] that there exists one gravitational lens for every ten thousand galaxies. This makes the 10% increase in sources with features extracted a very important result; for a data set of 1 million sources for example, this would *on average*, result in up to 10 additional gravitational lenses. In addition these changes also resulted in up to 20% more anomalous sources being located within the top 2000 sources. All of these small changes and additions work together and improve the detection rate of the anomalous sources by a substantial amount. From the results it is also seen that there are additional anomalous sources returned when using the adaptations; some of the feature extraction failures that occurred initially included anomalous sources.

6.2 Active Learning

For the subsets of the DECaLS data used, the anomalous sources have been identified beforehand by visual inspection in order to be able to test the performance of Astronomy. As such, the machine learning algorithms do not detect *new* sources since they are identified already, but rather returned them in a more favoured order. The output of Astronomy is an ordered list of the data, where the sources are ranked from most anomalous to least anomalous based on the scores given to them by the machine learning algorithm. The first, or top, 2000 sources in this list are looked at to evaluate the performance of the algorithm. This is compared to 2000 sources selected randomly within the subsets. Throughout the three subsets used, applying Astronomy provided a significant increase, up to 600%, in the number of anomalous sources located within the top 2000 sources when compared to the number of sources in the random sample. An even greater improvement is seen when applying active learning, especially towards the very top of the ranked list, with almost 40% more anomalous sources located within the top 500 sources.

Additionally, it was found that active learning performs better when the sources are labelled in bulk sets rather than incremental increases, although this can be an issue when using very large data sets. For large volumes of data, inspecting each and every source individually is time consuming and more often than not, is not possible. Having a larger number of anomalous sources concentrated within the top subset of

the data set makes them easier to detect and identify. This is the fundamental goal of Astronomy; it is not a classifier used to identify specific sources, but rather it is meant to identify anomalous sources throughout the relevant data set and to return the more anomalous sources higher up in the output.

Coupled with the previous result, the active learning aspect that forms a part of Astronomy allows user specific targets to be weighted higher than other sources, allowing the focus to be put on the targets of interest. The outcome in this study indicates that active learning provides a further improvement in the detection rate of the anomalous sources, even with a minimal number of sources labelled for retraining. It was shown that even a small amount of active learning, as little as 1% of the overall data set size, shows a visible improvement in performance. Active learning has been shown to improve results to varying degrees, but not once has it degraded the performance of the algorithm.

The gains achieved from active learning outweigh the additional time required labelling the sources and provide the additional benefit of reducing false positives that are caused by artefacts and masked sources. Similar improvements in identifying anomalies and outliers with the use of active learning have been confirmed within other studies done [118, 119].

Variations in the way that the labelling was done indicates that a single, larger number of sources labelled before retraining provides a better result compared to labelling and retraining multiple times on smaller amounts. However, it is seen that the improvements (of the rank weighted score and the anomaly recall) gained from labelling 3% compared to 4% or higher is not as significant as the gain from 2% to 3%, or from 1% to 2%. While labelling more sources provides a better result, it might not be as effective due to the extra time required to label. A trade-off exists between better results and manually labelling more sources. It must be taken into consideration as more anomalous sources can be identified by extending the number of sources looked at. A choice must be taken as to whether the time and effort will be spent on labelling more sources, or investigating a larger number of sources in the output in order to identify more sources. There are many factors that play a role in this decision, but the two main ones are the overall volume of data and the time it takes to label the sources manually and so it would depend on the individual case in question.

6.3 Computational Performance Of Algorithms Used

Some basic tests were done on the scalability of the actual machine learning algorithms, iForest and LOF, in order to determine if there were any limitations on the number of sources and dimensions that they could handle. The main focus was to determine the run-time and memory usage of the algorithms, although other aspects were also measured such as precision, accuracy and recall. The results indicate that the iForest algorithm scales linearly for both the run-time and memory usage when increasing the number of sources while keeping the dimensions constant, as well as when increasing the dimensions and keeping the number of sources constant. Similarly, the LOF algorithm also showed a linear relationship between run-time and number of sources and between memory usage and the number of sources. For both algorithms it is seen that the computational processing power available is the limiting factor when increasing the number of sources investigated.

Additional tests were done to determine the variability in the results of the application of the iForest algorithm. Multiple runs were done using the same set up on the same data set, yet the results varied for each run. This is not unexpected since the algorithm selects a random value when making the feature cuts which will produce slightly different results each time. Comparing the variations to the original Astronomy paper reveal that some performed marginally better, while others performed worse but they were all fairly consistent throughout [108].

6.4 Performance On The DECaLS Data

A comparison was carried out with a known galaxy merger catalogue regarding the anomalous sources detected in the output of the DECaLS subsets. A few of the sources within the subsets investigated overlapped with the merger catalogue and are known. Several other sources were identified that are similar in nature to the previously identified mergers, but they are not within the merger catalogue. The results found within this study from applying Astronomy fall in line with similar studies done regarding machine learning applications to detect anomalies within astronomical data [120–122]. While Astronomy will not present all of the anomalies within the data set within the topmost output, it is clearly capable of reducing the amount of time and effort required to detect a majority of the anomalous sources within a given data set. This is a strong indication that machine learning is a suitable solution to the issue of tackling big volumes of data.

It should be noted that the results for the Brightest subset (the subset of DECaLS selected based on highest flux values) used are not fully accurate. Detection rates failed in some instances and the main cause of this was the lack of proper image sizes for the sources in question. The image sizes obtained were incorrect and resulted in an incorrect detection rate along with a high number of incorrectly identified sources. This occurred despite using the image sizes recommended from the models available within the DECaLS catalogues, and despite applying the open ellipse function to reduce these types of errors. As a result of this, it is possible that some anomalous sources within this particular data set have not been correctly identified. The results based on those that were identified however, reflect the general results found from all other testing, but to a smaller scale. There is still an improvement in detection rates when applying Astronomy and when utilising the active learning capabilities.

Determining whether or not *new* types of anomalous sources can be detected using such techniques have not been confirmed. The results suggest that new sources would *potentially* be detected, but due to the limited size of the data sets explored, this has not been conclusively confirmed. The subset size also limits the sources expected therein. For instance, gravitational lenses are estimated to occur once in every 10 000 massive galaxies only [88], suggesting a high possibility of there not being such an event in the relatively small subsets used. For merger events, a comparison was carried out to the Catalog of Galaxy Mergers [116], showing that there are known, faint galaxy merger events within the very limited subsets of the data used. Similar sources have been located within the data sets as well, but a full follow up investigation has not been done to identify these sources so it is not known whether or not they are also merger events. Only by applying the techniques on larger volumes of data and doing full followup investigations will it be possible to determine if the method can detect *new* source types.

A limitation regarding the data sets used is based on the minimum flux level restriction applied. The restriction was placed since it became difficult to identify the faintest of sources, but this also goes against the depth achieved by the DECaLS data. A large amount of the faintest sources are newly observed, or have not been observed to such an extent as has been done in the Legacy Surveys, but by placing the lower flux limits, a significant amount of these sources are ruled out when selecting the data to be investigated. In turn this restricts the possibility of identify new anomalies. Future studies would require alternative feature extraction methods, or would require improved observations on these faint sources in order to identify them. For the Legacy Surveys, the latter has been implemented on most of the faint sources as they are marked for follow up investigations.

Astronomy has proven to be a robust and versatile package that returns significant results on the optical data. While it is not meant to work as a classifier, applying active learning on selective sources can increase the detection rates of these specific sources only. Additionally, active learning improves the chances of detecting sources similar to ones that have already been identified. Astronomy is also not limited to optical image data, but can be used on other types of data as well, even data obtained from other fields. The results obtained in this study reinforce the idea that machine learning provides a suitable solution to handling large volumes of data and is applicable to nearly every instance where such large volumes are used.

6.5 Future Work

The features used, obtained from the ellipse fitting process, are largely dependent upon the image itself and not just on the source in question. The majority of the images used throughout the thesis are small and fitting contours accurately can be difficult for these image sizes. There are also anomalies such as rings galaxies for which the features used here are insensitive to which would thus be difficult to detect using this method. Alternative features could be incorporated, either in addition to, or as a complete replacement for, the current features used. It can be seen that representation learning can be very effective and could be a suitable replacement, although it has only been tested on a very specific subset of DECaLS data set [89]. Representation learning on larger and more diverse subsets will be investigated in future work, and other feature extractors for the DECaLS data will also be explored.

The faintest of sources within the DECaLS data set could not be used in the study due to the difficulty in identifying them visually. This places a limitation on the source magnitude that can be investigated, in turn restricting the chance of identifying new anomalies. However, known galaxy mergers were identified amongst the fainter sources that were included and several possible merger candidates similar to these were detected as well. This indicates that more focus should be given to the fainter sources in future investigations.

To determine whether new types of anomalies could be detected using this method it is recommended that future studies be done on much larger volumes of data, thereby increasing the chances of actually identifying anomalies. Large scale data sets would also provide a true test for the active learning aspect, especially regarding how much labelling is enough to make a significant impact.

Further research can also be done on multi-wavelength data since the anomaly detection techniques to date have mostly been limited to single data types. Sources that appear to be ordinary within the optical spectrum might be anomalous within the radio spectrum for instance. Correlations between properties in different wavelengths could also possibly be used to identify anomalies. Multi-wavelength studies into anomalies can provide greater insight into their properties and warrants future investigations.

Machine learning has proven to be a successful method for handling large volumes of data and has great potential for anomaly detection in current and future data sets. Detecting anomalies not only provides the opportunity to expand the knowledge about known, yet rare phenomena, it also provides the possibility of detecting previously unidentified phenomena. The work done here provides a solid foundation in anomaly detection on astronomical data with substantial results on real data and includes the identification of key criteria and shortcomings of data processing and feature extraction. Astronomaly is a flexible framework and can be applied to data of different wavelengths, making it an ideal tool to utilise for upcoming data from the Square Kilometer Array and from the Vera C. Rubin Observatory [5, 59].



Bibliography

- [1] Russ Taylor, Fabio Porto, Cui Chenzou, Yogesh. Wadadekar, and Oleg Malkov. Big data research infrastructure collaboration toward the ska (bricska). May 2021. ISSN 1678-2690. doi: <https://doi.org/10.1590/0001-3765202120201027>. URL <https://doi.org/10.1590/0001-3765202120201027>.
- [2] M. R. Blanton, M. A. Bershad, B. Abolfathi, F. D. Albareti, C. Allende Prieto, and et al. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. , 154:28, July 2017. doi: 10.3847/1538-3881/aa7567. URL <http://adsabs.harvard.edu/abs/2017AJ....154...28B>.
- [3] Dalya Baron and Dovi Poznanski. The weirdest SDSS galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465(4):4530–4555, 11 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw3021. URL <https://doi.org/10.1093/mnras/stw3021>.
- [4] A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins. Observation of a Rapidly Pulsating Radio Source. , 217(5130):709–713, February 1968. doi: 10.1038/217709a0. URL <https://ui.adsabs.harvard.edu/abs/1968Natur.217..709H>.
- [5] P.N. Wilkinson, K.I. Kellermann, R.D. Ekers, J.M. Cordes, and T. Joseph W. Lazio. The exploration of the unknown. *New Astronomy Reviews*, 48 (11):1551–1563, 2004. ISSN 1387-6473. doi: <https://doi.org/10.1016/j.newar.2004.09.036>. URL <https://www.sciencedirect.com/science/article/pii/S1387647304001344>. Science with the Square Kilometre Array.
- [6] Kenneth I Kellermann, James M. Cordes, Ronald D Ekers, Joseph Lazio, and P Wilkinson. The Exploration of the Unknown. *PoS*, sps5:005, 2010. doi: 10.22323/1.099.0005.
- [7] M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, and et al. Galaxy zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, 9(1), Dec 2010. ISSN 1539-1515. doi: 10.3847/aer2009036. URL <http://dx.doi.org/10.3847/AER2009036>.

- [8] Charles J. Hailey, Kaya Mori, Franz E. Bauer, Michael E. Berkowitz, Jaesub Hong, and et al. A density cusp of quiescent x-ray binaries in the central parsec of the galaxy. *Nature*, 556(7699):70–73, 2018. URL https://EconPapers.repec.org/RePEc:nat:nature:v:556:y:2018:i:7699:d:10.1038_nature25029.
- [9] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. doi: 10.1007/bf02478259.
- [10] Dalya Baron. Machine learning in astronomy: a practical overview, 2019.
- [11] Ronald J. Buta. Galaxy morphology, 2011.
- [12] Edwin. Hubble. No. 324. Extra-galactic nebulae. *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington*, 324:1–49, January 1926. URL <https://ui.adsabs.harvard.edu/abs/1926CMWCI.324...1H>.
- [13] Curtis Struck. Galaxy collisions. *Physics Reports*, 321(1-3):1–137, Nov 1999. ISSN 0370-1573. doi: 10.1016/s0370-1573(99)00030-7. URL [http://dx.doi.org/10.1016/S0370-1573\(99\)00030-7](http://dx.doi.org/10.1016/S0370-1573(99)00030-7).
- [14] M. Romano, P. Cassata, L. Morselli, G. C. Jones, M. Ginolfi, A. Zanella, M. Béthermin, P. Capak, A. Faisst, O. Le Fèvre, and et al. The alpine-alma [cii] survey. *Astronomy Astrophysics*, 653:A111, Sep 2021. ISSN 1432-0746. doi: 10.1051/0004-6361/202141306. URL <http://dx.doi.org/10.1051/0004-6361/202141306>.
- [15] Ihsan Hafez. *Abd al-Rahman al-Sufi and his book of the fixed stars: a journey of re-discovery*. PhD thesis, James Cook University, October 2010. URL <https://ui.adsabs.harvard.edu/abs/2010PhDT.....295H>.
- [16] E. P. Hubble. *Realm of the Nebulae*. 1936. URL <https://ui.adsabs.harvard.edu/abs/1936rene.book.....H>.
- [17] F. Bournaud, C. J. Jog, and F. Combes. Multiple minor mergers: formation of elliptical galaxies and constraints for the growth of spiral disks. *Astronomy Astrophysics*, 476(3):1179–1190, Oct 2007. ISSN 1432-0746. doi: 10.1051/0004-6361:20078010. URL <http://dx.doi.org/10.1051/0004-6361:20078010>.
- [18] Roberto G. Abraham and Sidney van den Bergh. The morphological evolution of galaxies. *Science*, 293(5533):1273–1278, Aug 2001. ISSN 1095-9203. doi: 10.1126/science.1060855. URL <http://dx.doi.org/10.1126/science.1060855>.

- [19] Gerard de Vaucouleurs. Classification and Morphology of External Galaxies. *Handbuch der Physik*, 53:275, January 1959. doi: 10.1007/978-3-642-45932-0_7. URL <https://ui.adsabs.harvard.edu/abs/1959HDP...53..275D>.
- [20] E. Holmberg. A photographic photometry of extragalactic nebulae. *Meddelanden fran Lunds Astronomiska Observatorium Serie II*, 136:1, January 1958. URL <https://ui.adsabs.harvard.edu/abs/1958MeLuS.136....1H>.
- [21] Rui Guo, Cai-Na Hao, Xiaoyang Xia, Yong Shi, Yanmei Chen, and et al. Toward an understanding of the massive red spiral galaxy formation. *The Astrophysical Journal*, 897(2):162, Jul 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab9b75. URL <http://dx.doi.org/10.3847/1538-4357/ab9b75>.
- [22] Gerard de Vaucouleurs. Recherches sur les Nebuleuses Extragalactiques. *Annales d'Astrophysique*, 11:247, January 1948. URL <https://ui.adsabs.harvard.edu/abs/1948AnAp...11..247D>.
- [23] N. Caon, M. Capaccioli, and R. Rampazzo. Photographic and CCD surface photometry of 33 early-type galaxies in the Virgo cluster. I. The data. , 86:429, December 1990. URL <https://ui.adsabs.harvard.edu/abs/1990A&AS...86..429C>.
- [24] N. Caon, M. Capaccioli, and M. D'Onofrio. On the shape of the light profiles of early-type galaxies. , 265:1013–1021, December 1993. doi: 10.1093/mnras/265.4.1013. URL <https://ui.adsabs.harvard.edu/abs/1993MNRAS.265.1013C>.
- [25] John Kormendy, David B. Fisher, Mark E. Cornell, and Ralf Bender. Structure and Formation of Elliptical and Spheroidal Galaxies. , 182(1):216–309, May 2009. doi: 10.1088/0067-0049/182/1/216. URL <https://ui.adsabs.harvard.edu/abs/2009ApJS...182..216K>.
- [26] Kevin Bundy, Richard S. Ellis, and Christopher J. Conselice. The mass assembly histories of galaxies of various morphologies in the GOODS fields. *The Astrophysical Journal*, 625(2):621–632, jun 2005. doi: 10.1086/429549. URL <https://doi.org/10.1086/429549>.
- [27] Alice E. Shapley. Physical properties of galaxies from $z = 2-4$. *Annual Review of Astronomy and Astrophysics*, 49(1):525–580, 2011. doi: 10.1146/annurev-astro-081710-102542. URL <https://doi.org/10.1146/annurev-astro-081710-102542>.
- [28] O. J. Eggen, D. Lynden-Bell, and A. R. Sandage. Evidence from the motions of old stars that the Galaxy collapsed. , 136:748, November 1962. doi: 10.1086/147433. URL <https://ui.adsabs.harvard.edu/abs/1962ApJ...136..748E>.

- [29] S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. , 183:341–358, May 1978. doi: 10.1093/mnras/183.3.341. URL <https://ui.adsabs.harvard.edu/abs/1978MNRAS.183..341W>.
- [30] P. A. R. Ade, N. Aghanim, M. I. R. Alves, C. Armitage-Caplan, M. Arnaud, and et al. Planck2013 results. i. overview of products and scientific results. *Astronomy Astrophysics*, 571:A1, Oct 2014. ISSN 1432-0746. doi: 10.1051/0004-6361/201321529. URL <http://dx.doi.org/10.1051/0004-6361/201321529>.
- [31] Michael S. Turner. λ cdm: Much more than we expected, but now less than what we want, 2021.
- [32] A. Einstein. Die grundlage der allgemeinen relativitätstheorie. *Annalen der Physik*, 354(7):769–822, 1916. doi: <https://doi.org/10.1002/andp.19163540702>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19163540702>.
- [33] V. Smith. Concepts of particle physics vol 1. *Physics Bulletin*, 37:303–304, 1986.
- [34] Matthias Steinmetz and Julio F. Navarro. The hierarchical origin of galaxy morphologies. , 7(4):155–160, June 2002. doi: 10.1016/S1384-1076(02)00102-1. URL <https://ui.adsabs.harvard.edu/abs/2002NewA....7..155S>.
- [35] Leandros Perivolaropoulos and Foteini Skara. Challenges for λ cdm: An update, 2021.
- [36] C. Conselice and J. Arnold. The structures of distant galaxies – ii. diverse galaxy structures and local environments at $z=4-6$; implications for early galaxy assembly. *Monthly Notices of the Royal Astronomical Society*, 397:208–231, 2009.
- [37] Jürg Diemand and Ben Moore. The structure and evolution of cold dark matter halos. *Advanced Science Letters*, 4(2):297–310, Feb 2011. ISSN 1936-7317. doi: 10.1166/asl.2011.1211. URL <http://dx.doi.org/10.1166/asl.2011.1211>.
- [38] A Cibinel, E Daddi, M T Sargent, E Le Floch, D Liu, and et al. Early- and late-stage mergers among main sequence and starburst galaxies at $0.2 < z < 2$. *Monthly Notices of the Royal Astronomical Society*, 485(4):5631–5651, 03 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz690. URL <https://doi.org/10.1093/mnras/stz690>.
- [39] Joseph A O’Leary, Benjamin P Moster, Thorsten Naab, and Rachel S Somerville. Emerge: Empirical predictions of galaxy merger rates since $z=6$. *Monthly Notices of the Royal Astronomical Society*, Dec 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa3746. URL <http://dx.doi.org/10.1093/mnras/staa3746>.

- [40] Vivienne Wild, C. Jakob Walcher, Peter H. Johansson, Laurence Tresse, and et al. Post-starburst galaxies: more than just an interesting curiosity. *Monthly Notices of the Royal Astronomical Society*, 395(1):144–159, May 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.14537.x. URL <http://dx.doi.org/10.1111/j.1365-2966.2009.14537.x>.
- [41] J. Christopher Mihos and Lars Hernquist. Ultraluminous Starbursts in Major Mergers. , 431:L9, August 1994. doi: 10.1086/187460. URL <https://ui.adsabs.harvard.edu/abs/1994ApJ...431L...9M>.
- [42] Francoise Combes. Les galaxies. *L’Astronomie*, 130:14–19, August 2016. URL <https://ui.adsabs.harvard.edu/abs/2016LAstr.130h..14C>.
- [43] R. López Fernández, R. M. González Delgado, E. Pérez, R. García-Benito, R. Cid Fernandes, and et al. Cosmic evolution of the spatially resolved star formation rate and stellar mass of the califa survey. *Astronomy Astrophysics*, 615:A27, Jul 2018. ISSN 1432-0746. doi: 10.1051/0004-6361/201732358. URL <http://dx.doi.org/10.1051/0004-6361/201732358>.
- [44] M. Romano, P. Cassata, L. Morselli, G. C. Jones, M. Ginolfi, and et al. The alpine-alma [cii] survey. *Astronomy Astrophysics*, 653:A111, Sep 2021. ISSN 1432-0746. doi: 10.1051/0004-6361/202141306. URL <http://dx.doi.org/10.1051/0004-6361/202141306>.
- [45] Johan H. Knapen and Philip A. James. The H α Galaxy Survey. VIII. Close Companions and Interactions, and the Definition of Starbursts. , 698(2):1437–1455, June 2009. doi: 10.1088/0004-637X/698/2/1437. URL <https://ui.adsabs.harvard.edu/abs/2009ApJ...698.1437K>.
- [46] Matthias Bartelmann. Gravitational lensing. *Classical and Quantum Gravity*, 27(23):233001, Nov 2010. ISSN 1361-6382. doi: 10.1088/0264-9381/27/23/233001. URL <http://dx.doi.org/10.1088/0264-9381/27/23/233001>.
- [47] LSST Science Collaboration, Paul A. Abell, Julius Allison, Scott F. Anderson, and John R. Andrew et al. Lsst science book, version 2.0, 2009. URL https://www.lsst.org/sites/default/files/docs/sciencebook/SB_Whole.pdf.
- [48] F. W. Dyson, A. S. Eddington, and C. Davidson. A Determination of the Deflection of Light by the Sun’s Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919, January 1920. URL <https://doi.org/10.1098/rsta.1920.0009>.

- [49] D. Walsh, R. F. Carswell, and R. J. Weymann. 0957+561 A, B: twin quasistellar objects or gravitational lens? , 279:381–384, May 1979. doi: 10.1038/279381a0. URL <https://ui.adsabs.harvard.edu/abs/1979Natur.279..381W>.
- [50] Joachim Wambsganss. Gravitational lensing in astronomy. *Living Reviews in Relativity*, 1(1), Nov 1998. ISSN 1433-8351. doi: 10.12942/lrr-1998-12. URL <http://dx.doi.org/10.12942/lrr-1998-12>.
- [51] Johanna Miller. Gravitational-lensing measurements push hubble-constant discrepancy past 5 . *Physics Today*, 73:14–16, 03 2020. doi: 10.1063/PT.3.4424.
- [52] David J. Bacon, Alexandre R. Refregier, and Richard S. Ellis. Detection of weak gravitational lensing by large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 318(2):625–640, 10 2000. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2000.03851.x. URL <https://doi.org/10.1046/j.1365-8711.2000.03851.x>.
- [53] I. A. Bond, A. Udalski, M. Jaroszyski, N. J. Rattenbury, B. Paczynski, and et al. OGLE 2003-BLG-235/MOA 2003-BLG-53: A planetary microlensing event. *The Astrophysical Journal*, 606(2):L155–L158, apr 2004. doi: 10.1086/420928. URL <https://doi.org/10.1086/420928>.
- [54] Richard Ellis, Michael R. Santos, Jean-Paul Kneib, and Konrad Kuijken. A Faint Star-forming System Viewed through the Lensing Cluster Abell 2218: First Light at $z \sim 5.6$? , 560(2):L119–L122, October 2001. doi: 10.1086/324423. URL <https://ui.adsabs.harvard.edu/abs/2001ApJ...560L.119E>.
- [55] Jan-Torge Schindler, Xiaohui Fan, Ian D. McGreer, Jinyi Yang, Feige Wang, and et al. The extremely luminous quasar survey in the sloan digital sky survey footprint. iii. the south galactic cap sample and the quasar luminosity function at cosmic noon. *The Astrophysical Journal*, 871(2):258, Feb 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/aaf86c. URL <http://dx.doi.org/10.3847/1538-4357/aaf86c>.
- [56] S. G. Djorgovski, R. R. Gal, S. C. Odewahn, R. R. de Carvalho, R. Brunner, and et al. The Palomar Digital Sky Survey (DPOSS). In Stephane Colombi, Yannick Mellier, and Brigitte Raban, editors, *Wide Field Surveys in Cosmology*, volume 14, page 89, January 1998. URL <https://ui.adsabs.harvard.edu/abs/1998wfsc.conf...89D>.
- [57] I. N. Reid, C. Brewer, R. J. Brucato, W. R. McKinley, A. Maury, and et al. The second palomar sky survey. *Publications of the Astronomical Society of the Pacific*, 103:661, jul 1991. doi: 10.1086/132866. URL <https://doi.org/10.1086/132866>.

- [58] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, and et al. Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168, Apr 2019. ISSN 1538-3881. doi: 10.3847/1538-3881/ab089d. URL <http://dx.doi.org/10.3847/1538-3881/ab089d>.
- [59] Sarah Brough, Chris Collins, Ricardo Demarco, Henry C. Ferguson, Gaspar Galaz, and et al. The vera rubin observatory legacy survey of space and time and the low surface brightness universe, 2020.
- [60] Yanxia Zhang and Yongheng Zhao. Astronomy in the Big Data Era. *Data Science Journal*, 14:11, May 2015. doi: 10.5334/dsj-2015-011. URL <https://ui.adsabs.harvard.edu/abs/2015DatSJ..14...11Z>.
- [61] Giuseppe Longo, Erzsébet Merényi, and Peter Tiño. Foreword to the focus issue on machine intelligence in astronomy and astrophysics. *Publications of the Astronomical Society of the Pacific*, 131(1004):100101, Sep 2019. ISSN 1538-3873. doi: 10.1088/1538-3873/ab2743. URL <http://dx.doi.org/10.1088/1538-3873/ab2743>.
- [62] Cunshi Wang, Yu Bai, C. López-Sanjuan, Haibo Yuan, Song Wang, Jifeng Liu, David Sobral, P. O. Baqui, E. L. Martín, Carlos Andres Galarza, J. Alcaniz, R. E. Angulo, A. J. Cenarro, D. Cristóbal-Hornillos, R. A. Dupke, A. Edero-clite, C. Hernández-Monteagudo, A. Marín-Franch, M. Moles, L. Sodré Jr. au2, H. Vázquez Ramió, and J. Varela. J-plus: Support vector machine applied to star-galaxy-qsoclassification, 2021.
- [63] Snigdha Sen, Sonali Agarwal, Pavan Chakraborty, and Krishna Pratap Singh. Astronomical big data processing using machine learning: A comprehensive review. *Experimental Astronomy*, 53(1):1–43, February 2022. doi: 10.1007/s10686-021-09827-4. URL <https://ui.adsabs.harvard.edu/abs/2022ExA...53...1S>.
- [64] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.
- [65] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL <https://proceedings.neurips.cc/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf>.

- [66] L. Rokach and O. Maimon. Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005. doi: 10.1109/TSMCC.2004.843247.
- [67] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. doi: 10.1080/00031305.1992.10475879. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879>.
- [68] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000. ISSN 0163-5808. doi: 10.1145/335191.335388. URL <https://doi.org/10.1145/335191.335388>.
- [69] Elke Aichtert, Christian Böhm, and Peer Kröger. Deli-clu: Boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In Wee-Keong Ng, Masaru Kitsuregawa, Jianzhong Li, and Kuiyu Chang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 119–128, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33207-7.
- [70] M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient Parameter-free Clustering Using First Neighbor Relations. *arXiv e-prints*, art. arXiv:1902.11266, February 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190211266>.
- [71] Xiang Wang and Tie Liu. Multiple Sample Clustering. *arXiv e-prints*, art. arXiv:1910.09731, October 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191009731W>.
- [72] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning, 2020.
- [73] Hideitsu Hino. Active learning: Problem settings and recent developments, 2020.
- [74] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08*, page 413–422, USA, 2008. IEEE Computer Society. ISBN 9780769535029. doi: 10.1109/ICDM.2008.17. URL <https://doi.org/10.1109/ICDM.2008.17>.
- [75] Kent A. Spackman. Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. In Alberto M. Segre, editor, *ML*, pages 160–163. Morgan Kaufmann, 1989.

- [76] J.A. Hanley and Barbara Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 05 1982. doi: 10.1148/radiology.143.1.7063747.
- [77] E. Roberts, B. A. Bassett, and M. Lochner. Bayesian Anomaly Detection and Classification. <https://ui.adsabs.harvard.edu/abs/2019arXiv190208627R>, February 2019.
- [78] Rafael Garcia-Dias, Carlos Allende Prieto, Jorge Sánchez Almeida, and Ignacio Ordovás-Pascual. Machine learning in APOGEE. Unsupervised spectral classification with K-means. , 612:A98, May 2018. doi: 10.1051/0004-6361/201732134. URL <https://ui.adsabs.harvard.edu/abs/2018A&A...612A..98G>.
- [79] Ke Wang, Ping Guo, and A. Li Luo. A new automated spectral feature extraction method and its application in spectral classification and defective spectra recovery. , 465(4):4311–4324, March 2017. doi: 10.1093/mnras/stw2894. URL <https://ui.adsabs.harvard.edu/abs/2017MNRAS.465.4311W>.
- [80] Radamanthys Stivaktakis, Grigorios Tsagkatakis, Bruno Moraes, Filipe Abdalla, and et al. Convolutional neural networks for spectroscopic redshift estimation on euclid data. *IEEE Transactions on Big Data*, 6(3):460–476, Sep 2020. ISSN 2372-2096. doi: 10.1109/tbdata.2019.2934475. URL <http://dx.doi.org/10.1109/TBDATA.2019.2934475>.
- [81] Lior Shamir. Automatic identification of outliers in hubble space telescope galaxy images, 2021.
- [82] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.
- [83] J. Vega-Ferrero, H. Domínguez Sánchez, M. Bernardi, M. Huertas-Company, R. Morgan, and et al. Pushing automated morphological classifications to their limits with the dark energy survey, 2020.
- [84] Kate Storey-Fisher, Marc Huertas-Company, Nesar Ramachandra, Francois Lanusse, Alexie Leauthaud, and et al. Anomaly detection in astronomical images with generative adversarial networks, 2020.
- [85] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, and et al. Generative adversarial networks, 2014.
- [86] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv e-prints*, art. arXiv:1701.07875, January 2017. URL <https://ui.adsabs.harvard.edu/abs/2017arXiv170107875A>.

- [87] X. Huang, C. Storfer, V. Ravi, A. Pilon, M. Domingo, D. J. Schlegel, S. Bailey, A. Dey, R. R. Gupta, D. Herrera, and et al. Finding strong gravitational lenses in the desi decam legacy survey. *The Astrophysical Journal*, 894(1):78, May 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab7ffb. URL <http://dx.doi.org/10.3847/1538-4357/ab7ffb>.
- [88] X. Huang, C. Storfer, A. Gu, V. Ravi, A. Pilon, and et al. Discovering new strong gravitational lenses in the desi legacy imaging surveys. *The Astrophysical Journal*, 909(1):27, Mar 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abd62b. URL <http://dx.doi.org/10.3847/1538-4357/abd62b>.
- [89] Mike Walmsley, Anna M. M. Scaife, Chris Lintott, Michelle Lochner, Verlon Etsebeth, and et al. Practical galaxy morphology tools from deep supervised representation learning, 2021.
- [90] Franco D. Albareti, Carlos Allende Prieto, Andres Almeida, Friedrich Anders, Scott Anderson, and et al. The 13th Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-IV Survey Mapping Nearby Galaxies at Apache Point Observatory. , 233(2):25, December 2017. doi: 10.3847/1538-4365/aa8992. URL <https://ui.adsabs.harvard.edu/abs/2017ApJS...233...25A>.
- [91] M. Lochner and B.A. Bassett. Astronomy: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481, Jul 2021. ISSN 2213-1337. doi: 10.1016/j.ascom.2021.100481. URL <http://dx.doi.org/10.1016/j.ascom.2021.100481>.
- [92] François Lanusse, Quanbin Ma, Nan Li, Thomas E. Collett, Chun-Liang Li, and et al. Cmu deeplens: deep learning for automatic image-based galaxy–galaxy strong lens finding. *Monthly Notices of the Royal Astronomical Society*, 473(3): 3895–3906, Jul 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx1665. URL <http://dx.doi.org/10.1093/mnras/stx1665>.
- [93] H. Domínguez Sánchez et al. Transfer learning for galaxy morphology from one survey to another. *Mon. Not. Roy. Astron. Soc.*, 484(1):93–100, 2019. doi: 10.1093/mnras/sty3497.
- [94] Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Aboul Ella Hassanien, and I. M. Selim. Deep galaxy: Classification of galaxies based on deep convolutional neural networks, 2017.
- [95] P.H. Barchi, R.R. de Carvalho, R.R. Rosa, R.A. Sautter, M. Soares-Santos, and et al. Machine and deep learning applied to galaxy morphology - a comparative study. *Astronomy and Computing*, 30:100334, Jan 2020. ISSN 2213-1337. doi: 10.

- 1016/j.ascom.2019.100334. URL <http://dx.doi.org/10.1016/j.ascom.2019.100334>.
- [96] Mitchell K Cavanagh, Kenji Bekki, and Brent A Groves. Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs. *Monthly Notices of the Royal Astronomical Society*, 506(1):659–676, 06 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1552. URL <https://doi.org/10.1093/mnras/stab1552>.
- [97] Tharindu Jayasinghe, Don Dixon, Matthew S Povich, Breanna Binder, Jose Velasco, and et al. The milky way project second data release: bubbles and bow shocks. *Monthly Notices of the Royal Astronomical Society*, 488(1):1141–1165, Jun 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz1738. URL <http://dx.doi.org/10.1093/mnras/stz1738>.
- [98] Mike Walmsley, Chris Lintott, Tobias Geron, Sandor Kruk, Coleman Krawczyk, and et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies, 2021.
- [99] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.
- [100] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. doi: 10.1109/MIS.2009.36.
- [101] Adriana Dropulic, Bryan Ostdiek, Laura J. Chang, Hongwan Liu, Timothy Cohen, and et al. Machine learning the sixth dimension: Stellar radial velocities from 5d phase-space correlations. *The Astrophysical Journal Letters*, 915(1):L14, jul 2021. doi: 10.3847/2041-8213/ac09ef. URL <https://doi.org/10.3847/2041-8213/ac09ef>.
- [102] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.
- [103] Amulya Agarwal and Nitin Gupta. Comparison of outlier detection techniques for structured data, 2021.
- [104] DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, and et al. The desi experiment part i: Science, targeting, and survey design, 2016.
- [105] Edward L. Wright, Peter R. M. Eisenhardt, Amy K. Mainzer, Michael E. Ressler, Roc M. Cutri, and et al. The Wide-field Infrared Survey Explorer (WISE): Mission

- Description and Initial On-orbit Performance. , 140(6):1868–1881, December 2010. doi: 10.1088/0004-6256/140/6/1868. URL <https://ui.adsabs.harvard.edu/abs/2010AJ....140.1868W>.
- [106] Desi dr8 z-band depth magnitude. <https://datalab.noao.edu/ls/ls.php>. Accessed: 2021-08-04.
- [107] Decals depth plots for various bands. <https://www.legacysurvey.org/dr8/description/>. Accessed: 2021-09-01.
- [108] M. Lochner and B. A. Bassett. ASTRONOMALY: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481, July 2021. doi: 10.1016/j.ascom.2021.100481. URL <https://ui.adsabs.harvard.edu/abs/2021A&C....3600481L>.
- [109] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, and et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [110] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.
- [111] David Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 03 1996. doi: 10.1162/neco.1996.8.7.1341.
- [112] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [113] G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [114] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.
- [115] Dr. M. Lochner. Astronomaly. <https://github.com/MichelleLochner/astronomy>, 2020.
- [116] Chorng-Yuan Hwang and Ming-Yan Chang. A Catalog of Morphologically Identified Merging Galaxies. , 181(1):233–237, March 2009. doi: 10.1088/0067-0049/181/1/233. URL <https://ui.adsabs.harvard.edu/abs/2009ApJS..181..233H>.
- [117] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel. Big universe, big data: Machine learning and image analysis

- for astronomy. *IEEE Intelligent Systems*, 32(2):16–22, Mar 2017. ISSN 1541-1672. doi: 10.1109/mis.2017.40. URL <http://dx.doi.org/10.1109/MIS.2017.40>.
- [118] Noble Kennamer, Emille E. O. Ishida, Santiago Gonzalez-Gaitan, Rafael S. de Souza, Alexander Ihler, and et al. Active learning with respect: Resource allocation for extragalactic astronomical transients, 2020.
- [119] E. E. de Oliveira Ishida, M. V. Kornilov, K. L. Malanchev, M. V. Pruzhinskaya, A. A. Volnova, and et al. Active anomaly detection for time-domain discoveries. *Astronomy Astrophysics*, Apr 2021. ISSN 1432-0746. doi: 10.1051/0004-6361/202037709. URL <http://dx.doi.org/10.1051/0004-6361/202037709>.
- [120] Lars Doorenbos, Stefano Cavuoti, Massimo Brescia, Antonio D’Isanto, and Giuseppe Longo. Comparison of outlier detection methods on astronomical image data. *Emergence, Complexity and Computation*, page 197–223, 2021. ISSN 2194-7295. doi: 10.1007/978-3-030-65867-0_9. URL http://dx.doi.org/10.1007/978-3-030-65867-0_9.
- [121] Lior Shamir. Automatic identification of outliers in hubble space telescope galaxy images. *Monthly Notices of the Royal Astronomical Society*, 501(4):5229–5238, Jan 2021. ISSN 1365-2966. doi: 10.1093/mnras/staa4036. URL <http://dx.doi.org/10.1093/mnras/staa4036>.
- [122] Liang Xiong, Barnabás Póczos, Andrew J. Connolly, and Jeff G. Schneider. Anomaly detection for astronomical data. 2010.