# Galaxy Cluster Detection and Characterisation in the Big Data Era

Matthew C. Chan


Physics

Department of Physics

Lancaster University

August 2022

A thesis submitted to Lancaster University for the degree of Doctor of Philosophy in the Faculty of Science and Technology

*Supervised by Dr. John P. Stott*

# Abstract

In this thesis, we present proof-of-concept studies that describe how data-driven techniques can be applied to observational data in order to detect and estimate properties of galaxy clusters, which are the largest gravitationally bound objects to have assembled in the Universe. Given the significance of clusters in astrophysics and cosmology, it is important to develop automated methods that are able to efficiently detect and process a large sample of clusters from existing photometric datasets, in preparation for upcoming large-scale galaxy surveys. This can be achieved by employing machine learning algorithms that are suited at solving the tasks at hand. In particular, algorithms that can self-learn the importance of features from the labelled data of known clusters, which minimises the amount of manual input required to make accurate predictions.

Initially, we demonstrate how a popular object detection algorithm can be applied to wide-field colour images to identify and predict the astronomical coordinates of clusters. We then demonstrate how a novel ensemble regression algorithm can be applied to line-of-sight galaxies within colour-magnitude space to estimate the photometric redshift of clusters. Finally, we present a hybrid empirical and analytical model that performs background subtraction of field galaxies along the line-of-sight of clusters within colour-magnitude space and then estimates the richness of clusters within a characteristic radius.

We also compare our findings with the results of existing conventional techniques to examine the overall predictive performance of our methods at generalising to unseen instances. Furthermore, we note that our methods can be combined together into a sequential data pipeline

to create a comprehensive catalogue that contains key characteristics (e.g. position, distance, mass) of observed clusters for conducting astrophysical and cosmological research.

This thesis is dedicated to all those that have supported me throughout my life.

# Acknowledgements

If I could travel back in time and give myself some advice before beginning the PhD, I would say "If you are hardworking and have a strong determination to succeed, then with proper guidance, confidence and patience from a supportive, understanding and accommodating environment, anything can be achieved". I would never have imagined of reaching this stage if these words were not down-to-earth.

Firstly, I would like to thank my PhD supervisor, Dr. John P. Stott, for his scientific insight, invaluable feedback and ensuring that obtaining a positive experience is the most important aspect when embarking on a PhD journey at Lancaster University. Secondly, I would like to thank my partner, Fong Fong Lee, for her endless emotional support to give me the motivation and inspiration needed to complete this thesis. Thirdly, I would like to thank my family for their unwavering physical support to enable me to effectively conduct my research during the COVID-19 pandemic. Fourthly, I would like to thank my internship supervisor, Ray Eitel-Porter, for his inspiring leadership and mentoring during my six-month data science placement at Accenture (see the following links to view a published paper that I coauthored as well as a blog of my personal experiences while I was working at Accenture: https://arxiv.org/abs/2207.09833 and https://www.accenture.com/gb-en/blogs/blogs-my-internship-journey-galaxies-ethical). Lastly, I would like to thank all of my PhD office peers, friends, colleagues, acquaintances and staff at Lancaster University for their thoughtful recommendations and interesting discussions on a variety of different topics over the last several years.

In addition, I would like to give gratitude to the Science and Technology Facility Council for providing studentship funding to make this

opportunity possible. Furthermore, I would also like to give gratitude to the 4IR Centre for Doctoral Training for hosting data science workshops and events.

# Declaration

This thesis is my own work and no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other institute of learning.

The research presented in this thesis has also been published in paper format by a peer-reviewed scientific journal. In particular, chapter 2 has been published as Chan & Stott (2019, MNRAS, https://doi.org/10.1093/mnras/stz2936), chapter 3 has been published as Chan & Stott (2021, MNRAS, https://doi.org/10.1093/mnras/stab858) and chapter 4 has been published as Chan & Stott (2022, MNRAS, https://doi.org/10.1093/mnras/stac2210).

"*What we know is a drop, what we don't know is an ocean.*"
Sir Isaac Newton

"*No great discovery was ever made without a bold guess.*"
Sir Isaac Newton

"*If I have seen further it is by standing on the shoulders of Giants.*"
Sir Isaac Newton

# Contents

# List of Figures

# Relevant Publications by the Author

Chapter 2

- "Deep-CEE I: Fishing for galaxy clusters with deep neural nets", **M. C. Chan**, J. P. Stott, 2019, MNRAS, https://doi.org/10.1093/mnras/stz2936.

Chapter 3

- "Z-Sequence: Photometric redshift predictions for galaxy clusters with sequential random k-nearest neighbours", **M. C. Chan**, J. P. Stott, 2021, MNRAS, https://doi.org/10.1093/mnras/stab858.

Chapter 4

- "AutoEnRichness: A hybrid empirical and analytical approach for estimating the richness of galaxy clusters", **M. C. Chan**, J. P. Stott, 2022, MNRAS, https://doi.org/10.1093/mnras/stac2210.

# Chapter 1

# Introduction

## 1.1 Key characteristics of galaxy clusters

Galaxy clusters sit at the top of the hierarchical mass assembly of gravitationally bound objects. The formation of clusters has been taking place since the early epochs of the Universe from initial density perturbations, through to a series of mergers and accretion of matter (Kravtsov & Borgani, 2012). This build-up process is considered as complete when a cluster has become virialised, where the total gravitational potential energy of the cluster is equal to twice the negative total kinetic energy of its galaxy members, in order to establish a stable and self-gravitating equilibrium.

It is common to categorise clusters based on their optical morphology (e.g. Zwicky et al. 1961; Abell 1965; Bautz & Morgan 1970; Rood & Sastry 1971). These categories can be broadly grouped as regular and irregular, where regular clusters have a spherical appearance and a noticeable core region (e.g. the Coma cluster of galaxies), whilst irregular clusters have a non-spherical appearance and no noticeable core region (e.g. the Hercules cluster of galaxies). It is also typical to find the brightest cluster galaxy situated at the bottom of the gravitational potential well of clusters (Lin & Mohr, 2004), where brightest cluster galaxies are one of the most luminous and largest types of galaxies to exist in the Universe.

The intrinsic properties of clusters encompass a wide range of values, such

that clusters can be further divided into subcategories based on their optical richness (Bahcall, 1999). These subcategories can be broadly grouped as 'poor' and 'rich', where poor clusters have intrinsic properties towards the lower end of the scale (e.g. the Fornax cluster of galaxies) when compared with rich clusters (e.g. the Perseus cluster of galaxies). For instance, the number of galaxies found in poor and rich clusters can vary from tens to thousands inside a relatively small volume, with a radius that can stretch from hundreds of kpc to several Mpc. The velocities of galaxy members in poor and rich clusters can reach from hundreds of km $s^{-1}$ to thousands of km $s^{-1}$, with the total mass of clusters spanning from $10^{13}$ to $10^{15}$ solar masses.

It is important to note that galaxies make up the smallest fraction (i.e. approximately 3 per cent) of the total mass in clusters (Pratt et al., 2019). A large fraction (i.e. approximately 12 per cent) of the total mass in clusters is found in the hot gas that permeates the intra-cluster medium, where temperatures can range from several keV to tens of keV due to the shock heating of infalling gas into the deep potential well of clusters (Takizawa & Mineshige, 1998). The dominant fraction (i.e. approximately 85 per cent) of the total mass in clusters exists in the form of dark matter, which has a strong gravitational influence within clusters such that it constrains the majority of galaxy members from dispersing due to their relatively high peculiar velocities (Zwicky, 1937). Although, for the work in this thesis we will mainly focus on describing the optical attributes of clusters since we only employ optical data to study clusters.

Observational measurements of clusters are essential for validating the physical behaviour of simulated clusters, where simulations are powerful tools to pragmatically examine the inner workings of clusters as well as probe the nature of large-scale structure. For example, numerous models have been proposed to describe the density profile (e.g. Navarro et al. 1997; Moore et al. 1999; Navarro et al. 2004; Merritt et al. 2006) of hypothetical dark matter halos, where dark matter halos contain relatively high concentrations of dark matter that fully encompass clusters such that they are not affected by cosmological expansion. These models have subsequently been applied to simulations to unravel the evolution of galaxies in overdense environments (e.g. Sensui et al. 1999; Ghigna et al. 2000; Bullock et al. 2001; Takahashi et al. 2002). Moreover, numerous models of the

halo mass function (e.g. Press & Schechter 1974; Sheth & Tormen 1999; Jenkins et al. 2001; Tinker et al. 2008) have been derived from cosmological simulations, where the halo mass function describes the number density of dark matter halos per unit mass at different redshifts for specific cosmological assumptions. These models have subsequently been applied to observations of clusters to constrain the precision of cosmological parameters in cosmological models (e.g. Rozo et al. 2009; Vikhlinin et al. 2009; Planck Collaboration et al. 2016a; Bocquet et al. 2019).

The population of galaxies in a cluster can be roughly divided into two distributions that are nominally known as the 'red-sequence' and 'blue-cloud' (Eales et al., 2018). The term red-sequence refers to an apparent narrow ridgeline that is visible within colour-magnitude space consisting of early-type galaxies (i.e. elliptical and lenticular galaxies) that are 'red' in colour due to the lack of star-formation in these galaxies. It should be noted that there is a slight tilting of the red-sequence within colour-magnitude space, where brighter red-sequence galaxies are redder than fainter red-sequence galaxies. This is mainly due to the brighter red-sequence galaxies being more massive than the fainter red-sequence galaxies (Kodama & Arimoto, 1997), such that these galaxies retain more metals within their systems after internal supernovae events and thus there is a higher concentration of metallicity in the atmospheres of their stellar populations. Correspondingly, the term blue-cloud refers to an apparent loose clumping that is visible within colour-magnitude space consisting of late-type galaxies (i.e. spiral and irregular galaxies) that are 'blue' in colour due to high star-formation rates in these galaxies.

Environmental processes within clusters also impact the star-formation rate of galaxies by reducing the amount of cold gas and dust available for the galaxies to create new stars via gravitational collapse. This effect can be seen in the morphology-density relation (Dressler 1980a; Dressler et al. 1997) of galaxies in clusters, where early-type galaxies are typically found in high density environments whilst late-type galaxies are typically found in low density environments. The environmental processes that drive this effect mainly involve ram-pressure stripping (Gunn & Gott, 1972), tidal stripping (Gallagher & Ostriker, 1972) and strangulation (Larson et al., 1980). For ram-pressure stripping, this process blows

the cold gas and dust out from galaxies when the galaxies move through the cluster and collide with the hot gas of the intra-cluster medium. For tidal stripping, the cold gas and dust are tugged out from galaxies when they gravitationally interact with each other (e.g. flybys or mergers) within a cluster. Strangulation restricts the access of galaxies to fresh cold gas and dust, due to the hot gas in the intra-cluster medium heating up any infalling gas, which prevents it from accumulating to form new stars. We note that these environmental processes have a more profound influence towards the inner regions of a cluster, where the density of galaxies and the density of hot gas in the intra-cluster medium is considerably higher.

Many of the known characteristics of clusters have been learned through the tireless efforts of researchers that have mostly worked with relatively small and selective samples of clusters. It is needless to say that there are still many clusters in the Universe that have yet to be discovered or properly examined. It is therefore extremely important to develop cluster cataloguing methods that can perform efficient and effective data-processing from modern observing strategies. As a result, this will enable larger and more diverse samples of clusters to be studied and thus further enhance our current understandings of astrophysics and cosmology.

## 1.2 History of galaxy cluster cataloguing

One of the earliest documented observations of a galaxy cluster in the literature was by Charles Messier in 1781 as part of his catalogue of nebulae and star clusters (Messier, 1781). He noted that there was a large collection of nebulae in the direction of the Virgo constellation, which would later become known as the Virgo Cluster. Further observations of nebulae would continue to be catalogued over the subsequent centuries by various researchers (e.g. Herschel 1785; Dunlop 1828; Herschel 1864; Lassell & Marth 1867; Schultz 1875; Searle 1880; Dreyer 1888; Bailey 1908; Curtis 1918; Innes 1924).

It was not until 1929 that nebulae were widely recognised as extra-galactic objects, when Edwin Hubble published his work (Hubble, 1929b) showing that the

Andromeda nebula was most likely an entirely separate galaxy. He calculated that it was too distant to be part of our own galaxy by using the period-luminosity relationship of observed Cepheid variable stars within the Andromeda nebula. This finding helped to pave the way for a new era of discoveries and exploration in extra-galactic astronomy.

From 1947 to 1957, Charles Shane and Carl Wirtanen conducted one of the first large-scale galaxy surveys from the Lick Observatory to count galaxies in the northern hemisphere. In 1954, they published a preliminary study of their observations (Shane & Wirtanen, 1954), which indicated that galaxies are not randomly distributed but tend to cluster together into groupings. This result subsequently signified the importance of conducting cluster cataloguing as an effective means of examining the overall distribution of matter in the Universe.

In the following years, the first major cluster catalogue to be published was by George Abell in 1958 (Abell 1958, N.B. for the remainder of this thesis we refer to the clusters in this catalogue as Abell clusters), where clusters were detected through manual inspection of photographic plates from the Palomar Sky Survey (Abell, 1959). For this catalogue, a successful cluster detection had to satisfy the following selection criteria: the cluster must have at least 50 members within a (1.7 arcminute)/($z$) radius of the cluster centre; the cluster must have at least 50 members within two magnitudes fainter than the third brightest member; the cluster must be between a redshift range of $0.02 \leq z \leq 0.2$. He documented the key characteristics of 2712 observed clusters, such as the position (i.e. based on astronomical coordinates), distance (i.e. based on the magnitude of the tenth brightest galaxy member) and richness (i.e. based on the number of observed galaxies within a magnitude range and within a fixed search radius after applying a statistical background subtraction to account for field galaxies). It should be noted that richness is a direct proxy of the mass of clusters via the mass-richness relation (Johnston et al., 2007).

Not long afterwards, another major cluster catalogue was published by Fritz Zwicky et al. in 1961 (Zwicky et al. 1961, N.B. for the remainder of this thesis we refer to the clusters in this catalogue as Zwicky clusters), where clusters were also detected through manual inspection of photographic plates from the Palomar Sky Survey. For this catalogue, a successful cluster detection had to satisfy the

following selection criteria: the cluster must have at least 50 members within a density boundary that is twice the density of the local field; the cluster must have at least 50 members within three magnitudes fainter than the brightest member. They also documented the key characteristics of 9700 observed clusters, such as the position (i.e. based on astronomical coordinates), distance (i.e. based on the magnitude and angular size of galaxy members) and richness (i.e. based on the number of observed galaxies within a magnitude range and within a density isopleth[1] map after applying a statistical background subtraction to account for field galaxies).

However, it should be noted that the manual inspection of photographic plates to detect the Abell and Zwicky clusters is an extremely inefficient method. This is because it is very time-consuming to physically examine every single object within the images. In addition, it is difficult for other researchers to replicate this method in order to check for systematic errors in human-made measurements. Although, since the creation of these catalogues, numerous researchers have proposed new automated approaches that can efficiently perform large-scale cataloguing of clusters. We will now briefly describe several notable automated cluster cataloguing methods that have been developed for determining the key characteristics of clusters.

Turner & Gott (1976) developed a surface density searching algorithm that detects clusters by examining the measured astrometry information of individual galaxies from galaxy surveys. To obtain position measurements, this approach initially involves defining a circular border around each galaxy, where the radius of the circular border varies depending on the local surface density within a user-specified angular radius. Lastly, regions that have galaxies with overlapping circular borders are merged together as galaxy members of a clump, which indicates the presence of a cluster. To obtain distance measurements, this approach relies on cross-matching identified galaxy members with external reference catalogues to determine their radial velocities and thus the average radial velocity or average redshift of the identified galaxy members in clusters can be computed, which are both direct proxies of the distance of clusters via Hubble's law (Hubble,

[1] An isopleth is a contour line that highlights positions of similar values (e.g. density) within a map.

1929a). However, this approach does not specifically describe how to determine the mass of clusters but the number of identified galaxy members found within the radius of the clumps could be used to determine cluster richness.

Huchra & Geller (1982) developed a galaxy linking algorithm that detects clusters by examining the measured astrometry, photometry and distance information of individual galaxies from galaxy surveys. To obtain position measurements, this approach involves applying a friends-of-friends algorithm that links pairs of galaxies together based on whether the observed lengths of their angular and redshift separations are less than or equal to corresponding angular and redshift separation thresholds. These thresholds vary depending on the local number density, which is computed by integrating a galaxy luminosity function with a faint-end absolute magnitude limit that corresponds to the limiting magnitude of the galaxy survey at the average redshift of the galaxy pairing. Lastly, overdense regions are identified if they contain a grouping of linked galaxies, which indicates the presence of a cluster. To obtain distance measurements, this approach adopts the redshift values of individual galaxies that have already been measured by external reference catalogues and thus the average redshift of identified galaxy members could be computed to determine the distance of clusters. To obtain mass measurements, this approach computes the mass-to-light ratio based on the total luminosity of identified galaxy members, velocity dispersion of identified galaxy members and the size of the grouping of linked galaxies. We note that later on in this thesis we use cluster catalogues that have been created from an adapted version of this approach by Wen et al. (2012) and Wen & Han (2015) (N.B. for the remainder of this thesis we refer to the clusters in these catalogues as the WHL12 and WH15 clusters respectively).

Postman et al. (1996) developed a filter matching algorithm that detects clusters by examining the measured astrometry and photometry of individual galaxies from galaxy surveys. To obtain position measurements, this approach initially constructs filter profiles that are based on analytical models of the surface density and luminosity of cluster and field galaxies at specific redshift intervals. The filter profiles are combined into a likelihood function that is then applied onto grids of the survey area, which results in the generation of a likelihood map. Lastly, each grid is examined to identify regions containing galaxies that maximise the

likelihood of the filter profiles at the best fit redshift interval, which indicates the presence of a cluster. To obtain distance measurements, this approach involves matching the observed filter profiles to the expected filter profiles at specific redshift intervals to determine the redshift of clusters. To obtain mass measurements, this approach establishes a scaling relation to compute cluster richness based on the strength of the signal for the observed filter profiles. We note that later on in this thesis we use a cluster catalogue that has been created from an adapted version of this approach by Szabo et al. (2011) (N.B. for the remainder of this thesis we refer to the clusters in this catalogue as AMF11 clusters).

Annis et al. (1999) developed a centroid finding algorithm that searches for the brightest cluster galaxy of clusters by examining the measured astrometry and photometry of individual galaxies from galaxy surveys. To obtain position measurements, this approach initially measures the likelihood of all galaxies being brightest cluster galaxies by comparing their filter magnitudes and colours with the photometric properties of typical brightest cluster galaxies at specific redshift intervals, where galaxies that have the highest likelihood out of all the neighbouring galaxies are assumed to be brightest cluster galaxies. The likelihood of the remaining neighbouring galaxies being associated with the identified brightest cluster galaxies is then measured. This involves examining whether the galaxies have filter magnitudes and colours that are consistent with an expected colour-magnitude relation (i.e. red-sequence galaxies) originating from the identified brightest cluster galaxies. Lastly, overdense regions are identified if they contain galaxies that maximise the likelihood of being red-sequence galaxies around the brightest cluster galaxies, which indicates the presence of clusters. To obtain distance measurements, this approach involves matching the observed brightest cluster galaxies and red-sequence galaxies to the expected brightest cluster galaxies and red-sequence galaxies at specific redshift intervals to determine the redshift of clusters. To obtain mass measurements, this approach counts the number of identified galaxy members within a radius and brightness range to determine cluster richness.

Gladders & Yee (2000) developed a red-sequence fitting algorithm that detects clusters by examining the measured astrometry and photometry of individual galaxies from galaxy surveys. To obtain position measurements, this ap-

proach initially defines overlapping slices within colour-magnitude space based on a model of the red-sequence at specific redshift intervals. The probability that a galaxy belongs within each slice is then computed, where galaxies that have low probabilities within a slice are removed from the slice. Next, weights are computed for each galaxy within the slices based on their absolute magnitude and probability. The surface density of the projected angular positions of galaxies is also computed and then combined with the weights to form a new probability map. Lastly, a user-specified probability threshold is applied to the probability map to identify overdense regions that have a strong red-sequence, which indicates the presence of a cluster. To obtain distance measurements, this approach involves finding out which slice contains the most identified galaxy members to determine the redshift of clusters. However, this approach does not specifically describe how to determine the mass of clusters but the number of probable galaxy members found within a slice could be used to determine cluster richness. We note that later on in this thesis we use a cluster catalogue that has been created from an adapted version of this approach by Rykoff et al. (2014) (N.B. for the remainder of this thesis we refer to the clusters in this catalogue as redMaPPer clusters).

Goto et al. (2002) developed a cut and enhancement algorithm that detects clusters by examining the measured astrometry and photometry of individual galaxies from galaxy surveys. To obtain position measurements, this approach initially applies a multitude of cuts to galaxies within colour space, where the cuts are designed to detect galaxies at specific redshift intervals. It should be noted that each cut is applied independently, such that any galaxy found within the cuts is considered as a detected galaxy. The angular separation and colour differences between the detected galaxies is then computed to generate a weighted map, where galaxies that are close together within angular and colour space receive large weighting. Lastly, a density-searching algorithm is applied to the weighted map to identify overdense regions that are above a user-specified density threshold and contain a minimum number of galaxies within the cuts, which indicates the presence of a cluster. To obtain distance measurements, this approach involves finding out which cut contains the most detected galaxies to determine the redshift of clusters. To obtain mass measurements, this approach counts the

number of identified galaxy members within a radius and brightness range to determine cluster richness.

It is worth noting that these described approaches for cluster cataloguing work mainly in optical/near-infrared wavelengths, where there has been an abundance of data collected by large-scale galaxy surveys over the past few decades (e.g. the Las Campanas Redshift Survey, Shectman et al. 1996; the Centre for Astrophysics 2 Redshift Survey, Falco et al. 1999; the IRAS PSCz Redshift Survey, Saunders et al. 2000; the Canadian Network for Observational Cosmology Field Galaxy Redshift Survey, Yee et al. 2000; the Sloan Digital Sky Survey, York et al. 2000; the Two Degree Field Galaxy Redshift Survey, Colless et al. 2001; the DEEP Extragalactic Evolutionary Probe 2 Redshift Survey, Davis et al. 2003; the Millennium Galaxy Catalogue Survey, Liske et al. 2003; the VIsible Multi Object Spectrograph Very Large Telescope Deep Survey, Le Fèvre et al. 2005; the United Kingdom Infra-red Telescope Infrared Deep Sky Survey, Lawrence et al. 2007; the zCOSMOS Redshift Survey, Lilly et al. 2007; the WiggleZ Dark Energy Survey, Blake et al. 2008; the Six Degree Field Galaxy Survey, Jones et al. 2009; the Galaxy And Mass Assembly Redshift Survey, Baldry et al. 2010; the Wide-field Infrared Survey Explorer All Sky Survey, Wright et al. 2010; the Two Micron All-Sky Survey Redshift Survey, Huchra et al. 2012; the Dark Energy Survey, Dark Energy Survey Collaboration et al. 2016; the Dark Energy Spectroscopic Instrument Legacy Imaging Surveys, Dey et al. 2019; the Hyper Suprime-Cam Subaru Strategic Program, Ishikawa et al. 2020).

In recent times, there has also been growing efforts to conduct cluster cataloguing with observational data from X-ray (e.g. Voges et al. 1999; Ebeling et al. 2001; Böhringer et al. 2004; Mehrtens et al. 2012; Liu et al. 2022b), Sunyaev-Zeldovich (e.g. Reichardt et al. 2013; Planck Collaboration et al. 2014; Planck Collaboration et al. 2016b; Ricci, M. et al. 2020; Hilton et al. 2021) and weak gravitational lensing surveys (e.g. Miyazaki et al. 2002; Hetterscheidt et al. 2005; Wittman et al. 2006; Dietrich et al. 2007; Gavazzi & Soucail 2007), where these surveys are especially convenient for providing a means of validating cluster detections in optical/near-infrared galaxy surveys or vice versa. We will now briefly describe how clusters can be catalogued in each of these survey types.

For X-ray surveys, X-ray radiation is mainly emitted through the thermal bremsstrahlung of charged particles in the hot gas of the intra-cluster medium, which results in the projection of extended X-ray sources that can be detected via X-ray telescopes. The distance of clusters can be directly determined by measuring the redshift of spectral features in X-ray gas spectra. The mass of clusters can be directly determined from combining measurements of the X-ray gas density with X-ray gas temperature (N.B. assuming that the X-ray gas in the intra-cluster medium is under hydrostatic equilibrium).

For Sunyaev-Zeldovich surveys, line-of-sight cosmic microwave background photons experience inverse Compton scattering when they encounter energetic electrons in the intra-cluster medium of clusters (i.e. the Sunyaev-Zeldovich [SZ] effect, Sunyaev & Zeldovich 1972), which results in fluctuations of the cosmic microwave background that can be detected via microwave telescopes. The distance of clusters can be approximated from combining measurements of the angular size of the cosmic microwave background fluctuation with X-ray surface brightness of the intra-cluster medium. The mass of clusters can be indirectly determined by establishing a scaling relation between the strength of the cosmic microwave background fluctuations with cluster masses that have been measured from other techniques (e.g. weak gravitational lensing).

For weak gravitational lensing surveys, the theory of general relativity suggests that larger masses have a greater influence on the curvature of the surrounding spacetime and thus there is a greater bending of the path taken by light, which results in the apparent non-random alignment of background galaxies (N.B. assuming that background galaxies are randomly orientated) behind a cluster that can be detected via high-resolution wide-field telescopes. However, there is yet to be an effective way of measuring the distance of clusters with weak gravitational lensing. Although, the mass of clusters can be directly determined by computing the cluster mass required to produce the observed alignment of background galaxies.

The scale of cluster cataloguing is expected to increase significantly over the upcoming decades as larger and more sensitive state-of-the-art telescopes are currently being built/deployed for conducting future large-scale galaxy surveys (e.g. SPHEREx All-Sky Optical to Near-Infrared Spectral Survey, Doré et al.

2016; the Legacy Survey of Space and Time, Ivezić et al. 2019; the James Webb Space Telescope Advanced Deep Extragalactic Survey, Endsley et al. 2020; the Nancy Grace Roman Space Telescope High Latitude Survey, Eifler et al. 2021; the *Euclid* Wide Survey, Scaramella et al. 2021). The completion of these data-intensive surveys will enable the next generation of researchers to probe further down the halo mass function as well as investigate the cluster population at even higher redshifts.

## 1.3 Framework of machine learning

Computers have revolutionised the way scientists conduct experiments ever since their creation by Charles Babbage in 1822 (Babbage & Davy, 1822). Evidently, without the assistance of computers, any type of analyses must be done manually by scientists. This is not practical in the modern information era, where working with large amounts of data has become the norm. As such, the automatic processing of information is highly desirable, to enable scientists to instead prioritise on developing theories and drawing conclusions from results.

During the 1950s, initial breakthroughs in the field of microelectronics (Moore, 1965) were important for the future of computing. These early developments would enable computers to evolve from executing simple calculations to perform more complex functions. Ultimately, this lead to the birth of a new scientific field, that was named as 'artificial intelligence' during a research conference at Dartmouth College in 1956 (McCarthy et al., 2006), where the term 'artificial intelligence' relates to the usage of machines to mimic intelligent behaviour.

Soon afterwards, a new branch developed under this field that focused on exploring how algorithms could self-learn from data to make accurate and robust data-driven decisions. This branch was coined as 'machine learning' in a publication by Arthur Samuel in 1959 (Samuel, 1959), where he described how a computer can be programmed to learn to play a game of checkers better than an average player could. This showcased the potential of how computers could be effectively utilised to make well-informed future choices from learning underlying relationships in historical data.

A notable publication by Frank Rosenblatt in 1958 introduced a novel algorithm, known as a 'perceptron' (Rosenblatt, 1958), which was inspired from earlier works (e.g. Rashevsky 1935; Mcculloch & Pitts 1943; Hebb 1949; Culbertson 1950; McCulloch 1950; Ashby 1952; Hayek 1952; Kleene 1956; Minsky 1956; Uttley 1956) that described how neurons in the brain process signals. Essentially, the perceptron can be seen as a 'universal approximator' that learns to find optimal weight coefficients for input features by updating the weights (initialised with random values) until the difference between the predicted and expected outputs are minimised. In machine learning, this is known as optimising the objective function, where the objective function is defined by the user to instruct an algorithm on what it should learn to do (i.e. minimise or maximise an objective depending on the choice of function for a given task). From which, the perceptron makes decisions by determining whether the sum of the product between the data of the input features and learned weight coefficients is above or below a decision threshold, as shown in Figure 1.1. This initial work would subsequently influence the development of more advanced machine learning algorithms.

In more recent times, the use of state-of-the art machine learning algorithms has become popular and widespread amongst many scientific fields where data is in abundance. For example, searching for exotic particles from high-energy collisions (Baldi et al., 2014); mapping of geological rocktypes (Harvey & Fotopoulos, 2016); optimising superconductor circuit designs (Menke et al., 2018); predicting isotropic lifetimes of heavy nuclei (Pérez & Balatsky, 2019); discovering chemicals with desired attributes (Tkatchenko, 2020); monitoring animal biodiversity (Chalmers et al., 2021) and modeling biological sequences (Muntoni et al., 2021).

The branch of machine learning can be further divided into the following three sub-branches: supervised learning (Cunningham et al., 2008); unsupervised learning (Ghahramani, 2004) and reinforcement learning (Arulkumaran et al., 2017). These sub-branches employ different learning approaches to govern the overall behaviour of machine learning algorithms, where an appropriate choice of sub-branch to use is dependent on the given task. We will now briefly discuss the learning approach of each sub-branch.

A supervised learning approach uses labelled data to evaluate the prediction errors of an algorithm, where it is the preferred approach in classification and re-

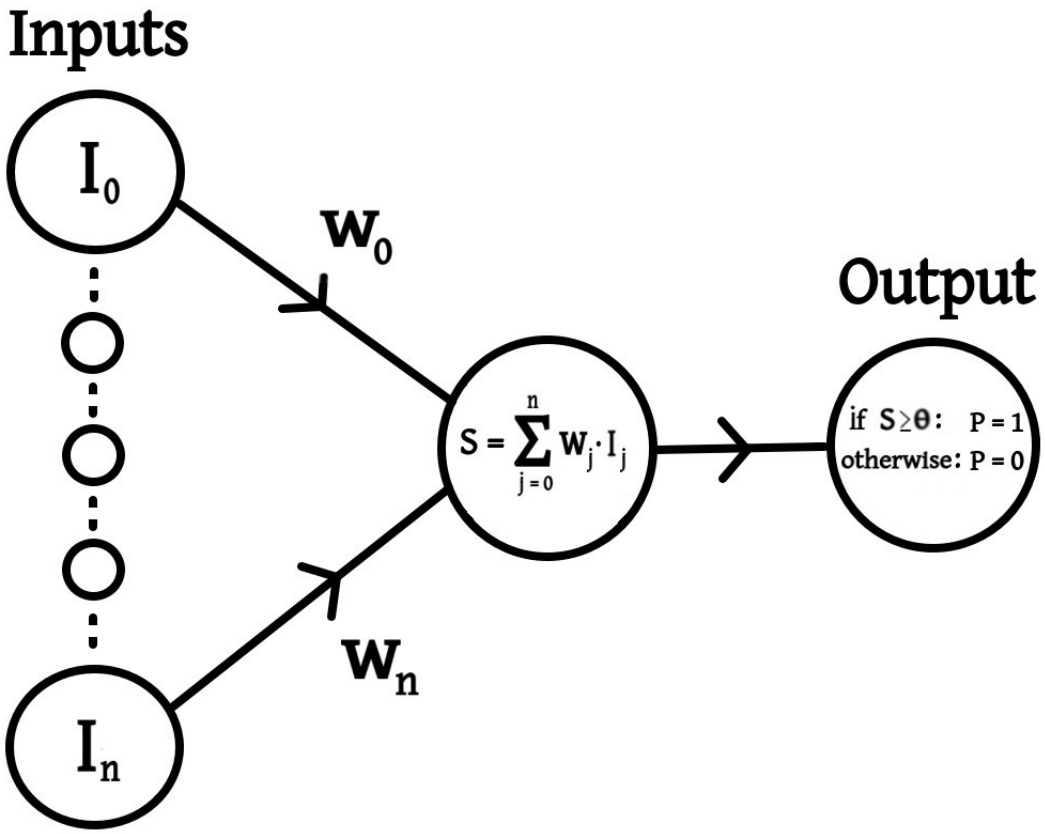


**Figure 1.1:** This figure displays a schematic diagram of the overall functionality of the perceptron algorithm at processing information from the input to output nodes, where n represents an input feature, $I_n$ represents an input feature value, $w_n$ represents a weight for the input feature, S represents the dot product between the input feature values and their corresponding weight values, $\theta$ represents the decision threshold and P represents the output class prediction. We note that all of the weights are updated at the same time during training, where the change in weight value is determined by the resultant class prediction error of an instance. The decision threshold is also learned during training by setting $I_0 = 1$ and $w_0 = -\theta$, which is known as the bias term. In addition, the output node is based on a unit step function. This diagram was inspired by Raschka & Mirjalili (2017).

gression tasks. Classification-based algorithms (e.g. decision trees, Breiman et al. 1984; support vector machine, Cortes & Vapnik 1995; naive Bayes, Rish 2001) should be provided with training data containing discrete labels to constrain the algorithm to learn to accurately predict the class of instances, such as classifying different types of amino acids based on their molecular properties (Barati Farimani & Aluru, 2018). Regression-based algorithms (e.g. ridge regression, Hoerl & Kennard 1970; lasso regression, Tibshirani 1996; linear regression, Altman & Krzywinski 2015) should be provided with training data containing continuous labels to constrain the algorithm to learn to accurately predict the outcome of instances, such as estimating the amount of rainfall based on observed cloud conditions (Meyer et al., 2016). We note that some machine learning algorithms are capable of conducting classification and regression tasks, such as decision trees and support vector machine.

An unsupervised learning approach allows an algorithm to explore and discover patterns or structures within unlabelled data, where it is the preferred approach in clustering and dimensionality reduction tasks. Clustering-based algorithms (e.g. k-means, Macqueen 1967; Gaussian mixture, Duda & Hart 1973; agglomerative hierarchical, Müllner 2011) attempt to group together instances that have similar attributes, such as identifying plants that may be part of the same species based on physical features of the plants (Hall et al., 2017). Dimensionality reduction-based algorithms (e.g. principal component analysis, Pearson 1901; singular value decomposition, Stewart 1993; kernel principal component analysis, Schölkopf et al. 1997) attempt to transform the feature space of data into a compressed representation that retains important information whilst also lessening noise, such as reducing the significance of redundant features in sensor data (Varshney & Willsky, 2011).

A reinforcement learning approach associates environmental interactions of an algorithm with rewards to encourage desirable behaviour, where it is the preferred approach in tasks with sequential actions. Sequential action-based algorithms (e.g. temporal difference learning, Sutton 1988; Q-learning, Watkins & Dayan 1992; policy gradient, Sutton et al. 1999) are trained by assessing the state of an environment after specific objectives have been met from a series of decisions by the algorithm. This encourages actions that result in positive rewards and

discourages actions that result in negative rewards, such as deploying a self-driving car that performs maneuvers based on real-time road and traffic conditions (Kiran et al., 2022).

In order to maximise the performance of machine learning algorithms, it is important to tune their hyper-parameters (i.e. external algorithmic settings that cannot be automatically learned but need to be manually configured). This can be achieved by using popular tuning strategies such as grid search (Hsu et al., 2003), random search (Bergstra & Bengio, 2012) and Bayesian optimisation (Snoek et al., 2012). Although, the appropriate choice of tuning strategy to use is dependent on the complexity of the hyper-parameter search space. We will now briefly discuss the methodology of each tuning strategy.

A grid search strategy examines the performance of every hyper-parameter combination, to determine the optimal hyper-parameter configuration. This strategy is only suitable for machine learning algorithms with a small number of hyper-parameters that are known to be highly influential, since it becomes more computationally expensive to examine every hyper-parameter combination when the size of the search space increases.

A random search strategy examines the performance of randomly sampled hyper-parameter combinations, to determine a near-optimal hyper-parameter configuration after conducting a specified number of sampling iterations. This strategy is effective for machine learning algorithms with an intermediary to large number of hyper-parameters, since it does not need to examine every hyper-parameter combination. It is also a cost efficient strategy when computational power is an issue.

A Bayesian optimisation strategy examines the performance of hyper-parameter combinations that have been sampled by a selection function, to determine a near-optimal hyper-parameter configuration. The selection function samples new hyper-parameter values that are likely to improve the performance of a machine learning algorithm based on a probability distribution of the performance of previously examined hyper-parameter combinations. This strategy is effective for machine learning algorithms with an intermediary number of hyper-parameters that have complex relationships, since it constructs a model of the probability space.

It is also important to ensure that machine learning algorithms do not significantly overfit (e.g. only learns to model noise) or underfit (e.g. performs worse than random guessing) to the data during training, as this reduces the overall effectiveness of algorithms at conducting a given task. This issue can be addressed by splitting the data into training and testing sets, which would allow algorithms to be cross-validated across different subsamples of the data to measure the true model performance. Some commonly used cross-validation strategies to assess algorithms are known as holdout cross-validation (Kohavi, 1995), k-fold cross-validation (Bengio & Grandvalet, 2004) and Monte Carlo cross-validation (Xu & Liang, 2001), where an appropriate choice of cross-validation strategy to use mainly depends on the available computing power. We will now briefly discuss the methodology of each cross-validation strategy. For simplicity, we will only mention the training and test sets but it should be noted that a validation set is also used when hyper-parameter tuning is involved.

A holdout cross-validation strategy involves randomly partitioning the data into a training set and a testing set, where the algorithm is trained exclusively on the training set and tested exclusively on the test set. This cross-validation strategy has low computational costs, which makes it suitable when working with large amounts of data, since the algorithm is only trained and assessed once on these two sets. However, due to the immutable nature of this cross-validation strategy, it would result in a less accurate evaluation of the true model performance when compared to the k-fold and Monte Carlo cross-validation strategies.

A k-fold cross-validation strategy involves randomly partitioning the data into a 'k' number of folds, where the algorithm is tested on each fold whilst also being trained on the remaining folds to measure the average performance across the folds. The computational cost of this cross-validation strategy is dependent on the number of folds used, where a larger number of folds would yield a more accurate evaluation of the true model performance but at higher computational costs since the algorithm will need to be trained and assessed on all folds. In addition, due to the mutable nature of this cross-validation strategy, it would result in a more accurate evaluation of the true model performance when compared to the holdout cross-validation strategy.

A Monte Carlo cross-validation strategy involves randomly partitioning the data into a training set and a testing set over a 'N' number of iterations, where the algorithm is trained and tested on the sets created within each iteration to measure the average performance across the iterations. The computational cost of this cross-validation strategy is dependent on the number of iterations used, where a larger number of iterations would yield a more accurate evaluation of the true model performance but at higher computational costs since the algorithm will need to be trained and assessed on 'N' training and test sets. Similar to k-fold cross-validation, this cross-validation strategy also has a highly mutable nature, which would result in a more accurate evaluation of the true model performance when compared to the holdout cross-validation strategy.

The deployment of a machine learning algorithm into production can be summarised in a series of development stages. The first stage is typically the data and algorithmic pre-processing stage, which may involve some of the following events: exploring relationships in the data; deciding whether machine learning is appropriate for the given task; deciding the learning approach; sampling the data; transforming the data; removing outliers from the data; selecting relevant features; splitting the data into training, validation and test sets; selecting relevant machine learning algorithms; selecting evaluation metrics. The second stage is typically the learning stage, which may involve some of the following events: training machine learning algorithms on instances in the training set; tuning hyper-parameters of the selected machine learning algorithms on instances in the validation set; selecting the machine learning algorithm that displays the best overall performance on instances in the validation set. The third stage is typically the assessment stage, which may involve some of the following events: applying a cross-validation strategy to examine how well the trained machine learning algorithm performs on unseen instances in the test set; deciding whether the trained machine learning algorithm is ready for deployment or requires additional training/testing.

## 1.4 History of machine learning in astronomy

The application of machine learning in astronomy has been relatively successful from when it was first utilised over a few decades ago. Given the vast number of individual objects that can be observed in astronomical surveys, it would be extremely time-consuming to catalogue objects without the assistance of automated tools. Especially, in the era of modern astronomy, where the total amount of data collected from astronomical surveys can range from terabytes to petabytes. As such, the integration of machine learning into data processing pipelines is an essential step towards automating historical procedures, which subsequently enables scientists to quickly gather and study large samples.

The earliest applications of machine learning in astronomy can be traced back to the 1990s, where machine learning was increasingly being used to classify astronomical objects (e.g. Fayyad et al. 1993; Djorgovski et al. 1994; Serra-Ricart 1994; Naim 1995; Weir et al. 1995; Owens et al. 1996).

A notable example is the task of distinguishing between stars and galaxies, which was traditionally approached using methods such as deciding a line of separation in density plots of the observed distributions (e.g. Kron 1980; Dickey et al. 1987; Kurtz et al. 1985; Heydon-Dumbleton et al. 1989) or matching the shape of the observed distributions with templates (e.g. Sebok 1979; Valdes 1982; Oegerle et al. 1986; Maddox et al. 1990). However, these methods do not make use of all the available information of objects, which meant that they would only be practical under specific conditions (e.g. within restricted brightness and image size regimes). Instead, Odewahn et al. (1992) described how to train an artificial neural network, using a set of 14 different image parameters, to classify whether images from the Palomar Sky Survey contained stars or galaxies. From which, they obtained a classification accuracy of approximately 95 per cent when identifying stars and galaxies across a wide range of brightnesses and image sizes without needing to make prior assumptions.

Another notable example is the task of classifying galaxies based on their morphological appearance, which was traditionally done by the manual eyeballing of human experts (e.g. Hubble 1926; Morgan 1958; de Vaucouleurs 1959; van den Bergh 1960). Undeniably, this method is impractical to replicate for very large

samples of galaxies. Instead, Storrie-Lombardi et al. (1992) described how to train an artificial neural network, using a set of 13 distance-independent measurements, to classify the morphology of galaxies from the ESO-LV catalogue (Lauberts & Valentijn, 1989). From which, they showed that the predictions of the trained neural network algorithm was in good agreement with manual eyeball classifications of five different galaxy morphologies (i.e. E, S0, Sa+Sb, Sc+Sd and Irr). Subsequently, these initial examples showcase the potential of employing machine learning in astronomy applications and how it can be effectively utilised to improve existing procedures.

During the 2000s, machine learning began to be applied to a wider variety of astronomy tasks as a means of gaining further insight into data or replacing existing conventional methods. We will now outline several applications of machine learning in astronomy from this period.

Fuentes & Gulati (2001) described how to train a k-nearest neighbour algorithm to predict stellar atmospheric parameters when given either spectra, spectral indices or spectral lines as an input. They demonstrate that a machine learning approach can be used to quantify the impact of different spectral features for obtaining accurate parameter estimates.

Estrada-Piedra et al. (2004) described how to train an ensemble of locally weighted regression algorithms to determine the age of stellar populations in galaxies when given high resolution spectra as an input. They demonstrate that the prediction time of a machine learning approach was much more efficient than an exhaustive search approach with similar levels of precision for predictions.

Wadadekar (2005) described how to train a support vector machine algorithm to estimate the photometric redshift of galaxies when given filter magnitudes and flux radii as inputs, where the support vector machine algorithm attempts to fit a hyperplane that minimises the margin of error (i.e. satisfying a user-specified prediction error threshold) between the data points and the hyperplane within the given feature space. They demonstrate that the overall prediction errors yielded by a machine learning approach was much lower than conventional template fitting methods.

d'Abrusco et al. (2007) described how to train a multi-stage machine learning model to generate a three dimensional map of astronomical objects in the

Universe. Initially, they trained an artificial neural network algorithm to estimate the redshift of galaxies using photometric data, which helped to break degeneracies in photometry of nearby and distant objects. Then, they applied a dimensionality reduction algorithm to transform a photometric feature space into a low-dimensional projection, which visually compressed similar objects together within the projection. Lastly, they used an agglomerative hierarchical algorithm to trace large-scale structure based on the apparent groupings along with the estimated photometric redshifts of galaxies, which separated the spatial distributions of different object types. Overall, they demonstrate that a machine learning approach can replace human interpretations of the visual appearance of large-scale structure in the Universe.

By the 2010s, further technology advancements in improving computer processing power via hardware accelerators (Thompson et al., 2020) would enable deep learning algorithms (LeCun et al., 2015) to be applied more frequently to astronomy tasks, where deep learning algorithms had evolved from research on experimenting with deeper configurations of the artificial neural network algorithm. One of the primary benefits of employing deep learning algorithms over traditional machine learning algorithms is that deep learning algorithms can automatically learn the importance of input features whereas traditional machine learning algorithms have a greater reliance on feature engineering[2] to perform well. This means that using a deep learning approach further reduces the amount of human input required, especially when working with raw datasets that contain many input features. We will now also outline several applications of deep learning in astronomy from this period.

Charnock & Moss (2017) described how to train a recurrent neural network algorithm to classify different types of supernovae when given photometric light curves as an input, where the unique looping architecture of a recurrent neural network algorithm enables it to process a sequence of inputs from different time steps to make predictions. They demonstrate that a deep learning approach is

[2]Feature engineering is part of the data pre-processing stage that involves selecting relevant features or transforming the feature space in order to reduce the dimensionality and complexity of an existing dataset. This helps machine learning algorithms to learn more efficiently by removing noise that could negatively impact the overall predictive power.

more time efficient than other conventional methods with similar levels of accuracy for examining early-epoch light curves of supernovae, which is advantageous for deciding whether to conduct spectroscopic follow-up observations before the supernovae become too faint to observe.

Shen et al. (2017) described how to train an autoencoder algorithm to recover gravitational wave signals that are immersed within actual instrumental and environmental noise. They demonstrate that a deep learning approach significantly outperforms other conventional denoising methods at processing low signal-to-noise data containing weak gravitational wave signals, which is valuable since the majority of observed gravitational waves have very weak signals.

Ackermann et al. (2018) described how to train a convolutional neural network algorithm to distinguish between a non-interacting galaxy and galaxy mergers when given a multi-band colour image as an input. They demonstrate that only a deep learning approach can match accuracy levels of human classifiers, which was not achievable with conventional galaxy merger detection methods.

Mishra et al. (2019) described how to train a generative adversarial network algorithm to simulate temperature anisotropy maps of the cosmic microwave background. They demonstrate that a deep learning approach can replicate the outputs of traditional cosmological simulation software within a much shorter running time-frame, which will be beneficial for future large-scale studies of the cosmic microwave background due to traditional cosmological simulation software being computationally expensive and time-consuming to run.

The examples we have described so far, have shown the wealth of astronomy tasks that machine learning can be applied to. We expect that in the remainder of the 2020s and beyond, machine learning will continue to play a major role in replacing many existing conventional approaches in astronomy with state-of-the-art automated tools. In particular, for the remainder of this thesis we focus our attention on the tasks of galaxy cluster detection, cluster redshift estimation and cluster richness estimation, where the application of machine learning in these tasks is still in its infancy.

## 1.5 This thesis

The overall aim of the work in this thesis is to demonstrate how modern data science methods can be utilised to develop automated algorithms that serve to efficiently and accurately process large quantities of observational data of clusters in large-scale galaxy surveys. This provides researchers with a set of powerful data-driven tools that minimise the need for making strong prior assumptions when measuring the key characteristics of clusters throughout the Universe, which ultimately improves our current understandings of astrophysics and cosmology.

Each of the main chapters in this thesis can be briefly summarised as follows:

- In chapter 2 we trained an object detection algorithm on a set of wide-field colour images of cross-matched Abell and WHL12 clusters between $0.1 < z < 0.2$ that had been observed in the Sloan Digital Sky Survey Data Release 9. We utilised transfer learning of internal parameters in the object detection algorithm, which reduced training time by reusing tuned internal parameters from an already pre-trained object detection model. In addition, we constrained the object detection algorithm to learn to recognise the observable characteristic features of clusters within a fixed radius from the cluster cores. We applied the trained object detection algorithm on unseen redMaPPer clusters to examine the overall accuracy at detecting clusters as well as examining the overall precision of our astronomical coordinate estimates.

- In chapter 3 we created a photometric dataset of line-of-sight galaxies within a fixed radius of cross-matched WHL12 and redMaPPer cluster cores between $0.05 \leq z \leq 0.6$ that had been observed in the Sloan Digital Sky Survey Data Release 9, where the clusters had their photometric redshifts determined by WHL12. We trained an ensemble regression algorithm to estimate the redshift of clusters by computing the average redshift of the k-nearest neighbour line-of-sight galaxies found in bootstrapped versions of the photometric dataset. In addition, we integrated a sequential feature selection strategy into the ensemble regression algorithm, which ensured that it only utilised photometry features that minimised the redshift prediction

error. We applied the trained ensemble regression algorithm on unseen WHL12 and redMaPPer clusters to examine the overall precision of our redshift estimates.

- In chapter 4 we first trained a reconstruction algorithm to replicate the photometry of individual cluster galaxies that were identified by AMF11 between $0.1 \leq z \leq 0.35$ in the Sloan Digital Sky Survey Data Release 16. We also sampled individual field galaxies from manually identified field regions. The algorithm then learned to distinguish between cluster and field galaxies based on the resultant reconstruction error, where cluster galaxies yielded smaller reconstruction errors and field galaxies yielded larger reconstruction errors. Next, we used cross-matched WH15 and redMaPPer clusters to establish a scaling relation that approximated the characteristic radius of clusters, where the scaling relation was between characteristic radius values determined by WH15 and the number of cluster galaxies identified by the algorithm within a fixed radius at the cluster redshift. We reapplied this learned scaling relation to the cross-matched WH15 and redMaPPer clusters and then resampled galaxies within the characteristic radius. Lastly, we obtained best fit parameter values from fitting the Schechter function to a composite luminosity distribution of identified cluster galaxies from cross-matched WH15 and redMaPPer clusters. We then used the best fit parameter values of the Schechter function to estimate the richness of individual clusters within a characteristic radius by refitting and integrating the Schechter function on the luminosity distribution of identified cluster members belonging to individual clusters. We measured the overall accuracy of the trained algorithm on unseen individual cluster and field galaxies from AMF11 and the field regions respectively. We also applied the best fit parameter values of the Schechter function on unseen cross-matched WH15 and redMaPPer clusters to examine the overall precision of our richness estimates.

Throughout this thesis, we adopted a Lambda cold dark matter cosmology with $H_0 = 71$ km s$^{-1}$ Mpc$^{-1}$, $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$.

# Chapter 2

# Deep-CEE I: Fishing for galaxy clusters with deep neural nets

# Abstract

We introduce Deep-CEE (**Deep** Learning for Galaxy **C**luster **E**xtraction and **E**valuation) in this proof-of-concept study of a novel deep learning technique that works directly with wide-field colour imaging to search for galaxy clusters without the need for photometric catalogues. This technique is complementary to traditional methods and could also be used in combination with them to confirm cluster candidates. We use a state-of-the-art object detection algorithm, that is adapted to localise and classify clusters from other astronomical objects in SDSS imaging. As there is an abundance of labelled data for clusters from previous classifications in publicly available catalogues, we do not need to rely on simulated data. This means we keep our training data as realistic as possible, which is advantageous when training a deep learning algorithm. Ultimately, we will apply our model to surveys such as the Legacy Survey of Space and Time and *Euclid* to probe wider and deeper into unexplored regions of the Universe. This will produce large samples of both high redshift and low mass clusters, which can be utilised to constrain both environment-driven galaxy evolution and cosmology.

## 2.1 Introduction

Throughout the 1950s to 1980s the astronomer George Abell and several others composed the Abell catalogue of rich galaxy clusters in the northern and southern hemispheres, where the completed catalogue ended up containing 4073 clusters in total (Abell et al., 1989). George Abell used a magnifying glass to manually examine photographic plates and looked specifically for over-dense regions of galaxies. This would be one of the last times a major wide-field cluster search was conducted manually by eye.

Since then a variety of techniques have been developed to search for clusters. One of the primary techniques for extracting clusters from imaging data is red-sequence fitting. Unlike the manual search method used by George Abell, this technique is applied to photometric catalogue data extracted from imaging, as opposed to the images themselves. In the charge-coupled device era, this catalogue-based technique has proven to be an efficient alternative to by-eye searches. In addition, both X-ray emission and the SZ effect reveal the presence of clusters through the properties of the hot intracluster medium. Furthermore, we remind the reader that clusters contain large concentrations of dark matter as well as tens to thousands of individual member galaxies, which means that their presence can be inferred via weak gravitational lensing.

However X-ray, SZ and weak gravitational lensing techniques will also need to optically confirm their candidate clusters, as there are contaminants (e.g. active galactic nuclei or nearby galaxies) and line-of-sight coincidences (e.g. unrelated low mass groups at different redshifts) that can conspire to give false positive detections. This confirmation has typically been done manually, which is time inefficient and can introduce biases that result in an uncertain selection function. Therefore, an approach that can produce fast and precise analysis of imaging data would be advantageous to both search for or confirm clusters.

The Legacy Survey of Space and Time (LSST, Ivezić et al. 2019) is soon to begin operations at the Vera Rubin Observatory, which is currently under construction in Chile with an expected first-light in 2023. LSST will be the deepest wide-field galaxy survey ever conducted, performing multiple scans of the entire southern sky over ten years, with an estimated 15 TB of data generated per night.

Correspondingly, *Euclid* (Scaramella et al., 2021) is a wide-field space telescope that is due to commence operation in 2023. It will conduct a weak gravitational lensing and galaxy clustering survey to probe the nature of dark matter and dark energy across a 15000 square degree region of sky with an estimated 0.3 PB of data generated per year over six years. This means that data mining techniques will be required to analyse the enormous outputs of these telescopes. LSST and *Euclid* are expected to observe thousands of previously unknown clusters across a wide range of masses and redshifts but cataloguing them presents a significant challenge.

Deep learning is very applicable in modern astronomy due to the abundance of data collected from past and present telescope surveys. This makes it a preferable technique when conducting data mining tasks such as classification and regression. However, at the time of this work a deep learning approach had yet to be developed for detecting and determining properties of clusters from optical wide-field imaging data.

A convolutional neural network (CNN, Fukushima 1980) is a particular type of deep learning algorithm that has been widely successful in the field of computer vision, where CNNs are designed to mimic the human brain at learning to perceive objects by activating specific neurons upon visualising distinctive patterns and colours. It is typical to train and utilise CNNs for processing high-dimensional features directly from digital images into a meaningful output with minimal human involvement. LeCun et al. (1998) first introduced a deep learning approach using CNNs to classify uniquely handwritten digits in images from the Modified National Institute of Standards and Technology (MNIST, LeCun et al. 2010) dataset achieving a classification error rate of less than 1 per cent.

It should be noted that conventional CNNs are adept at learning to recognise the visual features of objects but rather naive at self-determining their positions in an image since they process an image as a whole instead of its constituent parts. In this work, we aim to develop a deep learning approach that can efficiently localise and classify clusters in images. Szegedy et al. (2013) first demonstrated a deep learning approach to perform object detection in images by modifying the architecture of CNNs into modules that are specific to classification and local-

isation tasks, where objects with importance are classed as 'foreground' whilst everything else is considered as 'background'.

TensorFlow (Abadi et al., 2015) is an open source data science library that provides many high level application programming interfaces (API) for deep learning. One of these is the object detection API[3] (Huang et al., 2016) that contains multiple state-of-the-art deep learning algorithms, which are specifically designed to either enhance the speed or accuracy of an object detection model. These include Single Shot Detection (SSD, Liu et al. 2015) and Faster Region-based CNN (Faster R-CNN, Ren et al. 2015).

Huang et al. (2016) tested different object detection algorithms on images from the Common Objects in Context (COCO, Lin et al. 2014) dataset. In particular, they found that the Faster R-CNN algorithm returned high precision for predictions and was suitable for large input images during training and testing. However, the algorithm took a relatively long time to train and was slow at generating predictions. Correspondingly, they found that the SSD algorithm was relatively quick to train and produced fast predictions, but the overall precision of predictions was lower when compared to the Faster R-CNN algorithm. From which, we decided to choose the Faster R-CNN algorithm as we preferred accuracy over speed for predictions.

We organise this chapter in the following format. We split §2.2 into two subsections to outline our methodology. In §§2.2.1 we explain the concept behind the Deep-CEE model and in §§2.2.2 we describe the procedure to create the training and test sets. We also split §2.3 into two subsections to outline our results. In §§2.3.1 we analyse the performance of our model with the test set and in §§2.3.2 we assess our model on an unseen dataset. In §2.4 we discuss the limitations and future applications of our model. Finally, in §2.5 we summarise this work.

[3]The full list of object detection algorithms can be found via https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md and https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md

## 2.2 Method

### 2.2.1 Deep learning method

We used a supervised learning approach to train the Faster R-CNN algorithm (Ren et al., 2015) by providing it labelled images. The architecture of the algorithm can be seen in Figure 2.1. It is comprised of three different individual networks that work collectively to make predictions. These three networks are called the Feature Network (FN), Region Proposal Network (RPN) and Detection Network (DN). To train our model, we used a joint end-to-end training approach, which means we allowed the outputs from all the networks to be generated before all trainable layers are updated. In addition, we assigned one ground truth box per image, where the definition of the ground truth box can be found in §§2.2.2. We note that the learnable parameters in the RPN and DN were learned from scratch whereas pre-trained parameters were used for the FN. Throughout this subsection we adopted a similar methodology and hyper-parameters[4] as described in Ren et al. (2015) and Huang et al. (2016). We set a learning rate of 0.0002, momentum of 0.9, gradient clipping threshold of 10 and a mini-batch size of one. We used random initialisation of weights from a zero-mean truncated Gaussian distribution with a standard deviation of 0.01 for weights in the RPN. We used variance scaling initialisation (Glorot & Bengio, 2010) from a uniform distribution for weights in the DN. We also initialised bias values for the trainable layers in the RPN and DN to be zero.

[4]The definitions of typical hyper-parameters in a neural network is explained in more detail in Ruder (2016).

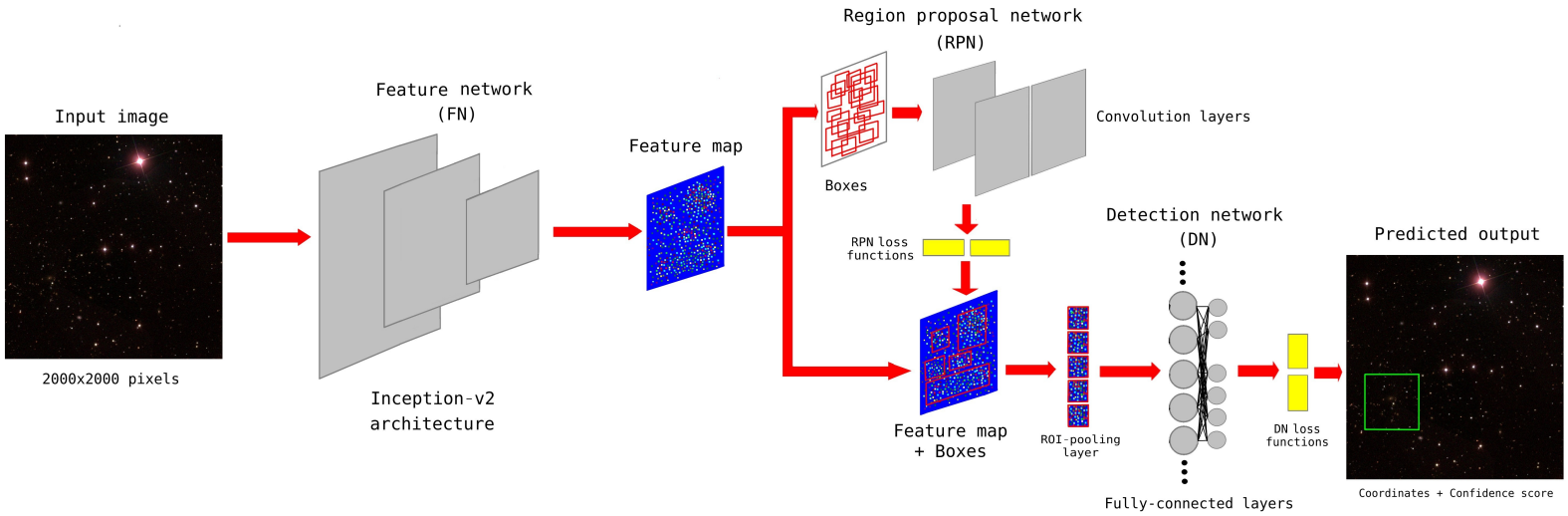


**Figure 2.1:** This figure displays a high-level overview of the architecture for the Faster R-CNN algorithm which contains the Feature Network, Region Proposal Network and Detection Network. The output from each network is used as the input for the next network. The architecture of the algorithm is similar to the system demonstrated in Figure 2 from Ren et al. (2015). For simplicity, the INCEPTION-V2 architecture is not displayed fully but it should be noted that '*Mixed_4e*' is used as the final layer of the Feature Network (Huang et al., 2016). The full details of the INCEPTION-V2 architecture can be found in Ren et al. (2018). The RPN and DN loss functions are only active during the training phase. A softmax activation function (Nwankpa et al., 2018) is used for performing classification of proposed boxes whilst a linear activation function is used for performing regression of coordinates.

#### 2.2.1.1 Feature network

The FN is found at the beginning of the Faster R-CNN algorithm and takes an image as its input. We applied transfer learning (Torrey & Shavlik, 2009) by using a pre-trained CNN called Inception-v2 (Szegedy et al., 2015) as the architecture of the FN. Inception-v2 consists of convolution layers, rectified linear unit (ReLU, Nair & Hinton 2010) activation functions, pooling layers, fully-connected (FC) layers and a softmax activation function. We note that we did not include the FC layers and softmax activation function for the FN since we only wanted to perform feature extraction in this network. The convolution layers and ReLU activation function were responsible for extracting non-linear features from an image (e.g. straight lines, edges, curves, blobs). The pooling layers were responsible for down-sampling the image to form a compressed feature map. The reason we chose to use the Inception-v2 architecture as opposed to other architectures (e.g. VGG16 (Simonyan & Zisserman, 2014) and AlexNet (Krizhevsky et al., 2012)) is that it has been specifically designed to reduce the overall number of parameters that need to be learned. This requires less computational cost to train the algorithm, while still retaining high accuracy. We note that Inception-v2 had been pre-trained to recognise objects from the COCO dataset, which contains images of commonly found objects in daily life (e.g. vehicles, animals, digital devices, accessories). This means we did not need to fully recalibrate the weights and biases in this network since they were sufficiently optimised at finding generic structures, where training every single weight and bias from scratch in this network would be computationally inefficient. Furthermore, we did not alter the architecture of Inception-v2.

#### 2.2.1.2 Region proposal network

The RPN is found after the FN and it consists of a shallow architecture of convolution layers with a ReLU activation function that is specific only to the first convolution layer. The weights and biases in the first convolution layer were shared for classification and localisation tasks whilst the remaining convolution layers were separated into parallel convolution layers, with independent weights and biases for each task. The RPN takes the feature map output from the FN as

its input. The role of the RPN was to determine the position of meaningful objects within an image. In the first convolution layer, a $3 \times 3$ pixel sliding window was used with zero-padding[5] and a pixel stride of one, which translates to every sixteenth pixel in the original image. At the centre of each sliding window, an 'anchor' is placed. Each anchor has a set number of different sized boxes generated around it, where the box dimensions and number of boxes was dependent on user-specified scaling and aspect ratios. From which, we set scaling ratios of 0.25, 0.5, 1.0 and 2.0 and aspect ratios of 0.5, 1.0 and 2.0. These ratios should reflect the dimensions of the ground truth boxes across all images. We note that a scaling ratio of 1.0 relates to a box of $256 \times 256$ pixels in the original image, where setting other values for the scaling ratio results in additional larger or smaller boxes at each anchor whilst setting other values for the aspect ratio results in boxes that have adjusted widths and heights with respect to each scaling ratio. In total, there were twelve boxes of different sizes at each anchor. In the final convolution layers, a $1 \times 1$ pixel sliding window was used with no-padding[6] and a pixel stride of one, which ensured a fixed dimensionality for the output of this layer.

The meaningfulness of objects was determined by calculating the amount of overlap between anchor boxes and the ground truth box in an image, where boxes that had a 70 per cent overlap or more with the ground truth box was assigned as a positive 'foreground' label whereas a negative 'background' label had a 30 per cent overlap or less. We calculated the overlap between two boxes as the percentage of the area overlap as a function of the total area of both boxes. In addition, boxes that had values between these overlap thresholds were ignored. We considered positive labelled boxes as meaningful objects and negative labelled boxes as irrelevant objects. From which, 128 positive labelled boxes and 128 negative labelled boxes were randomly chosen in each image to update the weights and biases. If there were fewer than 128 positive labelled boxes in an image, then additional negative labelled boxes with the next highest percentage overlap were chosen to represent positive labelled boxes. The RPN subsequently learned to

[5]This involved adding additional layers of pixels around the edge of an image with values of zero, which helped to preserve the dimensions of the input as it passed through the layer.

[6]This means that no additional pixels were added around the edge of an image.

identify positive 'foreground' objects and negative 'background' objects. Any box that was assigned a high probability by the RPN of containing a positive 'foreground' object was then passed onto the next stage in the Faster R-CNN algorithm. However, any box that was assigned a high probability by the RPN of containing a negative 'background' object was disregarded. Backpropagation (BP, Rumelhart et al. 1986) and stochastic gradient descent (SGD, Bottou 2010) was used to train the weights and biases in RPN[7]. We note that SGD is a variant of the conventional gradient descent (GD, Ruder 2016) algorithm, where the difference between SGD and GD is that SGD has a randomised component (e.g. random sampling/augmentation of the data) during each training iteration whilst GD has no randomised component.

Two additional steps were applied to limit the number of boxes for faster computation. Firstly, any box which extended outside the image borders was disregarded after the boxes were generated. Secondly, non-maximum suppression (NMS, Hosang et al. 2017) was used to keep the highest overlapping box with the ground truth box and also disregarded any remaining boxes that had a 70 per cent overlap or more with this box. These steps are repeated on the remaining boxes, such that the next box with the highest overlap was kept and any other box with a 70 per cent overlap or more with this box was also disregarded. This procedure continued until 300 boxes or fewer remained for each image.

We utilised two loss functions (log loss and smooth L1 loss) to calculate prediction errors of the RPN. The log loss function (Martinez & Stiefelhagen, 2018) worked with the output of a softmax activation function, which created a probabilistic distribution for each proposed box, such that the sum of the class probabilities for a proposed box equaled one. This function is described via the following equation:

$$L_{cls}(p_i, p_i^*) = -p_i^* log(p_i) - (1 - p_i^*) log(1 - p_i), \tag{2.1}$$

where $p_i$ is the predicted probability of a box and $p_i^*$ is zero or one depending

[7]We note that BP with respect to box coordinate proposals from the RPN was disabled as Huang et al. (2016) found that it caused the training to be unstable.

on whether the box was negatively or positively labelled respectively. The log loss function calculated the objectness error for boxes being predicted as 'foreground' and 'background'.

The smooth L1 loss function (Girshick, 2015) only considered positive labelled boxes in this work. It used the linear activation function to take into account the distance between the centre coordinates of the ground truth box and proposed boxes, and also the difference in size of the boxes when compared to the ground truth box. This function is described via the following equation:

$$L_{reg}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise,} \end{cases} . \tag{2.2}$$

where $x = (t_i - t_i^*)$ is the localisation error between the ground truth and proposed boxes. The smooth L1-loss function penalised the localisation error by taking the absolute value (behaves like L1-loss) for large errors and the square value (behaves like L2-loss) for small errors (Ng, 2004). This encouraged stable regularisation of the weights and biases.

The proposed boxes from the RPN were merged with the feature map from the FN, such that each box was overlaid on an 'object'. An ROI-pooling[8] (region-of-interest) layer was combined with remaining unused convolution layers of the Inception-v2 architecture to further extract features from the feature map. The ROI-pooling layer was ultimately responsible for merging the convolutional filter values within each box to ensure that every box contained the same number of values. This speeds up computations later on in the Faster R-CNN algorithm, as having fixed sized outputs leads to faster convergence.

#### 2.2.1.3 Detection network

The DN is found at the end of the Faster R-CNN algorithm and is composed of FC layers. The purpose of an FC layer was to combine all the outputs from the previous layer, this allowed for the algorithm to utilise all the processed information to effectively make decisions. The FC layers were run in parallel, such that

[8]It should be noted that Tensorflow's '*crop_and_resize*' operation was used for the ROI-pooling layer (Huang et al., 2016). This involved using bilinear interpolation to generate new cropped feature maps associated to each box from the previous feature map.

the weights and biases were split between classification and localisation. One of these two FC layers consisted of 2 neurons to categorise the outputs for classification, the other FC layer consisted of 4 neurons to predict the properties for box regression. Similar to the procedure for RPN, 16 positive and 48 negative labelled boxes were randomly chosen within each image to train the weights and biases in the DN, where boxes that had a 50 per cent overlap or more with the ground truth box was assigned as positive 'foreground' labels or otherwise negative 'background' labels. Additional negative boxes with the next highest overlap were assigned as positive labels if there were fewer than 16 positive labelled boxes. The log loss function, softmax activation function, smooth L1 loss function and linear activation function was again used to calculate the classification and localisation errors of the DN, where each loss function was also associated to its own FC layer. NMS was applied again using a 60 per cent threshold to reduce the number of overlapping boxes until 100 boxes or fewer remained per class. The weights and biases in the FC layers were also trained via BP and SGD[7]. We note that classification error was measured by comparing the assigned label of each box with the label of the ground truth box whilst localisation error was measured by calculating the difference between the pixel coordinates, height and width of the positive labelled boxes with the ground truth box.

Finally, since we decided to adopt a joint end-to-end training approach we could instead combine the loss functions of the RPN and DN into one multi-tasking loss function (Huang et al., 2016) to train the algorithm rather than training the RPN and DN separately. This is more computationally efficient as it simultaneously takes into account of all the prediction errors for the proposed boxes with the ground truth boxes. Therefore, the total loss of the algorithm was represented as the weighted sum of the objectness/classification and box regression losses. This multi-tasking loss function is described via the following equation:

$$L(\{p_i\},\{t_i\}) = \alpha\frac{1}{N}\sum_i L_{cls}(p_i,p_i^*) + \beta\frac{1}{N}\sum_i p_i^* L_{reg}(t_i,t_i^*), \tag{2.3}$$

where $i$ is the proposed box index number in each mini-batch, $p_i$ is the predicted probability of a proposed box, $t_i$ represents the height, width, centre x and y coordinates of a proposed box, $\alpha$ and $\beta$ are balancing weights for the objectness/classification and regression terms where $\alpha = 1$ and $\beta = 2$, $p_i^*$ signifies whether the proposed box has a positive or negative label, $t_i^*$ is the height, width, centre x and y coordinates of the ground truth box and $N$ is the number of anchors used for calculating this loss function in each mini-batch. In addition, $L_{cls}$ is objectness/classification loss and $p_i^* L_{reg}$ is box regression loss for only positive labelled boxes.

### 2.2.2 Galaxy cluster catalogue sample and image preprocessing

The WHL12 catalogue applied the friends-of-friends grouping algorithm (Huchra & Geller, 1982) on astrometry, photometry and distance data from the Sloan Digital Sky Survey Data Release 8 (SDSS-III DR8, Aihara et al. 2011) to detect clusters. They identified 132684 clusters in the redshift range of $0.05 \leq z < 0.8$, where the resultant catalogue used Monte Carlo simulations to obtain an estimated completeness of greater than 95 per cent for detecting clusters with mass greater than $1.0 \times 10^{14}$ $M_\odot$ inside an $r_{200}$ radius[9] and in the redshift range of $0.05 \leq z < 0.42$. We used Abell clusters that were identified by the WHL12 catalogue to obtain the labelled data needed to create a training set. We chose the Abell clusters because our technique uses visual inspection of images in a

[9] $r_{200}$ is the radius at which the mean density of the cluster is 200 times greater than the critical density of the Universe.

similar manner to that performed by George Abell and is therefore appropriate for this proof-of-concept work.

We did not train the Faster R-CNN algorithm on the entire WHL12 catalogue, as this was a pilot study to test the applicability of the Faster R-CNN algorithm at detecting clusters on a sample set. We utilised photometric redshift values that were estimated by the WHL12 catalogue as the photometric redshift of the Abell clusters. From which, we limited the photometric redshift range of clusters to $0.1 < z < 0.2$, as we wanted to maximise the signal-to-noise available and avoid nearby clusters that could fill the field-of-view. We set a $N_{200}$ threshold of more than 20 observed galaxy members inside $r_{200}$ to limit the number of poorly populated clusters, which may have a lower signal-to-noise. From applying these constraints, we obtained a sample set of 497 Abell clusters. We also utilised cluster richness values that were estimated by the WHL12 catalogue as the richness of the Abell clusters, where richness was originally defined by Wen et al. (2012) via the following equation:

$$R_{L^*} = \frac{L_{200}}{L^*}, \tag{2.4}$$

where $R_{L^*}$ is the cluster richness, $L_{200}$ is the total $r$-band luminosity of galaxy members within $r_{200}$ and $L^*$ is the typical luminosity of galaxies in the $r$-band.

We note that the brightest cluster galaxy (BCG) is a giant elliptical galaxy that is usually located in the vicinity of the spatial and kinematic centre of a cluster (Stott et al., 2008). We converted the right ascension (RA) and declination (Dec) of the BCGs, that had been determined by the WHL12 catalogue, to pixel coordinates. We adopted these pixel coordinates as the centre coordinates for the ground truth boxes. We also set a box size that had dimensions of approximately 250 kpc around the centre coordinates at each cluster's photometric redshift (i.e. box length and width of 500 kpc), where the box size dimensions roughly corresponded to the optical core radius of clusters (Girardi et al., 1995). We also restricted Dec to be greater than 0 degrees in order to reduce the amount of data near the galactic plane, where there is a higher concentration of stars that may introduce significant foreground contamination.

The imaging camera on the SDSS telescope has a pixel size scaling of 0.396 arcsec pixel$^{-1}$. The SDSS telescope utilises five broadband imaging photometry filters that are referred to as *u, g, r, i, z*. These photometry filters cover a wavelength range of approximately 3000 to 11000 Å (Fukugita et al., 1996). We decided to use the *g, r, i* filters but not the *u* and *z* filters, as the filter response of the SDSS telescope is poorer at these wavelengths. We fixed each image size to 2000 × 2000 pixels (approximately 1443 × 1443 kpc at redshift $z = 0.1$ and 2588 × 2588 kpc at redshift $z = 0.2$) to capture the wider context in an image. It should be noted that the Faster R-CNN algorithm also intrinsically lowers the resolution of input images to a user-defined fixed dimensionality of 1000 × 1000 pixels for computational efficiency.

To make our wide-field colour images, we first ensured that the images[10] taken from the publically available Sloan Digital Sky Survey Data Release 9 (SDSS-III DR9, Ahn et al. 2012) were set to the same scaling and aspect ratios. We then stacked the *g, r, i* filter images to RGB channels and applied a non-linear transformation to 'stretch' each image channel, where we experimented with linear, square root and logarithmic transformations. The transformations were responsible for adjusting the contrast of images within lower and upper flux limits after mapping the images onto an apparent brightness scale. These procedures helped to reduce background noise, dim extremely bright objects as well as make stars and galaxies more distinguishable. From which, we found that applying the square root function was better for visualizing galaxy structure and colour. This benefits the Faster R-CNN algorithm by decreasing the learning complexity of the features.

In Ren et al. (2015), it is stated that one of the properties of the Faster R-CNN algorithm is translational invariance, which means the algorithm should be robust at finding translated objects. Although, to ensure that the algorithm did not learn positional bias we exposed the algorithm to clusters that were situated at random image positions. Firstly, we applied a uniform random offset to the coordinates of clusters which resulted in an even spread of cluster positions across all images, where a cluster could be found anywhere within 270 arcseconds of the

[10]The imaging data for SDSS-III DR9 can be found via NASA's SkyView (`http://skyview.gsfc.nasa.gov`) online database (McGlynn et al., 1998).

x and y planes from the image centre. We applied this random offset to images in the sample set an additional three times, which increased the size of the sample set to 1988. This also introduced additional negative boxes for the algorithm to learn since each image contained a slightly different background. Secondly, we allowed the algorithm to horizontally flip images during training, where each image had a 50 per cent chance of being flipped. This approach can double the size of the training sample if all images are flipped once but this approach does not affect the size of the testing sample. Since clusters can be observed from any orientation in an image, we found these augmentation techniques to be appropriate.

We performed hold-out validation on the sample set to create a training set and a test set, which are approximated representations of the full population. The training set was made up of ∼90 per cent of clusters (i.e. 1784 clusters) from the sample set and the test set was made up of the remaining ∼10 per cent (i.e. 204 clusters). In Figure 2.2, we show the astronomical coordinates for clusters in the training and test sets as well as the astronomical coordinates for all the clusters in the WHL12 catalogue. In Figure 2.3, we show distributions of the photometric redshift, $r$-band magnitudes for BCGs and richness of clusters in the training set. Lastly, in Figure 2.4, we show the distribution of cluster positions in images from the training and test sets.

## 2.3 Results

### 2.3.1 Model analysis with test set

We trained our model with graphics processing unit (GPU) support for a maximum of 25000 steps to ensure the algorithm had enough training time to sufficiently learn to minimise prediction errors. We note that the number of steps is a tunable hyper-parameter that can shorten or extend the run-time of model training. In Figure 2.5, we found that the algorithm generalised well as the total loss stabilised at approximately 3000 steps, where a step represents one iteration per mini-batch size of one from the dataset through the algorithm. For a competent model, the total loss should not fluctuate significantly during training.

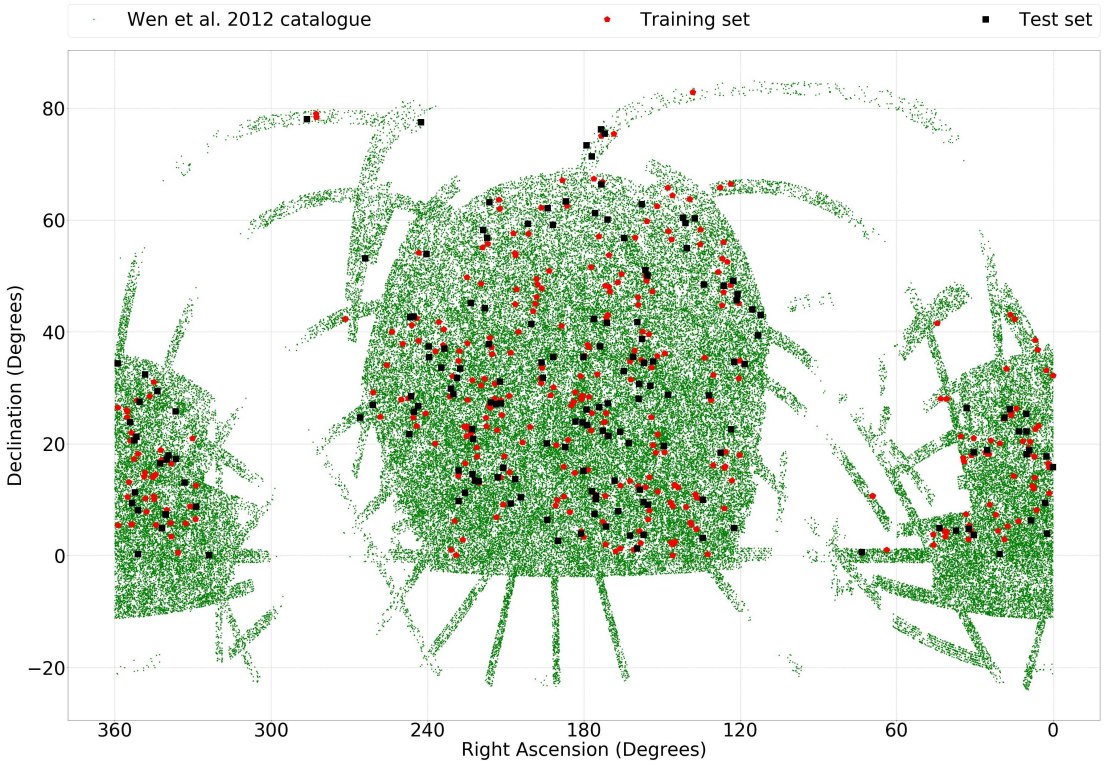


**Figure 2.2:** This figure displays a map of astronomical coordinates using the J2000 epoch system of clusters in the training set (red pentagons), test set (black squares) and full WHL12 catalogue (green circles).

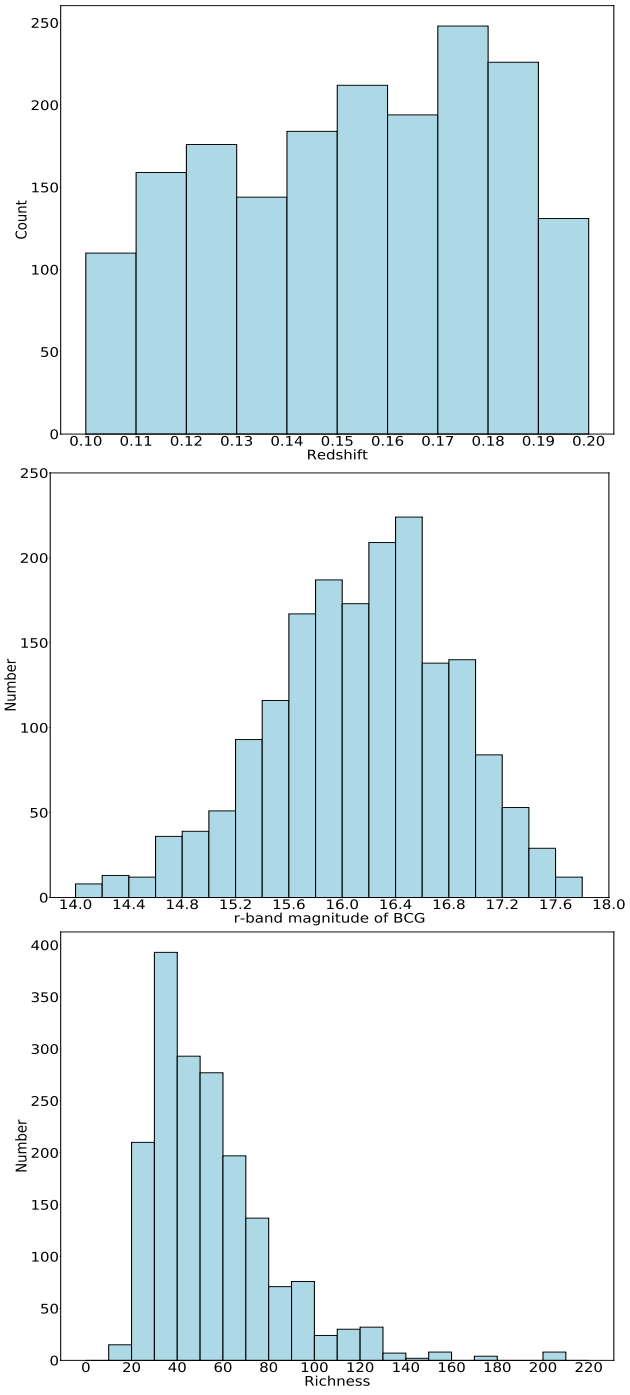


**Figure 2.3:** This figure displays the distributions of properties for clusters in the training set. This includes the photometric redshift, $r$-band magnitude of the BCG and richness (from top to bottom row respectively).

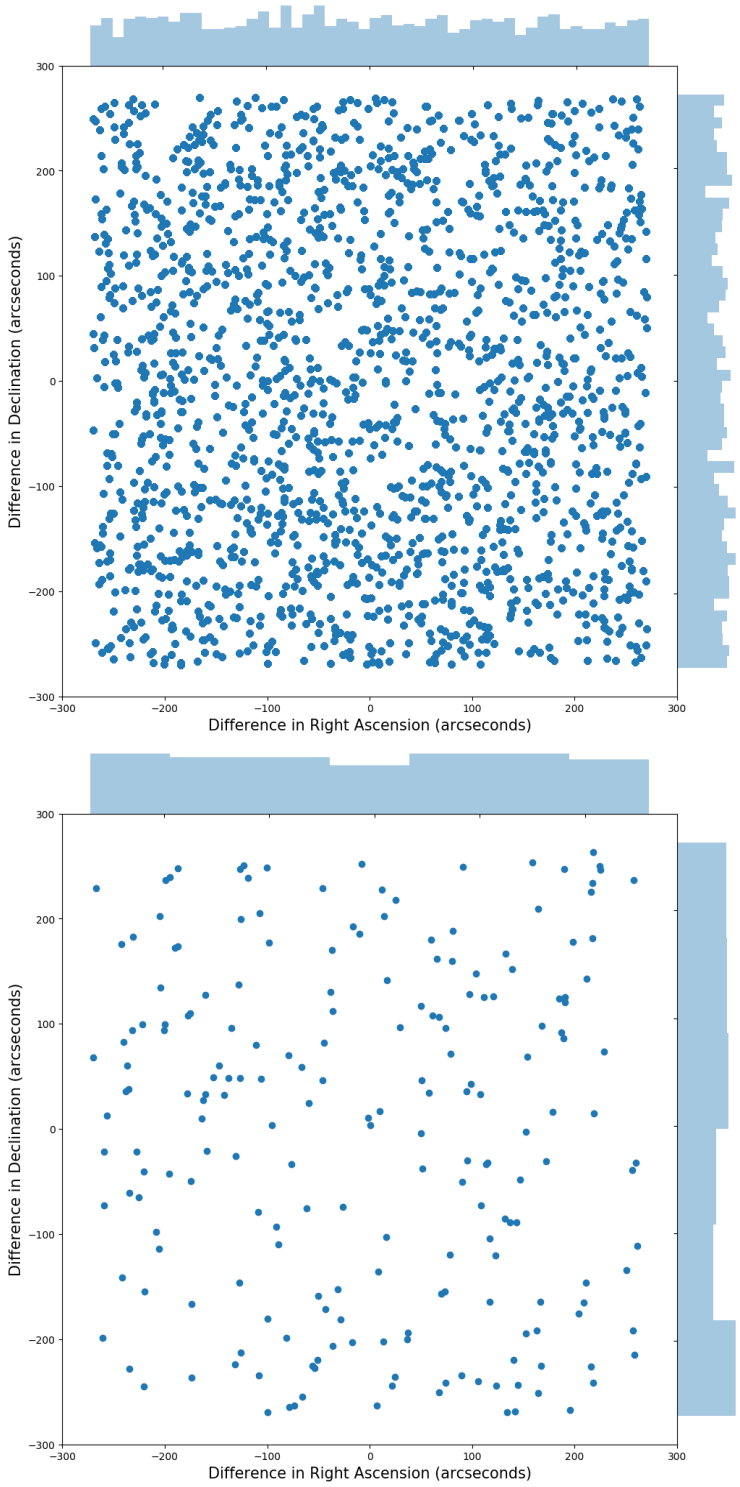


**Figure 2.4:** This figure displays the distribution of cluster positions in images from the training set (top row) and test set (bottom row). The points were determined by calculating the difference in arcseconds between the uniform random offset and the true coordinates of the clusters at their respective photometric redshift.

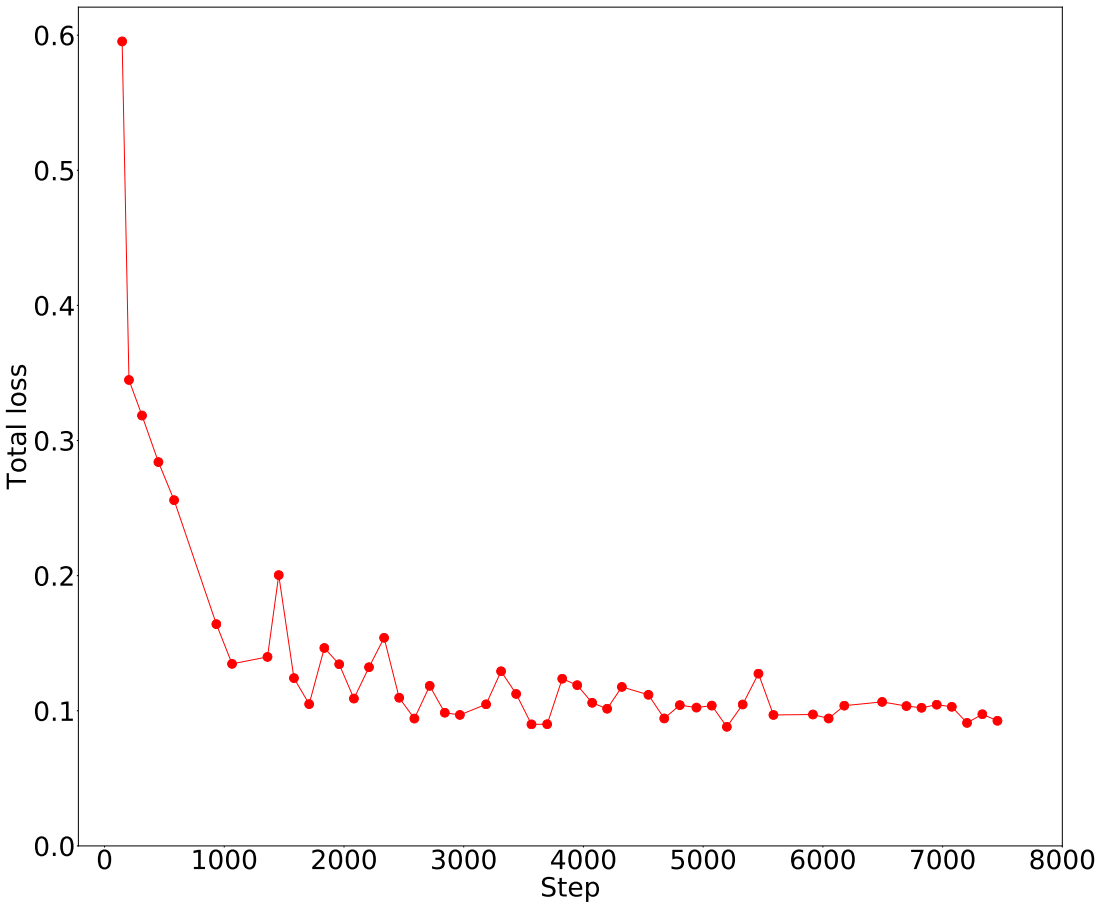


**Figure 2.5:** This figure displays the total loss (see Equation 2.3) which considers the objectness/classification and localisation errors of clusters in the test set during training. We stopped model training after 7458 steps since the total loss appeared to stabilise, where each point represents the total loss recorded at different step intervals. The values of these points can be found in Table A1.

We could also examine the training performance of the RPN and DN via their respective loss functions, where we measured the objectness and box regression loss in the RPN as well as the classification and box regression loss in the DN. Objectness/classification loss measured whether boxes were likely to contain ground truth objects whilst box regression loss measured the exactness of the dimensions between the positive labelled boxes and ground truth boxes. We note that smaller loss values indicated that prediction values were becoming more similar to the ground truth values. In Figure 2.6, we found that each of the losses also eventually stabilised at approximately 3000 steps, where the RPN appeared to be better generalised than the DN since there were fewer fluctuations in the losses at higher steps, in particular for the objectness/classification losses. This could be explained by two possible reasons: either the training was suited for identifying generic 'cluster-like' structures in images but improvements could be made to enhance the final classification of clusters, or we needed to allow our model to train for more steps.

We note that the Abell clusters in the test set had appeared in the training set too but since the image positions of the clusters and their surrounding image environments had been shifted, we assumed the images to be somewhat unique. This would be a useful test of the localisation performance of our model.

To evaluate our model we used common metrics such as precision, recall and F1 score (Goutte & Gaussier, 2005). Although, we did not measure the model performance on true negatives, since there were many other astronomical objects aside from clusters to consider in the images. We expected that finding a cluster would be relatively rare due to stars, galaxies and image artifacts populating the field-of-view. From which, our approach only searched for clusters rather than attempting to classify all objects, such that we treated any other object as non-clusters. The final output of our model was a 'confidence' score that was generated for every proposed box, where a high confidence score meant a high probability of an object being a 'real' cluster. We aimed to determine a threshold for the confidence score that returned high precision and high recall ratios. This involved running our model on the test set using different confidence score thresholds to examine the number of true positives (TP), false positives (FP) and false negatives (FN) returned.

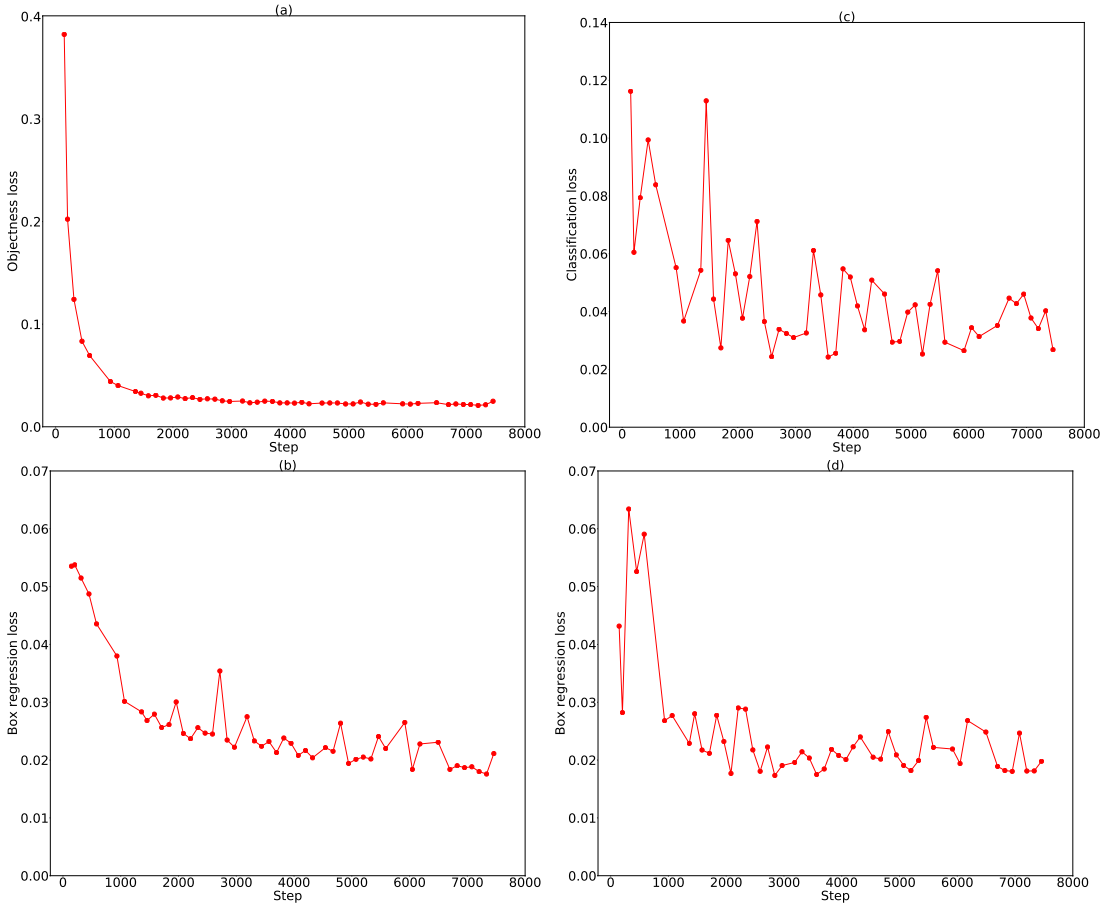


**Figure 2.6:** This figure displays training losses of the RPN and the DN are represented in (a), (b), (c) and (d). Where (a) displays the RPN objectness loss, (b) displays the RPN box regression loss, (c) displays the DN classification loss and (d) displays the DN box regression loss. The training of the model was stopped after 7458 steps when the total loss had minimal fluctuations, see Figure 2.5. The value for each point in (a),(b),(c) and (d) can be found in Table A1.

We also aimed to define a distance threshold by calculating a linear distance between the predicted and ground truth centre coordinates, where the predicted cluster centre was assumed to be at the same redshift as the ground truth cluster centre. We would only apply this distance threshold during the model analysis, as we wanted to distinguish whether a predicted object should be considered as a TP or FP detection in order to assess classification and localisation errors.

We note that TPs refers to the number of proposed boxes that score greater than the confidence score threshold and has a predicted centre within the distance threshold of the ground truth centre. FPs refers to the number of proposed boxes that score greater than the confidence score threshold but does not have a predicted centre within the distance threshold of the ground truth centre. FNs refers to the number of ground truth clusters without a successful prediction above the confidence score threshold within the distance threshold.

We calculated the precision and recall ratios using the number of TPs, FPs and FNs for each confidence score threshold. Precision (also known as purity) is a ratio that considers the total number of ground truth objects returned by our model when compared with the total number of predictions, where precision is described via the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.5}$$

Recall (also known as completeness) considers the total number of ground truth objects returned by our model when compared with the total original number of ground truth objects, where recall is described via the following equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.6}$$

Precision-Recall (PR) curves are typically used as visual tools to examine the performance of a model, especially when a class population imbalance exists in a dataset (Davis & Goadrich, 2006). We note that each point on a PR curve refers to the precision and recall ratio at a specific cut-off threshold. We explored eleven cut-off thresholds for confidence scores ranging from 0 to 100 per cent to calculate the corresponding F1 score at each confidence score threshold.

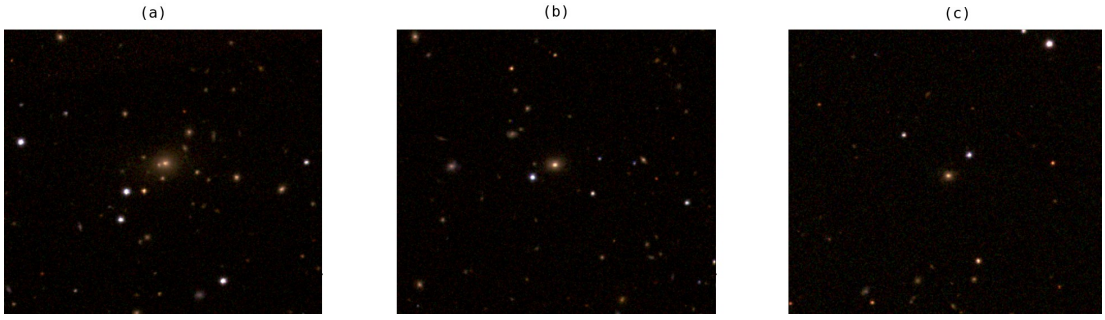


**Figure 2.7:** This figure displays colour images (a), (b) and (c) of three different Abell clusters from the test set. The J2000 coordinates for each cluster are as follows: (a) RA: 222.78917 and Dec: 14.61203, (b) RA: 180.19902 and Dec: 35.58229 and (c) RA: 137.49464 and Dec: 60.32841. The predicted confidence scores and properties for the clusters in (a), (b) and (c) can be found in Table 2.1.

F1 score is the harmonic mean between the precision and recall ratios at each confidence score threshold (Chase Lipton et al., 2014). We aimed to maximise the F1 score for our model in order to find the optimal balance between precision and recall, where F1 score is described via the following equation:

$$\text{F1 Score} = 2\left(\frac{\text{Precision x Recall}}{\text{Precision } + \text{ Recall}}\right) \tag{2.7}$$

We also analysed three individual clusters (a), (b) and (c) in the test set that had contrasting predicted confidence scores. The colour images of these three clusters can be seen in Figure 2.7. From Table 2.1, we found that (c) had the lowest confidence score, whilst (a) had the highest. This was likely because even though (c) had a high richness value, it was at a higher redshift which meant its galaxies would appear fainter. Additionally (b) had a lower richness than (c) but it was estimated to have a higher confidence score, this was likely because it was at a lower redshift. This indicated that clusters at a lower redshift with brighter galaxies were more likely to receive higher confidence scores than their fainter counterparts at higher redshift. However, it seemed that a cluster would also need to have high richness in order to achieve a very high confidence score. As a demonstration of our model, we chose a confidence score threshold of 80 per cent for the remainder of this subsection.

Furthermore, we investigated how the environment (actual or contaminants)

| ID | Confidence Score (%) | Photometric Redshift | $r$-band Magnitude of the BCG | Richness |
|---|---|---|---|---|
| (a) | 98 | 0.1474 | 14.99 | 78.49 |
| (b) | 51 | 0.1303 | 16.19 | 35.90 |
| (c) | 20 | 0.1875 | 16.73 | 58.44 |

**Table 2.1:** This table displays the predicted confidence scores and properties of each cluster in Figure 2.7.

surrounding a cluster in an image can influence the detections by our model. We visually inspected multiple high-scoring candidate clusters from four different images in the test set. In Figure 2.8, we observed a predicted candidate cluster that lies at a linear distance of 88 kpc from the ground truth centre, where we assumed that the predicted galaxy was at the same redshift as the ground truth galaxy when calculating the linear distance. We observed that multiple possible candidate galaxies could be classified as the BCG of the cluster. Since we trained our model to predict a BCG as the cluster centre, we would expect one of the galaxies in Figure 2.8 to be chosen. However, we found that our model was unable to definitively determine the ground truth cluster centre if there were multiple BCG-like galaxies close together within an image, such as in the event of an on-going cluster merger. Instead our model found an average position of the galaxies, which is likely more appropriate for such systems. Wen & Han (2013) developed an approach to determine the dynamical state of a cluster based on astrometry, photometry and distance data to quantify a relaxation parameter 'Γ', which considered factors such as morphological asymmetry, ridge flatness and normalised deviation. They defined the relaxation parameter for $\Gamma \geq 0$ as representing a relaxed cluster state and $\Gamma < 0$ as an unrelaxed cluster state, such that a more positive/negative value would imply a more relaxed/unrelaxed dynamical state respectively. We cross-matched the cluster in Figure 2.8 with Wen & Han (2013), which suggested that the cluster was in a very unrelaxed dynamical state with a relaxation parameter value of $\Gamma = -1.43 \pm 0.08$ indicating a possible on-going merger.

In Figure 2.9, we found another predicted candidate cluster that lies at a linear distance of 158 kpc from the ground truth centre. We used spectroscopic measure-

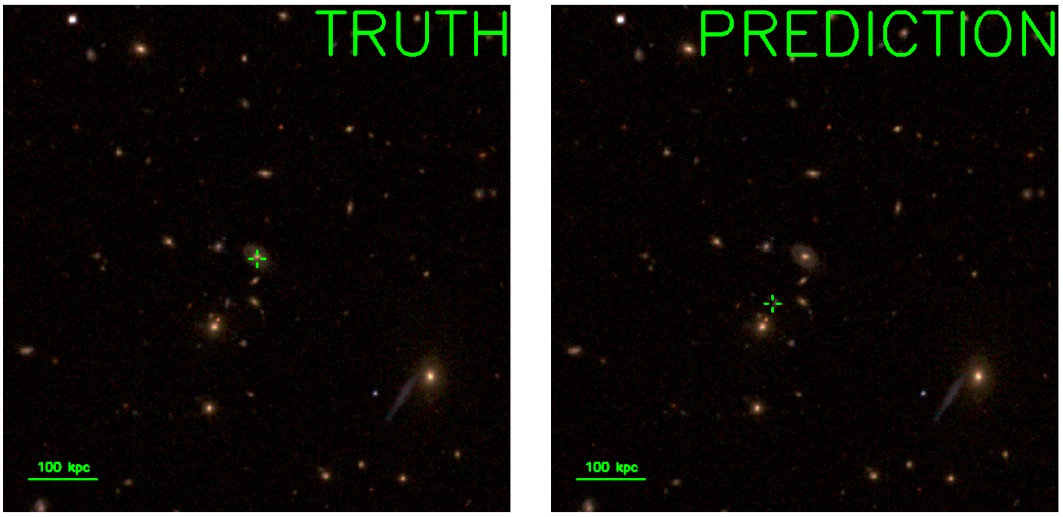


**Figure 2.8:** This figure displays the ground truth and predicted centre coordinates, where there is a linear separation of 88 kpc with respect to the photometric redshift of $z = 0.1788$ for the ground truth cluster. The J2000 coordinates of the ground truth cluster is RA: 191.85623 and Dec: 35.54509.

ments from SDSS/Baryon Oscillation Spectroscopic Survey (BOSS, Eisenstein et al. 2011) to determine the spectroscopic redshift of the galaxies. We identified the spectroscopic redshift of the predicted 'BCG' as $z = 0.15765 \pm 0.00003$ and the spectroscopic redshift of the ground truth BCG as $z = 0.19282 \pm 0.00003$. This meant that while the galaxies are in the same line-of-sight, they are likely not part of the same gravitationally-bound system. From which, our model was unable to determine the ground truth BCG since the predicted BCG-like galaxy had stronger visual features and was at a lower redshift.

We found that our model identified two BCG-like galaxies in Figure 2.10 as potentially two separate cluster centres that were within each other's respective optical core radius. One of the two candidate objects had a predicted centre coordinate that was nearby to the ground truth centre whilst the other object was further at a linear distance of 220 kpc away. We found the spectroscopic redshift of the predicted 'BCG' to be $z = 0.16209 \pm 0.00002$ whilst the ground truth BCG had a spectroscopic redshift of $z = 0.15990 \pm 0.00002$. We cross-matched this cluster with Wen & Han (2013), which suggested the cluster had a relaxation parameter value of $\Gamma = -0.61 \pm 0.09$. This value implied the cluster

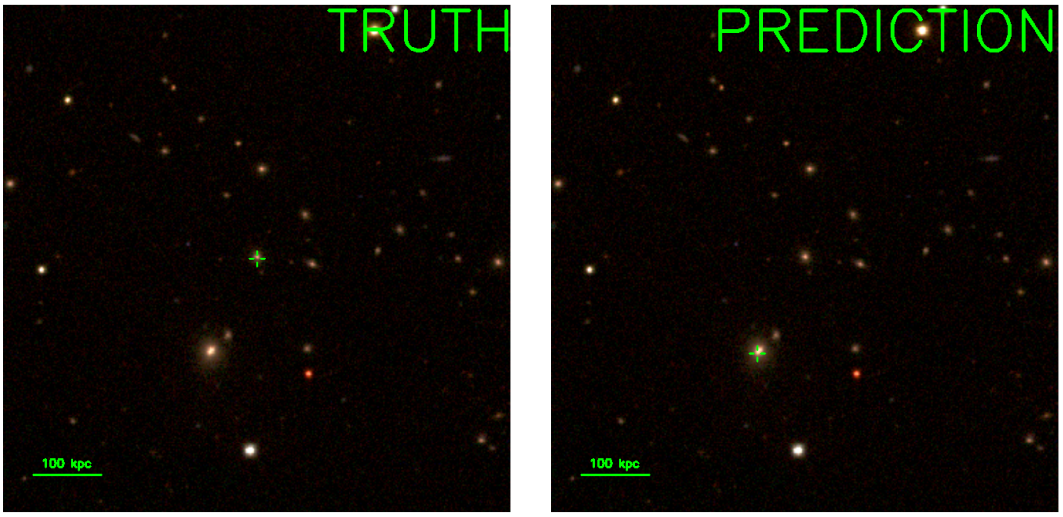


**Figure 2.9:** This figure displays the ground truth and predicted centre coordinates, where there is a linear separation of 158 kpc with respect to the photometric redshift of $z = 0.1618$ for the ground truth cluster. The J2000 coordinates of the ground truth cluster is RA: 161.20657 and Dec: 35.54042.

was in an unrelaxed dynamical state but to a lesser extent than the cluster seen in Figure 2.8, such that this cluster was not experiencing an on-going merger but was likely in a possible pre-merger or post-merger state.

In Figure 2.11, we found that our model was able to detect a cluster candidate which was far from the ground truth centre at a linear distance of 1163 kpc. We noticed that the predicted centre was located on top of a BCG-like object. We found the spectroscopic redshift of the predicted 'BCG' to be $z = 0.10608 \pm 0.00002$ and the ground truth BCG spectroscopic redshift to be $z = 0.14551 \pm 0.00003$. This indicated that the two objects were clearly physically unrelated, therefore we identified the candidate object as a separate cluster. Using the NASA/IPAC Extragalactic Database[11], we cross-matched this cluster candidate as 'MSPM 08519' with a heliocentric redshift of $z = 0.1055$ in the Smith et al. (2012) galaxy groups and clusters catalogue.

We decided that an appropriate distance threshold value would be between 88 and 158 kpc based on the predictions in Figures 2.8 and 2.9. We noticed at

[11]The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

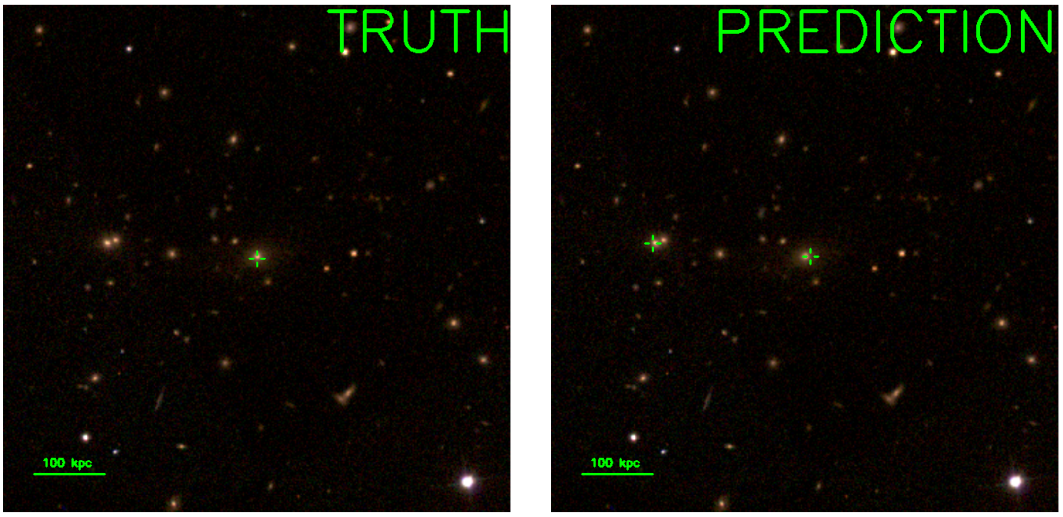


**Figure 2.10:** This figure displays the ground truth and predicted centre coordinates (the one that does not overlap directly with the ground truth), where there is a linear separation of 220 kpc with respect to the photometric redshift of $z = 0.1603$ for the ground truth cluster. The J2000 coordinates of the ground truth cluster is RA: 353.35867 and Dec: 9.42395.

these distances, the detections were visually far enough apart to cleanly constrain probable individual clusters from potential cluster mergers and line-of-sight overlapping clusters. This helped us to differentiate between cases of TP and FP detections in our model analysis. From which, we selected a distance threshold of 100 kpc for the remainder of this work.

For example, in Figures 2.10 and 2.11 we observed secondary detections made by our model. We would categorise these as FP detections, since the detections did not satisfy our set criteria for TPs. However, we note that FPs could also be actual clusters, such as is the case for Figure 2.11. This implied that FPs may consist of candidate clusters that need further verification or require cross-matching to existing cluster catalogues, where it would not be inappropriate to label these FP detections as new cluster candidates. Although, in Figure 2.9, we found that no successful detections were made on the ground truth cluster within the distance threshold. We would categorise this ground truth cluster as a FN whereas the detection would be categorised as a FP. This suggested that FNs may consist of clusters that do not appear to have an obvious overdensity of galaxies


**Figure 2.11:** This figure displays the ground truth and predicted centre coordinates (the one that does not overlap directly with the ground truth), where there is a linear separation of 1163 kpc with respect to the photometric redshift of $z = 0.1368$ for the ground truth cluster. The J2000 coordinates of the ground truth cluster is RA: 186.96341 and Dec: 63.38483.

| Confidence score threshold (%) | # TP | # FP | # FN | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 0 | 203 | 60997 | 1 | 0.003317 | 0.9951 | 0.006612 |
| 10 | 198 | 538 | 6 | 0.2690 | 0.9706 | 0.4213 |
| 20 | 197 | 391 | 7 | 0.3350 | 0.9657 | 0.4975 |
| 30 | 193 | 302 | 11 | 0.3899 | 0.9461 | 0.5522 |
| 40 | 191 | 243 | 13 | 0.4401 | 0.9363 | 0.5987 |
| 50 | 188 | 202 | 16 | 0.4821 | 0.9216 | 0.6330 |
| 60 | 181 | 163 | 23 | 0.5262 | 0.8873 | 0.6606 |
| 70 | 177 | 119 | 27 | 0.5980 | 0.8676 | 0.7080 |
| 80 | 165 | 72 | 39 | 0.6962 | 0.8088 | 0.7483 |
| 90 | 136 | 29 | 68 | 0.8242 | 0.6667 | 0.7371 |
| 100 | 0 | 0 | 204 | 0.00 | 0.00 | 0.00 |

**Table 2.2:** This table displays the total number of TPs, FPs and FNs returned by our model on images from the test set, where the precision, recall and F1 score ratios were calculated for each confidence score threshold.

and do not contain distinctive BCG-like galaxies. These findings contribute to how distinguishable clusters would appear at the output of the Faster R-CNN algorithm.

In Figure 2.12, we observed that high precision diminishes recall and high recall diminishes precision, where a low precision ratio would result in a large number of unverified candidate objects whilst a low recall ratio would result in many known clusters not being detected. From Table 2.2, we noticed that an 80 per cent confidence score threshold had the greatest F1 score, which indicated that this confidence score threshold was the most effective at balancing precision and recall for cluster detections in the test set.

We subsequently analysed the linear distance between all of the predicted centre coordinates from the ground truth centre using an 80 per cent confidence score threshold. We note that when we were initially deciding on an appropriate distance threshold value, we disregarded any detection further than 250 kpc from the ground truth centre from being considered as a TP prediction, since the detection would lie outside the optical core radius. In Figure 2.13, we found that

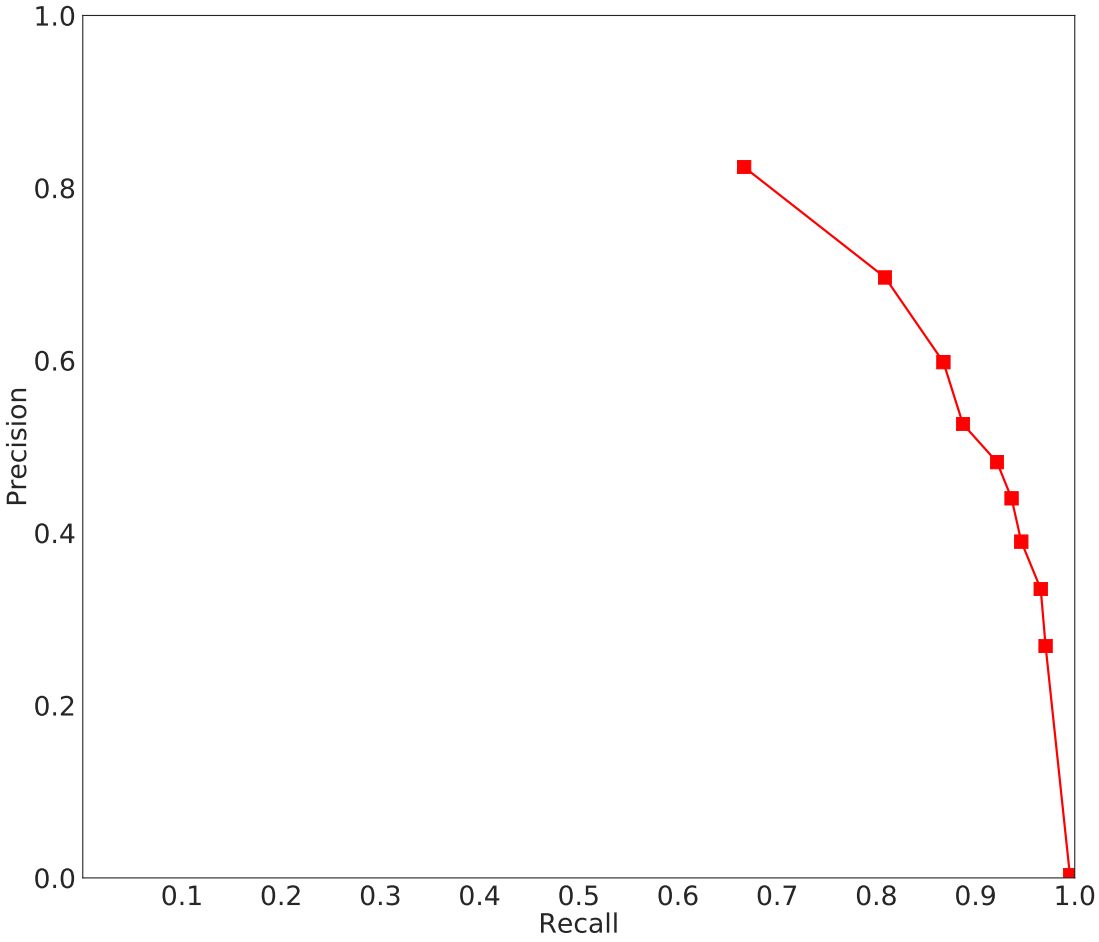


**Figure 2.12:** This figure displays the precision versus recall ratios from the test set, where each point represents the ratios at different confidence score thresholds. The values of each point can be found in Table 2.2. We did not include the precision and recall ratio for the 100 per cent confidence score threshold, as it provided no conclusive evaluation of the model performance.

only a minority of ground truth clusters had predicted centre coordinates near to 250 kpc, whereas a distance threshold of 100 kpc contained 70 per cent of all of the predictions and returned 81 per cent of the total ground truth clusters in the test set when using an 80 per cent confidence score threshold.

We calculated the standard error of regression (i.e. root mean squared error) for position estimates by our model to determine the average difference between the predicted and ground truth centre coordinates, where the standard error of regression is described via the following equation:

$$\sigma_{estimate} = \sqrt{\frac{\sum (Y - Y')^2}{N}}, \tag{2.8}$$

where $\sigma_{estimate}$ is the standard error of regression, $Y$ is the ground truth value, $Y^{'}$ is the predicted value and $N$ is the number of estimates (L. McHugh, 2008). We obtained a $\sigma_{estimate}$ value of 17.42 kpc for TP predictions in the test set. We could also estimate a 95 per cent confidence interval for all predicted centre coordinates to be approximately within $\pm 1.96 \times \frac{\sigma_{estimate}}{\sqrt{N}}$ of ground truth centre coordinates (Altman & Bland, 2005).

In Figure 2.14, we visually compared the positions of detected ground truth clusters from the test set with their original locations in Figure 2.4. We found that our model did not show bias towards any particular location in an image. This suggested that the uniform random offset in §§2.2.2 was effective at constraining positional bias.

In Figure 2.15, we compared the photometric redshift, $r$-band magnitude of the BCG and richness distributions of detected ground truth clusters with their original distributions from the test set. We aimed to determine whether our model exhibited bias towards any of these properties. From which, we performed a two sample Kolmogorov-Smirnov (KS) test (Smirnov, 1939) to test whether the original and detected distributions violated the null hypothesis. Since the KS test is non-parametric, the distributions did not need to have normality. We calculated test statistic values of 0.06275, 0.07335 and 0.02193 for the photometric redshift, $r$-band magnitude of the BCG and richness distributions respectively. We set $\alpha = 0.05$ as the level of significance to obtain a critical value of 0.1281 (see Equation 3.1 in Gail & Green (1976)). Since the test statistic values were

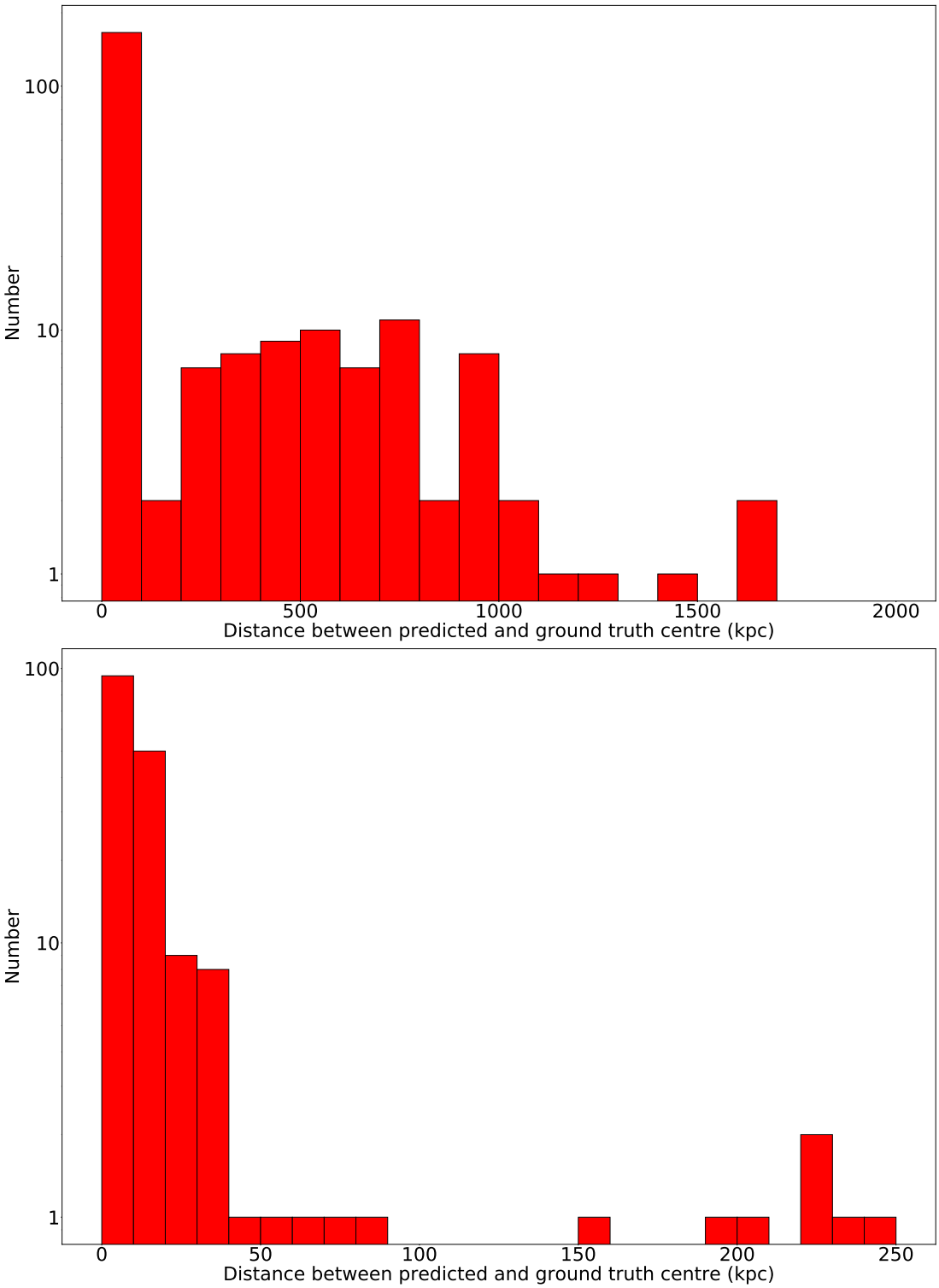


**Figure 2.13:** This figure displays the distribution of the linear distance between predicted and ground truth centre coordinates in test set images when using an 80 per cent confidence score threshold for all predictions (top row) and all predictions within the distance threshold (bottom row).

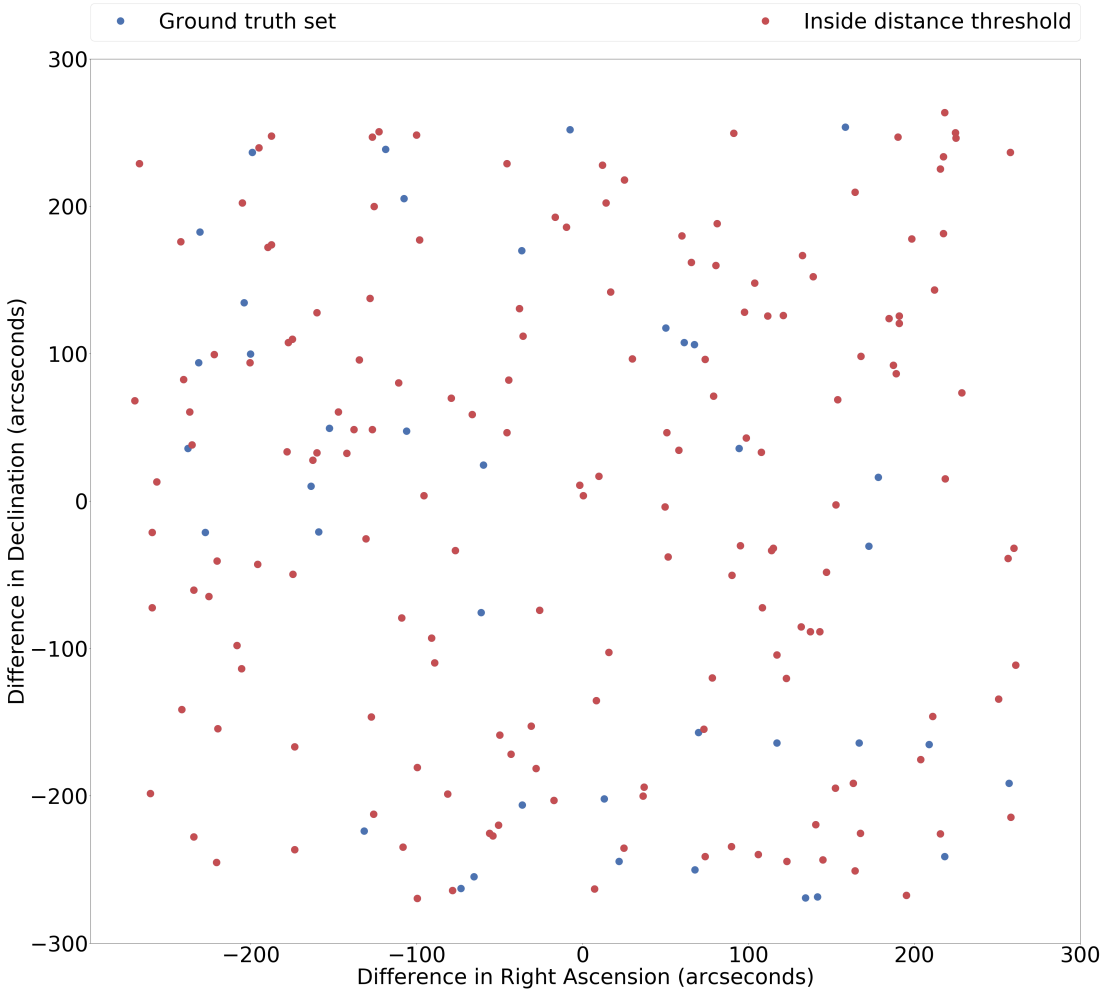


**Figure 2.14:** This figure displays a comparison of the centre coordinate offset between detected ground truth clusters (red circles) and the original list of uniform random offset values (blue circles) of ground truth clusters in the test set from Figure 2.4.

smaller than the critical value at $\alpha = 0.05$, we cannot reject that the original and returned distributions were statistically the same.

### 2.3.2 Comparison to redMaPPer galaxy clusters

The redMaPPer catalogue detected clusters by using the red-sequence fitting technique on the photometric data of individual galaxies to search for a distinctive red-sequence with their nearby line-of-sight galaxies in colour-magnitude space. Rykoff et al. (2014) also applied their method to galaxies in SDSS-III DR8, to create a catalogue of $\sim$25,000 candidate clusters in the redshift range of $0.08 < z < 0.55$. We reapplied the same testing constraints as used in §§2.2.2 on the redMaPPer clusters, where these clusters must be in the photometric redshift range of $0.1 < z < 0.2$. We did not need to apply an observed galaxy count constraint to the redMaPPer clusters since the redMaPPer catalogue by default only contained clusters with at least 20 member galaxies[12]. In Figure 2.16, we locate a 105 square degree region that contains 31 clusters identified by the redMaPPer algorithm. We note that this region did not contain any clusters from our training or test sets in the photometric redshift range of $0.1 < z < 0.2$. We used these redMaPPer clusters to create a redMaPPer test set for examining the localisation and classification performance of our model on unseen clusters.

We adopted the same procedure from §§2.2.2 to generate wide-field colour images of redMaPPer clusters. We subsequently applied our model on these images as well as reutilising the precision, recall and F1 score evaluation metrics. In Figure 2.17, we observed a precision and recall trade-off that is similar to Figure 2.12, where precision increased with larger confidence score thresholds whilst recall decreased. From Table 2.3, we found that a 70 per cent confidence score threshold had the greatest F1 score. This suggested that our model had not overfit since it performed better on an unseen dataset, where this confidence score threshold was smaller than the 80 per cent confidence score threshold in §§2.3.1. We note that using a smaller optimal confidence score threshold would increase the number of detected objects whilst still retaining high precision and recall.

[12]We note that Rykoff et al. (2014) did not define galaxy members within $r_{200}$ but instead from an optical radius cutoff that scaled with the number of galaxies found via percolation.

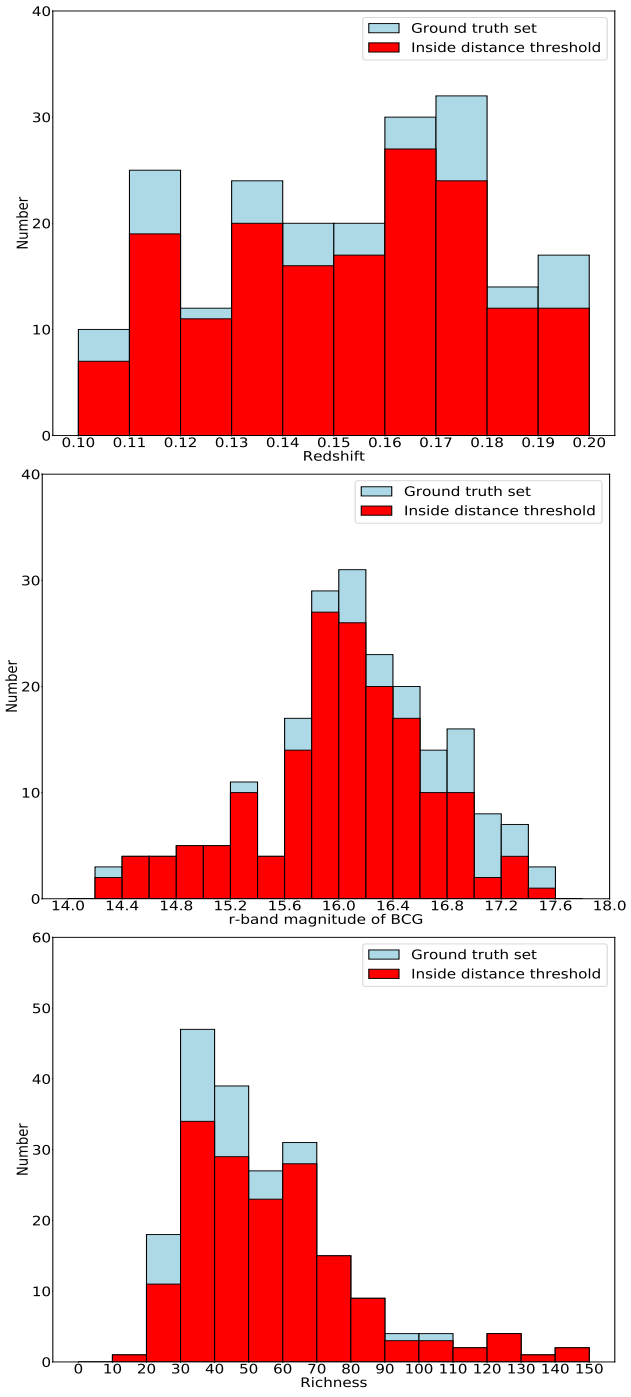


**Figure 2.15:** This figure displays the distributions of properties from the original (blue fill) and detected ground truth clusters (red fill) in the test set when using an 80 per cent confidence score threshold. In particular, the histograms present the photometric redshift, $r$-band magnitude of the BCG and richness of clusters (from top to bottom row respectively).

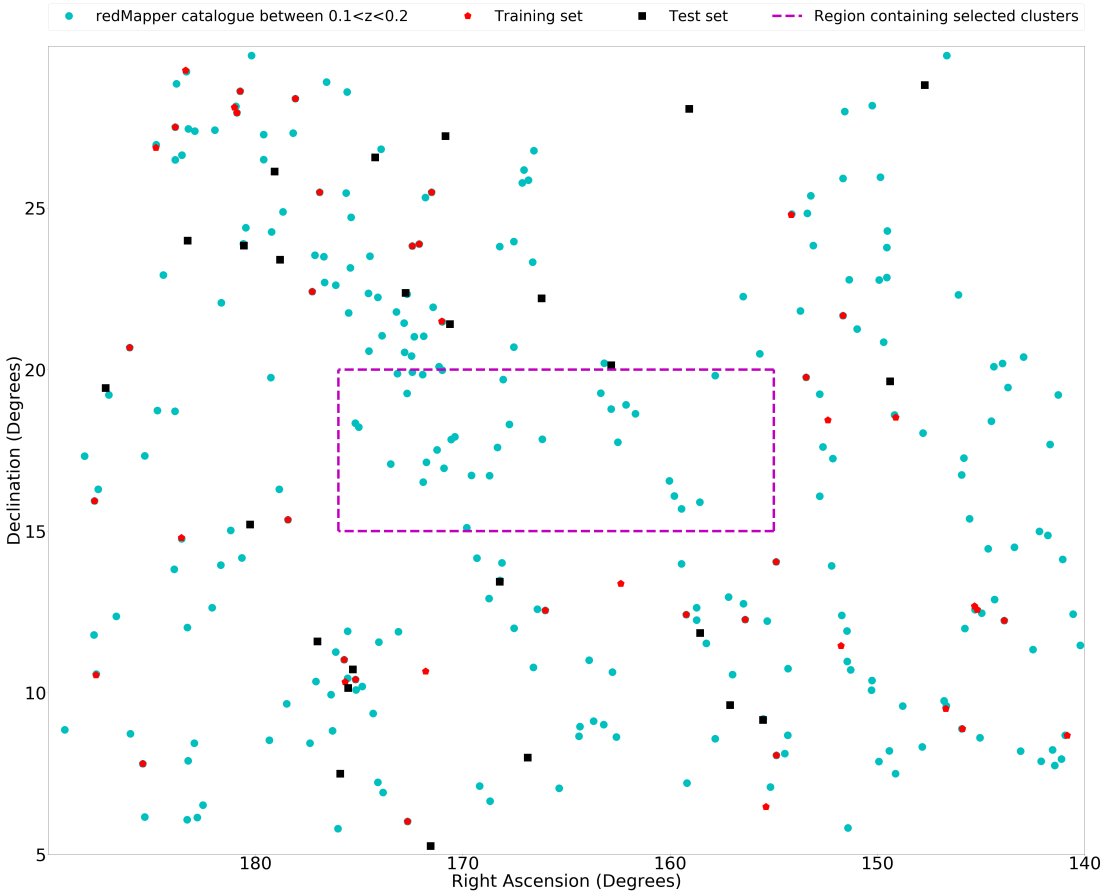


**Figure 2.16:** This figure displays a map of astronomical coordinates using the J2000 epoch system for clusters in the training set (red pentagons), test set (black squares) and redshift filtered redMaPPer catalogue (cyan circles). We highlight the region (purple dashed lines) of clusters in the redMaPPer test set, which were not already part of the training set or test set.

| Confidence score threshold (%) | # TP | # FP | # FN | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 0 | 30 | 9270 | 1 | 0.003226 | 0.9677 | 0.006430 |
| 10 | 29 | 50 | 2 | 0.3671 | 0.9355 | 0.5273 |
| 20 | 29 | 29 | 2 | 0.50 | 0.9355 | 0.6517 |
| 30 | 29 | 23 | 2 | 0.5577 | 0.9355 | 0.6988 |
| 40 | 28 | 18 | 3 | 0.6087 | 0.9032 | 0.7273 |
| 50 | 28 | 14 | 3 | 0.6667 | 0.9032 | 0.7671 |
| 60 | 28 | 11 | 3 | 0.7179 | 0.9032 | 0.80 |
| 70 | 27 | 8 | 4 | 0.7714 | 0.8710 | 0.8182 |
| 80 | 23 | 4 | 8 | 0.8519 | 0.7419 | 0.7931 |
| 90 | 17 | 0 | 14 | 1.00 | 0.5484 | 0.7083 |
| 100 | 0 | 0 | 31 | 0.00 | 0.00 | 0.00 |

**Table 2.3:** This table displays the total number of TPs, FPs and FNs returned by our model on images from the redMaPPer test set, where the precision, recall and F1 score ratios were calculated for each confidence score threshold.

When using a 70 per cent confidence score threshold, we obtained a $\sigma_{estimate}$ value of 12.33 kpc for TP detections in the redMaPPer test set.

We attempted to follow-up the FP detections from the redMaPPer test set images by cross-matching the FP detections with the redMaPPer and WHL12 catalogues. For this task, we used the 70 per cent confidence score threshold, which had eight FP detections as seen in Table 2.3. To perform the cross-matching, we applied a wider photometric redshift constraint range of $0.05 < z < 0.4$ to clusters in the redMaPPer and WHL12 catalogues. We also utilised a $\sim$1.61 arcminute radius for the predicted RA and Dec coordinates, which corresponded with the optical core radius of 250 kpc for a cluster at redshift $z = 0.15$, when searching through the catalogues. From which, we found that none of the FP detections matched with existing clusters in the redMaPPer catalogue. Although, we identified three FP detections as existing clusters in the WHL12 catalogue. This indicated that our model was capable of detecting clusters that had not been drawn from the same sample as the training set. The remaining five FP detections did not match with any known clusters in the redMaPPer and WHL12

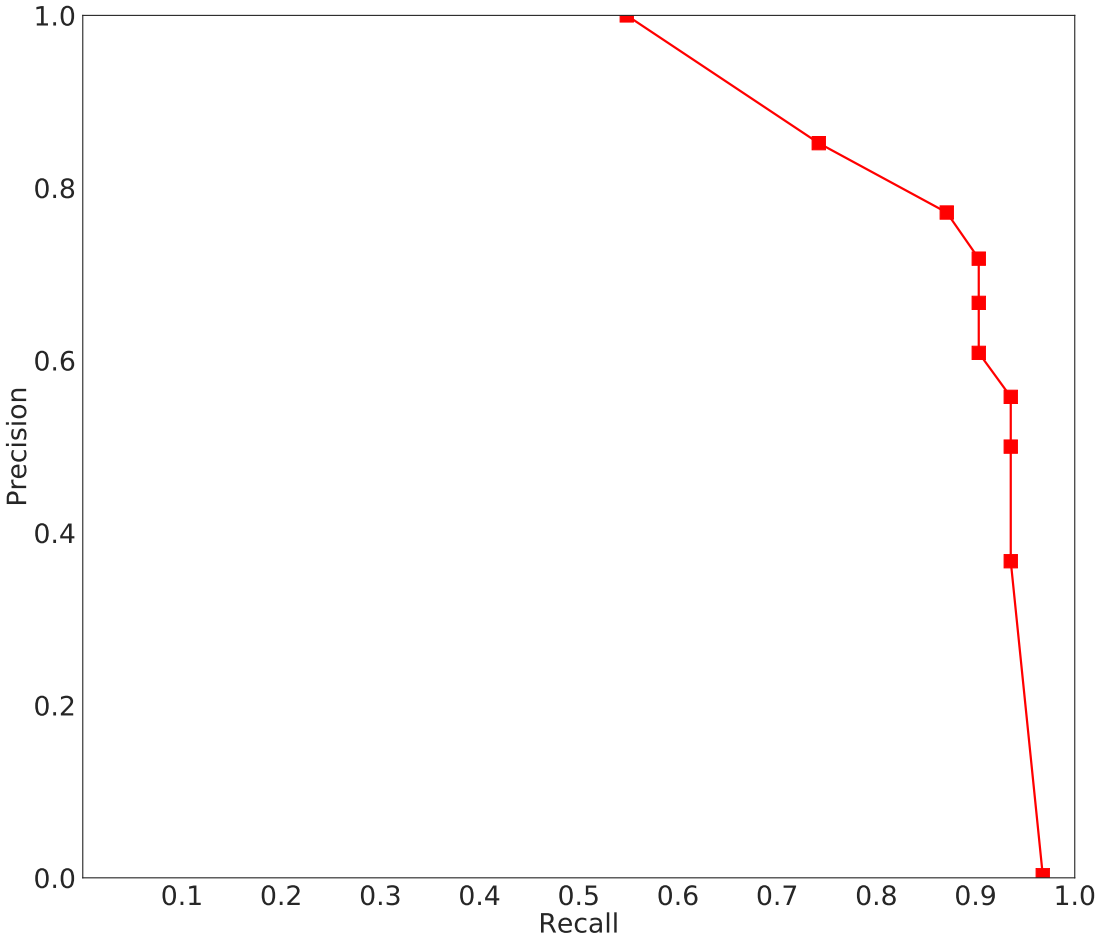


**Figure 2.17:** This figure displays the precision versus recall ratios from the redMaPPer test set, where each point represents the ratios at different confidence score thresholds. The values of each point can be found in Table 2.3. Similar to Figure 2.12, we did not include the precision and recall ratios for the 100 per cent confidence score threshold, as it again provided no conclusive evaluation of the model performance.

catalogues but these may exist as known clusters in other cluster catalogues.

## 2.4 Discussion

### 2.4.1 Limitations of our model

Feature engineering is a key process for increasing computational efficiency as well as aid in improving the predictive performance of a deep learning model. In this work, we applied constraints to the training set to reduce the complexity of visual features. For example, we used Abell clusters which intrinsically contained a minimum of 50 galaxies within a 1.5 $h^{-1}$ Mpc radius of the cluster centre (Abell, 1958). This meant that these clusters would have strong signal-to-noise and were likely to be real gravitationally bound clusters. However, not all Abell clusters had been verified as actual clusters, so one limitation of our approach is that our model is reliant on Abell clusters that were cross-matched with clusters in the WHL12 catalogue for training data. We note that Wen et al. (2012) used Monte Carlo simulations to determine a false detection rate of less than 6 per cent for the entire WHL12 catalogue. This is important for our model, where a higher false detection rate would lower the overall predictive power of our model because the training data would contain a higher proportion of contaminants that should not be classified as clusters. Similarly, this effect would influence the results yielded from the test set, as it would not give a true indication of how well our model detects clusters. Although, since we had a relatively large training set, it was impractical to directly check for contaminants in every image as well as perform spectroscopic follow-up of all ground truth cluster members. We must therefore assume that the Abell clusters in the WHL12 catalogue are actual clusters. In future work, it would also be important to account for astrometry errors in RA and Dec coordinates of clusters in the WHL12 catalogue, since we used these coordinates for our ground truth centre coordinates.

As with all deep learning algorithms, there are hyper-parameters that require either minor or extensive fine tuning depending on the task at hand. The main hyper-parameters of the Faster R-CNN algorithm include the learning rate,

momentum, gradient clipping threshold, mini-batch size, number of layers, number of neurons in each layer and architecture. However, to fully optimise every hyper-parameter would be computationally expensive without an efficient tuning strategy. For this work, we relied on the usage of transfer learning for partial optimisation of the hyper-parameters in our deep learning model. We had mostly adopted the defaulted values set for these hyper-parameters from the pre-trained model. From which, we show that the model is still capable of being adapted to perform generalised object detection of clusters. Although, in future work it would be valuable to explore hyper-parameter tuning of the Faster R-CNN algorithm for conducting cluster detection.

We adopted a specific methodology when generating all of our wide-field colour images. This involved using the same contrasting, image aspect and image scaling ratios for computational efficiency. From which, all future input images to our model will be somewhat restricted to using the same image pre-processing techniques in order to obtain maximum performance. However, this may create a trade-off between computational efficiency and bias from our image pre-processing. In future work, it would be beneficial to examine the effects from applying different image pre-processing constraints. We could also investigate whether applying additional image augmentation techniques (e.g. image rotation, vertical flipping) can increase the predictive performance of our model, where Perez & Wang (2017) showed that using simple image transformations (e.g. shifting, zooming, rotating, flipping, distorting, shading) can result in a more robust model.

We performed hold-out validation on the sample set to form a training set and a test set. However, this approach is limited to a simple approximation since we observed an imbalance of the cluster properties in Figure 2.3. For instance, we noticed that there were fewer lower redshift clusters compared to the number of higher redshift clusters. This meant our model could overfit to cluster properties that appeared more frequently. In future work, we could perform k-fold cross-validation or Monte Carlo cross-validation when splitting the sample set to examine the effects from using different compositions of training and test sets.

### 2.4.2 Future applications of this technique

LSST and *Euclid* are ideal surveys to apply our deep learning model to, as they will be wider and deeper than any survey conducted before them. This will likely result in the detection of many thousands of candidate high redshift or low mass clusters that are currently undiscovered. We expect that this will be an iterative process in practice when applying our model on these large datasets.

The Deep-CEE method will be of great use for confirming candidate clusters detected by X-ray, SZ or weak gravitational lensing surveys, as they often have many interlopers. In future work, we could also employ training sets that are composed of clusters found via different cluster finding techniques. For example, it would be interesting to compare clusters that are detected by a deep learning algorithm with a training set of X-ray selected clusters versus a training set of red-sequence fitting selected clusters. This may be a good way to test the various biases of different cluster detection methods, which can filter through to any cosmological predictions made with them.

In addition, our deep learning model could be adapted and applied to both optical imaging and other cluster detection methods at the same time. For example, we could train a multi-tasking algorithm to detect clusters from examining red-sequence fits and/or X-ray images alongside optical images. This would create a robust sample of clusters that are identified over a combination of different cluster finding techniques.

Finally, in order to prepare for LSST and *Euclid*, it would be practical to develop a data pipeline that is capable of processing the results from running our model on the entirety of SDSS as well as perform cross-matching of FP detections with existing cluster catalogues. This would be beneficial for constraining the purity and completeness of our model, as we expect many FP detections are likely to be actual clusters. In future work, we also aim to develop additional models that can predict the properties of clusters, such as redshift and richness. This will be vital for cataloguing the thousands of clusters discovered in upcoming galaxy surveys as well as improving our criteria for distinguishing between TP, FP and FN detections.

## 2.5 Conclusion

We present Deep-CEE, a novel deep learning model for detecting clusters in wide-field colour images and returning their respective RA and Dec. We used Abell clusters that were found by the WHL12 catalogue as the ground truth labels in images for our training and test sets. We trained our model to yield confidence scores of whether objects were likely to be clusters. We considered detections that had a predicted confidence score greater than a confidence score threshold as cluster candidates. We determined an optimal confidence score threshold that was based on the threshold value with the greatest F1 score. We initially found that an 80 per cent confidence score threshold was optimal for finding clusters in our test set, where we achieved a precision ratio of 70 per cent and a recall ratio of 81 per cent[13]. We then found that a 70 per cent confidence score threshold was the optimal threshold for detecting unseen clusters from the redMaPPer catalogue as part of another test set, where we achieved a precision ratio of 77 per cent and a recall ratio of 87 per cent. It should be noted that our precision ratios are specific to the test sets, since we were aware that some FP detections were likely to be actual clusters but we did not consider these as TP detections during our analysis. We showed that our model did not overfit to clusters in the training set, since we obtained a lower optimal confidence score threshold when we applied the model to unseen clusters. This suggested that our model was suitable for performing generalised object detection of clusters.

By applying Deep-CEE to upcoming wide-deep imaging surveys, such as LSST and *Euclid*, we expect to discover many new higher redshift and lower mass clusters. Our approach will also be a powerful tool when combined with catalogues or imaging data from other wavelengths (e.g. X-ray and SZ surveys). It is hoped that future cluster samples produced by Deep-CEE alone or in combination with other cluster finding techniques will be well-understood and therefore applicable for constraining cosmology as well as environmental galaxy evolution research. In future work, we will build upon this model by developing methods to estimate the

[13] An ideal confidence score threshold would have a precision and recall ratio of 100 per cent. (Tharwat, 2021)

properties of clusters (e.g. redshift and richness) in a similar manner to George Abell many years ago.

# Chapter 3

# Z-Sequence: Photometric redshift predictions for galaxy clusters with sequential random k-nearest neighbours

# Abstract

We introduce Z-Sequence, a novel empirical model that utilises photometric measurements of observed galaxies within a specified search radius to estimate the photometric redshift of galaxy clusters. Z-Sequence itself is composed of a machine learning ensemble based on the k-nearest neighbours algorithm. We implement an automated feature selection strategy that iteratively determines appropriate combinations of filters and colours to minimise photometric redshift prediction error. We intend for Z-Sequence to be a standalone technique but it can be combined with cluster finders that do not intrinsically predict redshift, such as our own Deep-CEE model. In this proof-of-concept study we train, fine-tune and test Z-Sequence on publicly available cluster catalogues derived from SDSS. We determine the photometric redshift prediction error of Z-Sequence via the median value of $|\Delta z|/(1+z)$ (across a photometric redshift range of $0.05 \leq z \leq 0.6$) to be $\sim 0.01$ when applying a small search radius. The photometric redshift prediction error for test samples increases by $30-50$ per cent when the search radius is enlarged, likely due to line-of-sight interloping galaxies. Eventually, we aim to apply Z-Sequence to upcoming imaging surveys such as the Legacy Survey of Space and Time to provide photometric redshift estimates for large samples of as yet undiscovered and distant clusters.

**Supplementary material for this chapter can be found in 2022chanphdchapter3supplementary.pdf**

## 3.1 Introduction

There are presently two approaches that are commonly used to determine galaxy redshifts, these are through spectroscopy and photometry (e.g. Walcher et al. 2011; Piattella 2018). However whilst the former is precise it is also time-consuming, expensive and difficult to perform for faint distant sources, which limits the number of observations with spectroscopic redshifts. Alternatively, photometric redshifts are fast to acquire and have been shown to be successful for faint distant sources (e.g. Ilbert et al. 2009). Conventional methods to estimate photometric redshift involve either empirical or template fitting algorithms. Empirical algorithms learn the underlying relationships between observed brightness, colour and spectroscopic redshift from a large training sample of galaxies (e.g. Weinstein et al. 2004; Lopes 2007; Carrasco Kind & Brunner 2013; Bilicki et al. 2018; Pasquet et al. 2019). Whilst, template fitting algorithms match observed fluxes to theoretical spectral energy distributions of different galaxy types at reference redshifts (e.g. Bolzonella et al. 2000; Babbedge et al. 2004; Gorecki et al. 2014; Fotopoulou & Paltani 2018). Nevertheless, photometric redshifts tend to have larger measurement errors than spectroscopic redshifts since photometric filters operate with low wavelength resolution, which means that individual spectral features cannot be utilised to determine redshift.

Photometric redshifts are often employed by imaging surveys to provide initial redshift estimates for many galaxies (e.g. Sánchez et al. 2014; Laigle et al. 2016; Beck et al. 2016; Tanaka et al. 2018), of which sub-samples can be followed up with spectroscopic redshifts. Similarly, it is important to develop models that will provide researchers with accurate initial redshift estimates for large and deep samples of the cluster population. In terms of predictive power for the low to intermediate redshift regime, empirical algorithms with sufficient training samples will generally outperform template fitting algorithms because template fitting algorithms require more physical assumptions when constructing spectral energy distributions to reflect possible observations. Whereas for the high redshift regime, template fitting algorithms will typically outperform empirical algorithms since high redshift training samples are more difficult to obtain due to observing limitations (Salvato et al., 2019).

In order to estimate redshifts for clusters, it is typically required to identify cluster members within a given search area. This can be conducted by utilising the red-sequence, which takes advantage of the fact that the red-sequence is seen as a well-defined linear relationship in colour-magnitude space (CMS) that evolves with redshift (Stott et al., 2009). From which, an empirical algorithm can estimate photometric redshift based on the observed red-sequence. This involves training an empirical algorithm to learn the redshifts from examples of known red-sequences, such that the red-sequence of an unknown cluster can be interpolated by the algorithm.

Additionally in order to break any colour-redshift degeneracies, where galaxies at different redshifts could have resembling colours, multi-dimensional CMS should be employed to reduce the reliance on specific colours. For example, a single colour that only utilises short wavelength optical filters would struggle to detect the red-sequence of a high redshift cluster since the filters would be unable to observe the redshifted 4000Å break[14], which is a distinctive broad spectral feature predominantly seen in the continuum spectrum of elliptical and lenticular galaxies (Dressler & Shectman, 1987). By utilising more colours, it is possible to straddle the 4000Å break to account for its transition at different redshifts (Rykoff et al., 2014).

For this work, we aim to employ an automated feature selection strategy that selects appropriate combinations of filters and colours in multi-dimensional CMS. We intend for this feature selection process to be fully data-driven based on observed galaxy photometry data, such that the selected features are effective at minimising photometric redshift prediction error. This method also comes with multiple practical benefits. Firstly, it is able to work with incomplete filter sets, as it does not rely on any specific filter. Secondly, it does not depend on galaxy photometric redshift catalogues. Thirdly, this approach can be combined with cluster finders that do not naturally predict redshift, such as Deep-CEE, since Z-Sequence only requires input astronomical coordinates and a photometry catalogue to predict photometric redshift of clusters.

[14]The 4000Å break is caused by the blanket absorption of photons at specific wavelengths from metals in the ionised atmospheres of old stellar populations (Kauffmann et al., 2003).

We structure this chapter with the following layout. In §3.2 we outline our methodology where §§3.2.1 describes our data pre-processing approach, §§3.2.2 describes our feature selection strategy plus machine learning algorithm and §§3.2.3 describes how we train our model. In §3.3 we present our results where §§3.3.1 describes the feature selection and filter magnitude-cut analysis, §§3.3.2 describes the hyper-parameter tuning and §§3.3.3 plus §§3.3.4 describes the tuned model performance on test sets. In §3.4 we review our findings where §§3.4.1 discusses the effectiveness of the tuned model at making predictions and §§3.4.2 discusses the practicality of the machine learning techniques used in this work. Finally, in §3.5 we summarise this work.

## 3.2 Methodology

### 3.2.1 Preparation of photometric datasets

We utilised candidate clusters that were detected in the SDSS-III DR8 by the WHL12 and redMaPPer cluster catalogues as part of our training, validation and test sets under a supervised learning approach. WHL12 used photometric redshifts of galaxies estimated by SDSS to identify overdense regions of galaxy clustering via the friends-of-friends grouping algorithm, in which a cluster redshift was calculated from the median value of determined cluster members. Whilst redMaPPer searched for the red-sequence within CMS across the SDSS sky coverage. The observed red-sequence profile of highly probable cluster members was then fit with a self-trained model of template red-sequences to estimate cluster redshift. It should be noted that the full WHL12 cluster catalogue has a photometric redshift range of $0.05 \leq z \leq 0.7846$ and the full redMaPPer cluster catalogue has a photometric redshift range of $0.0811 \leq z \leq 0.5983$.

Initially, we applied two selection criterion to the WHL12 cluster catalogue to identify clusters that had photometric redshifts between $0.0 < z < 0.6$ and also contained more than twenty observed member galaxies within an $r_{200}$ radius. This provided us with an approximation of the distribution of clusters found at different redshifts. From which, we calculated a mean photometric redshift

of $z = 0.3127$ based on the selected clusters. We used this mean photometric redshift to determine an angular distance of 54.96 arcseconds, which corresponds with the optical core radius of clusters of approximately 250 kpc. This angular distance also corresponds to a radius of approximately 100 kpc at $z = 0.1$ and 334 kpc at $z = 0.5$. We then cross-matched the clusters from the full WHL12 and redMaPPer cluster catalogues that were within 54.96 arcseconds and also within a photometric redshift range of $\pm 0.04(1 + z)$ as used by Wen et al. (2009)[15]. This ensured that we cleanly separated clusters to improve signal-to-noise in the dataset. The matching and non-matching clusters was then split into the following three datasets:

- **MWAR** - Cross-**m**atched **W**HL12 **a**nd **r**edMapper clusters.
- **WNMR** - **W**HL12 clusters with **n**o cross-**m**atched **r**edMapper clusters.
- **RNMW** - **r**edMapper clusters with **n**o cross-**m**atched **W**HL12 clusters.

Next, we reapplied our initial two selection criterion to all the clusters in the MWAR, RNMW and WNMR datasets. This split the clusters in each dataset into distinctive redshift and richness groupings, which can be used to examine how the Z-Sequence model performs on clusters that have these different properties. We set clusters that had properties within the selection criterion limits as the main training and test sets, whilst clusters that had properties outside the selection criterion limits were used as additional test sets. From which, the number of clusters within the selection criterion limits for the MWAR dataset was 8841 with a photometric redshift range of $0.0698 \leq z \leq 0.5986$, the WNMR dataset was 9723 with a photometric redshift range of $0.05 \leq z \leq 0.599$ and the RNMW dataset was 8646 with a photometric redshift range of $0.0811 \leq z \leq 0.5983$. The observed redshift distributions and positions of clusters from each dataset can be seen in Figures 3.1 and SA1 (available online).

We proceeded to cross-match the astronomical coordinates of clusters in each dataset to galaxies found in the SDSS-III DR9 photometric catalogue that were

[15] Wen et al. (2009) suggested that a photometric redshift gap of $\pm 0.04(1 + z)$ is a suitable indicator of true cluster richness, which corresponds to a rest frame velocity range of 24000 km $s^{-1}$ to account for the uncertainty of the photometric redshifts.

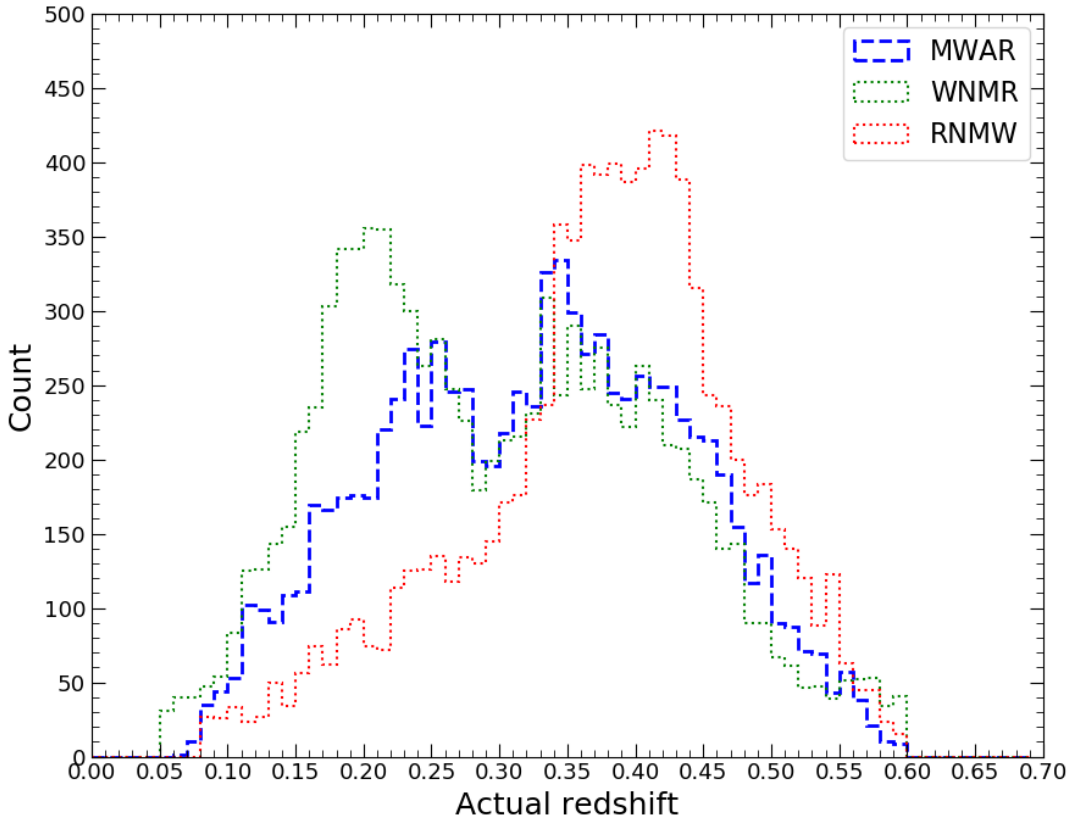


**Figure 3.1:** This figure displays a frequency histogram of the 'actual' redshift distributions of clusters, where photometric redshifts of clusters in the MWAR (blue dashed line) and WNMR (green dotted line) datasets were originally estimated by WHL12. Whilst the photometric redshifts of clusters in the RNMW (red dotted line) dataset were originally estimated by redMaPPer.

within the previously defined angular distance of 54.96 arcseconds. We selected 'primary' observations[16] of galaxies that had 'clean' photometry as determined by SDSS. This catalogue provided photometric measurements[17] for the following filters and colours:

- **Filters**: *u, g, r, i, z*,
- **Colours**: *u-g, g-r, r-i, i-z , u-r, g-i, r-z, u-i, g-z, u-z*,

where we used these filters and colours as our input features in §§3.2.2.

We assumed that any of the SDSS identified galaxies which lay along the line-of-sight and within 54.96 arcseconds of the input astronomical coordinates were part of the same cluster, from which we assigned each individual galaxy a cluster ID number for cross-referencing. To reduce the number of interloped galaxies, we empirically set multiple search radii of approximately 50, 100 and 150 kpc at the mean photometric redshift of $z = 0.3127$, which corresponds to angular distances of 10, 21 and 32 arcseconds respectively. The number of interlopers also depended on the position accuracy of the input cluster coordinates relative to the true cluster centroid. The reason we employed multiple search radii was to ensure that if the smallest search radius did not find a galaxy in the SDSS-III DR9 photometric catalogue, then the search radius would increase until a galaxy was found. This also provided a test for the effectiveness of the algorithm when given different views of the cluster core. It should be noted that this resulted in multiple forms of the training/validation/test sets containing additional galaxies in clusters found within each search radius.

We assigned the MWAR dataset as the training/validation sets and WNMR/RNMW datasets as test sets. The redshift distributions of the clusters in these datasets can be seen in Figure SA2 (available online) for each search radius. We chose

[16]The term 'primary' refers to the best imaging observation recorded for a survey object if it was seen multiple times during an observing run in an SDSS plate, whilst other observations of the object are called 'secondary'. A more in-depth explanation can be found on http://www.sdss3.org/dr9/help/glossary.php

[17]SDSS 'modelMag' measurements were used for filter magnitudes and colours of galaxies. This approach ensured the same aperture was used for all filters and the resultant magnitudes were calculated based off the best-fit model parameters observed in the r-band. For further details see http://www.sdss3.org/dr9/algorithms/magnitudes.php

| Filter | LM [mag] | LM-0.5 [mag] | LM-1.0 [mag] | LM-1.5 [mag] | LM-2.0 [mag] | LM-2.5 [mag] |
|---|---|---|---|---|---|---|
| $u$ | 21.6 | 21.1 | 20.6 | 20.1 | 19.6 | 19.1 |
| $g$ | 22.2 | 21.7 | 21.2 | 20.7 | 20.2 | 19.7 |
| $r$ | 22.2 | 21.7 | 21.2 | 20.7 | 20.2 | 19.7 |
| $i$ | 21.3 | 20.8 | 20.3 | 19.8 | 19.3 | 18.8 |
| $z$ | 20.7 | 20.2 | 19.7 | 19.2 | 18.7 | 18.2 |

**Table 3.1:** This table contains the SDSS limiting magnitude (LM) values of each filter with specified magnitude-cuts. The LM values were determined from 95 per cent completeness studies of point sources[18]. The filter magnitude values shown were converted from the SDSS *ugriz* magnitude system (Lupton et al., 1999) to AB magnitude system (Oke & Gunn, 1983). It should be noted that the SDSS *ugriz* magnitude system is very similar to the AB magnitude system but not exact, such that $u_{AB} = u_{SDSS} - 0.04$ and $z_{AB} = z_{SDSS} + 0.02$ (Abazajian et al., 2004).

the MWAR dataset as the training set since we expected that these clusters were more likely to host a populated core, where the red-sequence would be well-defined (Kodama et al. 1998; Gladders et al. 1998; Lidman et al. 2008; Mei et al. 2009; Newman et al. 2014; Strazzullo et al. 2016) in comparison to clusters in the WNMR/RNMW datasets, given the nature of the methods of WHL12 and redMaPPer. We aimed for our model to learn and utilise 'red-sequence'-like features found within high dimensional CMS to effectively predict photometric redshifts across a broad redshift range.

Finally, we investigated how varying the brightness for filter magnitude-cuts (see Table 3.1) could improve the accuracy of photometric redshift estimates, as this would remove galaxies from the less well-defined faint end of the red-sequence that had relatively large filter magnitude errors and filter magnitude values fainter than a specified limiting magnitude[18] value. In addition, we also compared the performance of using filter magnitude-cuts to a control group dataset that had no filter magnitude-cuts applied.

[18]Limiting magnitudes for the SDSS telescope were found by repeated observations of a patch of sky to obtain a magnitude value that provided approximately 95 per cent completeness of point sources (Strauss et al., 2002). See SDSS imaging camera scope at `http://www.sdss3.org/dr9/scope.php` for magnitude limits of each filter.

### 3.2.2 Model techniques

#### 3.2.2.1 Feature selection process

It should be noted that we had a total of 32767 possible combinations for the input features (see the filters and colours described in §§3.2.1) that could be tested. Due to the computational costs involved to examine all these combinations, we decided to employ an automated feature selection technique known as Sequential Forward Selection (SFS, Aha & Bankert 1995) to determine appropriate filters and colours. This technique is a 'greedy' iterative strategy that builds a subset of features via a bottom-up selection approach starting from an empty feature subset. Each iteration evaluates the performance of feature combinations, where SFS selects and stores the feature that best satisfies an objective function into the empty feature subset. From which, we employed a multi-objective function that checked if the following conditions were satisfied in each iteration of SFS:

- The following equation calculates the photometric redshift prediction error:

$$E_z = \frac{|P_i - A_i|}{(1 + A_i)}, \tag{3.1}$$

  where $E_z$ is the photometric redshift prediction error for each tested cluster, $P_i$ is the estimated photometric redshift for each cluster and $A_i$ is the 'actual'[19] photometric redshift for each cluster. In Figure SA3 (available online), we show a direct comparison of the photometric redshifts for cross-matched clusters from the WHL12 and redMaPPer cluster catalogues, where both catalogues appear to be in good agreement.

- The median of photometric redshift prediction errors produced during an iteration must be lower than the median of photometric redshift prediction errors from the previous iteration to continue SFS iterations.

[19]This depended on which dataset was used as the photometric redshifts of clusters in the MWAR and WNMR datasets were from the WHL12 cluster catalogue whilst photometric redshifts of clusters in the RNMW dataset were from the redMaPPer cluster catalogue.

- Filter magnitude-cuts were used to remove galaxies fainter than a specified magnitude threshold for each photometry filter to improve the signal-to-noise of the datasets. This can result in clusters with no galaxies remaining. We determined a percentage of clusters retained by counting the number of clusters that had galaxies remaining, after filter magnitude-cuts were applied, from the initial total in a dataset. From which, we set a threshold for the percentage of clusters retained in the MWAR dataset must be equal or greater than 95 per cent[20] to continue SFS iterations.

In Figure 3.2, we show that the SFS strategy is a computationally efficient approach as it searches through a reduced number of possible combinations, where all selected features are not included for reconsideration in subsequent SFS iterations. The process continues until the objective function is no longer satisfied with the remainder of the input features. We also compared the performance of these features to a control group of features that were not selected with SFS, where the control group features were *g, r, i, g-r, r-i, g-i*. We assumed that the control group features would perform well since these filters and colours would likely display 'red-sequence'-like features over a wide range of redshifts in CMS accounting for the shifting of the 4000Å break.

#### 3.2.2.2 Machine learning algorithm

We adopted the sequential random k-nearest neighbours (SRKNN, Park & Kim 2015) algorithm as the foundation of our model. The SRKNN algorithm is an ensemble (Dietterich, 2000) that aggregates multiple k-nearest neighbours (KNN, Fix 1951; Cover & Hart 1967) models into one global model (see Figure 3.3). The KNN algorithm is classed as a non-parametric learning method in the field of machine learning that can be used for non-linear regression tasks. This means that the algorithm has no pre-defined parameters to train, which is opposite to parametric learning methods (e.g. weights in a neural network algorithm). Predictions for the KNN algorithm are produced by averaging the labelled values of

[20] A tolerable percentage of data purposely excluded from the dataset should be low, otherwise systematic biases and sample misrepresentation induced by the missing data could be introduced (Kang, 2013).

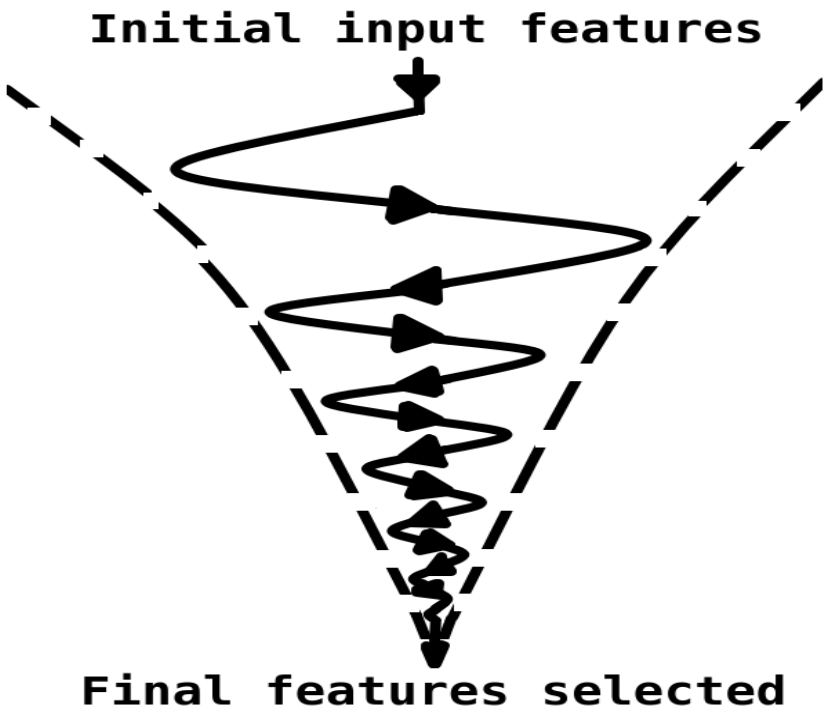


**Figure 3.2:** This figure displays a simplified perspective of the SFS strategy. The solid line with black arrows indicate the path taken by SFS to select features and the dashed lines represent the boundaries of feature space. It can be seen that as SFS progresses the feature space would shrink due to the reduced number of possible outcomes, where SFS would continue until it converges on a set of features. This diagram was inspired by Gutierrez-Osuna (2000).

the nearest neighbour training data points to the input data points, where we use the Euclidean distance metric[21] to compute distances. The main characteristics of the SRKNN algorithm involves bootstrap with replacement (Efron 1979; Efron & Tibshirani 1986) of the training set and random initialisation of input features to train each internal KNN model. These traits can improve the overall accuracy of predictions as a greater variety of features would be considered for each internal KNN model.

The SRKNN algorithm has three main hyper-parameter settings that should be optimised before deployment. These hyper-parameter settings are listed as follows:

- The number of internal KNN models (also equivalent to number of bootstrap resamples used).
- The number of randomly initialised input features.
- The number of nearest neighbours.

Park & Kim (2015) suggested that the performance of the SRKNN algorithm depends on the values assigned for each hyper-parameter setting, where the optimal values would vary for different datasets. In §§3.3.2, we examined and tuned each hyper-parameter setting with the MWAR validation set.

### 3.2.3 Outline of model training

Here, we describe the steps that were used to train and test our model for each search radius. The key points are summarised as follows:

1. Candidate clusters from the WHL12 and redMaPPer cluster catalogues were split into training, validation and test sets. The MWAR dataset was designated as the training/validation set (80:20 per cent split ratio), whilst

[21] It is known that distance comparisons in Euclidean space can become less effective with increasing dimensionality as the distance ratios become more uniform (Aggarwal et al., 2001). We note that other distance metrics such as cosine, Chi-squared, Manhattan and Minkowski could also be utilised for computing distances in the KNN algorithm (Hu et al., 2016).

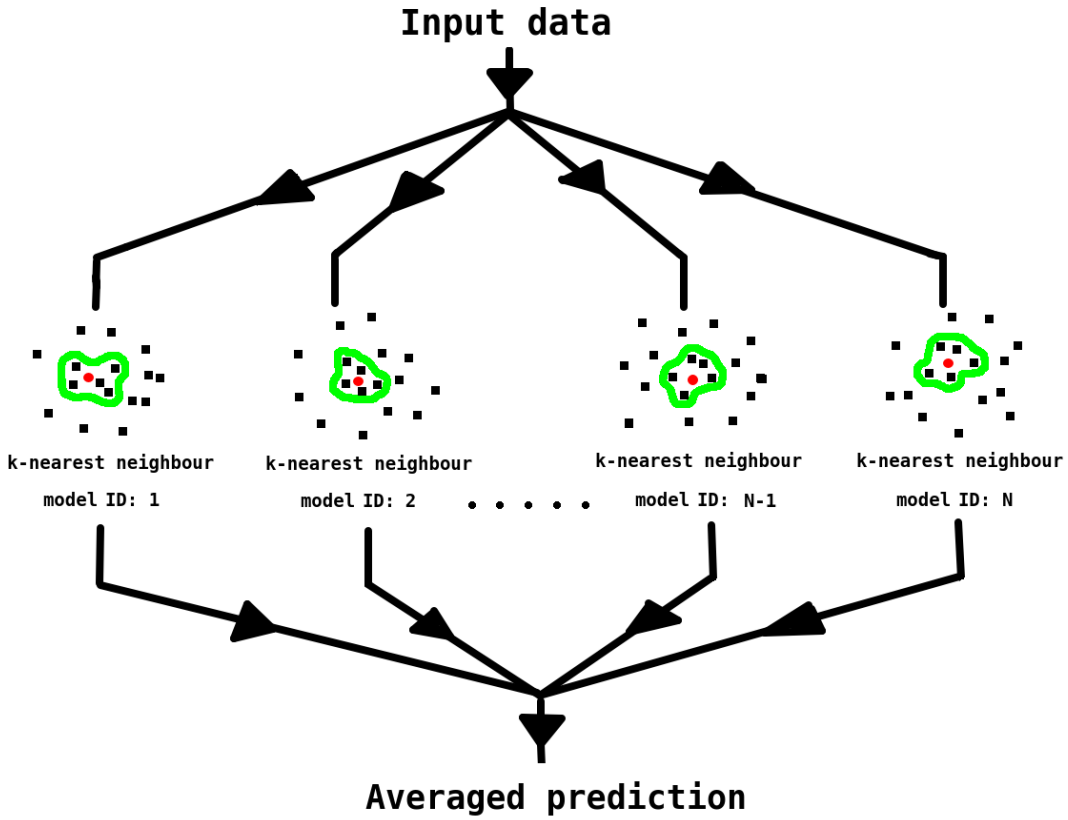


**Figure 3.3:** This figure displays a schematic diagram of the SRKNN algorithm. The solid lines with black arrows indicate the flow of input data to an 'N' number of internal KNN models. In this example diagram, we used a red circle in each internal KNN model to represent an input test data point, black squares represent training data points and the green outline show the nearest neighbour training data points from the input test data point. From which, the median of training label values for the corresponding nearest neighbour training data points was used as a prediction for an internal KNN model, where the global model prediction was approximated with the median of predictions across all internal KNN models. It should be noted that we utilised the Scikit-Learn machine learning library (Pedregosa et al., 2011) to construct the SRKNN algorithm.

the RNMW/WNMR datasets were used as test sets. Photometric measurements of observed galaxies in the clusters were obtained from the SDSS-III DR9 photometric catalogue and full-sky dust reddening maps (Schlegel et al. 1998; Schlafly & Finkbeiner 2011) were also used to account for galactic extinction.

2. All the filters and colours described in §§3.2.1 were assigned as input features to a single KNN algorithm for feature selection and filter magnitude-cut analysis. If a filter was used as part of an input feature, then the corresponding filter magnitude-cut was applied to exclude galaxies that had poor photometric measurements in that filter. The mean and standard deviation were also calculated for each feature in the MWAR training set to perform feature scaling[22]. From which, all input datasets to our model would require feature scaling with the same mean and standard deviation values determined for the MWAR training set.

3. Thirty repetitions of ten-fold cross-validation were computed with SFS for an individual KNN algorithm, where a single nearest neighbour was used[23]. This process was important for multiple reasons. Firstly, to analyse the stability of the KNN algorithm from minor changes to the training set. Secondly, to examine the relative frequency of features selected by SFS. Thirdly, to evaluate how filter magnitude-cuts affect the accuracy of photometric redshift predictions. Lastly, to provide a basis for comparing an individual algorithm with an ensemble algorithm.

4. The optimal filter magnitude-cuts determined for a single KNN algorithm were utilised for the SRKNN algorithm via transfer learning. From which,

[22] All photometric measurements of features were standardised with zero-mean centering and unit variance, which is necessary for the comparison of Euclidean distance measurements (Raschka, 2014).

[23] A single nearest neighbour minimises algorithmic biases but in turn maximises the variance of predictions (Friedman et al., 2001).

the training data for the internal KNN models of the SRKNN algorithm were built with bootstrap resamples, where bootstrap with replacement of the MWAR training set was used. Any clusters that were not used for bootstrapping of an internal KNN model were instead used for feature selection training of that internal KNN model with SFS. This ensured that all available training data was utilised.

5. The hyper-parameter settings of the SRKNN algorithm were tuned via a grid search strategy using hold-out validation of the MWAR validation set. This also examined how each of the hyper-parameter settings affected the model performance and generalisation.

6. Evaluation of the tuned model performance was obtained with the WNMR/RNMW test sets, which were all unseen clusters. Uncertainties for the photometric redshift estimate of each cluster were approximated with empirical bootstrap confidence intervals. Additionally, the tuned model was applied on clusters with low richness[24] and clusters at high redshift[25] to assess the response of the tuned model on clusters with unseen properties.

## 3.3 Results

### 3.3.1 Feature selection and filter magnitude-cut analysis

Following the procedures described in §3.2.3, we first examined the stability of photometric redshift predictions for a single KNN algorithm. As seen in Figure 3.4, we observed that for brighter filter magnitude-cuts the number of selected features by SFS were more contrasting, such that the resultant feature subsets

[24]We defined a cluster with low richness as a cluster that had twenty or fewer observed member galaxies.

[25]We defined a cluster at high redshift as a cluster that had a photometric redshift equal to or greater than 0.6, which was the upper limit of our training set.

for fainter filter magnitude-cuts were more strongly influenced by the observations in the MWAR training set itself. However, as seen from the corresponding photometric redshift prediction errors, we found that this did not significantly alter the stability of predictions. We also compared the performance of SFS selected features with the control group features (see §§§3.2.2.1), which had not been SFS selected. We repeated the same procedure used to analyse the SFS selected features for the control group features as well. From which, in Figure S1 (available online) we found that the control group features tended to have larger photometric redshift prediction errors in comparison to the SFS selected features for each search radius.

By repeatedly applying ten-fold cross-validation to the MWAR training set we could also examine the relative frequency of features selected by SFS. This was done by calculating the relative frequency of features observed in the best performing feature subsets across all thirty repeats. As seen from Table 3.2, we found that some of the features were frequently selected whilst other features were rarely chosen, such that certain features were more likely to be picked by SFS if they were present in the input features.

Next, we determined the optimal filter magnitude-cut for each search radius by identifying filter magnitude-cut values that returned the lowest photometric redshift prediction error and retained at least 95 per cent of clusters. In Figures 3.4 and S1 (available online), we found that the LM filter magnitude-cut was the optimal filter magnitude-cut for the 10 and 21 arcseconds search radii whilst the LM-0.5 filter magnitude-cut was the optimal filter magnitude-cut for the 32 arcseconds search radius. We also compared whether applying filter magnitude-cuts improved the predictive performance of the model. In Figures 3.4 and S1 (available online) we found that a dataset, NC, with no filter magnitude-cuts applied to it, was not the optimal filter magnitude-cut for any search radius whilst datasets with filter magnitude-cuts applied often had lower photometric redshift prediction errors.

We also assessed how magnitude-cuts of the filters themselves affected the percentage of clusters retained in the MWAR training set, where the optimal filter magnitude-cut for each search radius was applied. In Figure 3.5, we found that all filters, except for the $u$ filter, satisfied the 95 per cent cluster retainment

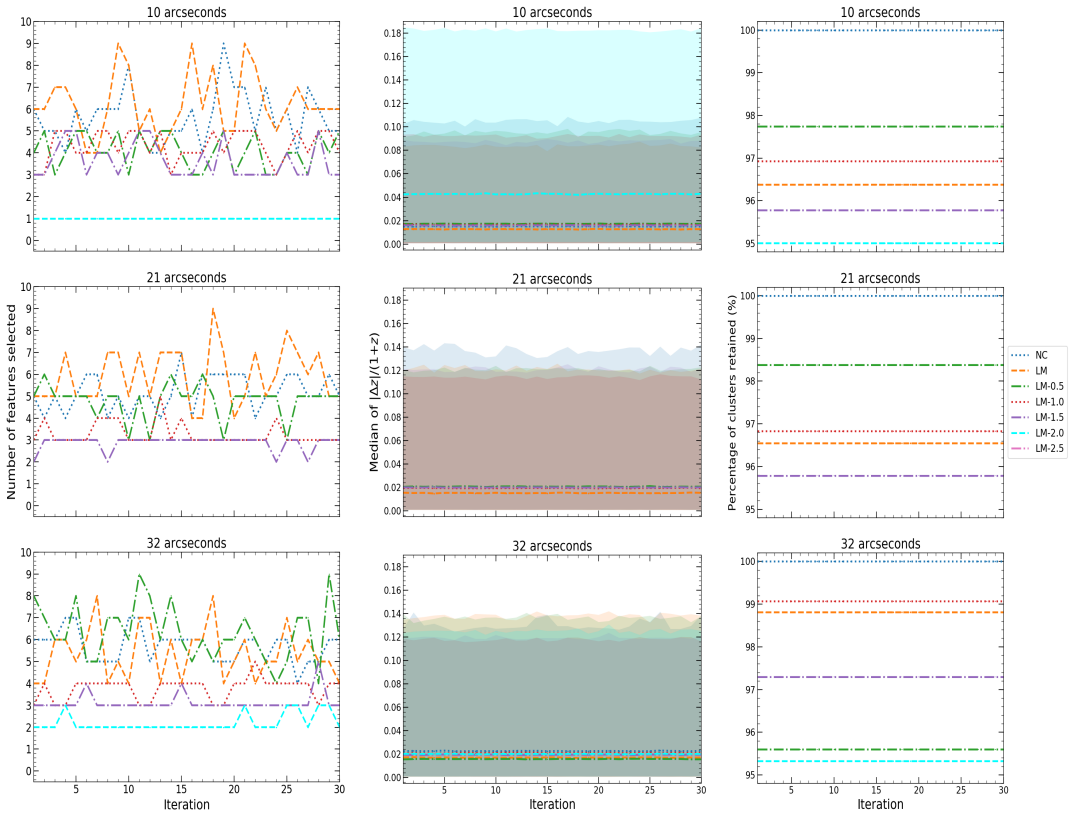


**Figure 3.4:** This figure displays the results from applying filter magnitude-cuts to the MWAR training set using a single KNN algorithm with SFS selected features for each search radius (10 arcseconds on the top row, 21 arcseconds on the middle row and 32 arcseconds on the bottom row). 'NC' represents a dataset with no filter magnitude-cuts applied and 'LM' represents the MWAR dataset with SFS selected features where filter magnitude-cuts were applied to the limiting magnitude of SDSS. In addition, 'LM' is the faintest filter magnitude-cut whilst 'LM-2.5' is the brightest filter magnitude-cut. Left column: Number of features selected for the best performing feature subset in ten-fold cross-validation across thirty repeats. Middle column: Median of photometric redshift prediction errors ($|\Delta z|/(1+z)$) across all tested clusters for the best performing feature subset in ten-fold cross-validation across thirty repeats, where the shaded regions represent 95 per cent confidence intervals. Right column: Percentage of test clusters retained after filter magnitude-cuts were applied with the best performing feature subset in ten-fold cross-validation across thirty repeats. It should also be noted that if the percentage of clusters retained, after filter magnitude-cuts were applied, did not satisfy the 95 per cent cluster retainment threshold we would not display the corresponding results in the other columns.

| Search Radius [arcseconds] | Optimal Filter Magnitude-Cut [mag] | SFS Selected Features | Relative Frequency Of SFS Selected Features (per cent) |
|---|---|---|---|
| 10 | LM | *r-i, g-z, r-z, g, g-i, z, r, i-z, g-r, i* | 100, 100, 90, 83, 67, 53, 47, 40, 30, 13 |
| 21 | LM | *z, r-i, g-i, g-z, r, g, g-r, i, r-z* | 87, 80, 80, 70, 63, 60, 60, 47, 47 |
| 32 | LM-0.5 | *g-z, r-i, g-i, g-r, g, i-z, z, r, i, r-z* | 93, 83, 83, 77, 70, 60, 50, 47, 43, 27 |

**Table 3.2:** This table displays the relative frequency of features selected by SFS across thirty repeats of ten-fold cross-validation on the MWAR training set with a single KNN algorithm at the optimal filter magnitude-cut for each search radius. The selected features were listed in the same order as the corresponding relative frequencies. It can be seen that the $z$ filter, rather than a colour, had the highest relative frequency amongst the features at the 21 arcseconds search radius for a single KNN algorithm but the relative frequency diminishes when the $z$ filter was instead used in an ensemble (see §§3.3.2).

threshold at each search radius. In addition, we observed in Table 3.2 that the $u$ filter did not appear in any final feature subset. From which, we decided that all input features which did not involve the $u$ filter would be used as the new input features for the SRKNN algorithm to reduce the computational cost of evaluating redundant features during feature selection training. One would expect the $u$ filter to be a poor predictor of redshift beyond very low redshift as it would probe further into the UV with increased redshift.

### 3.3.2 Hyper-parameter tuning analysis of the SRKNN algorithm

We combined the optimal filter magnitude-cuts learned in §§3.3.1 with a grid search strategy to fine-tune the SRKNN algorithm. We assumed that the knowledge learned for the KNN algorithm was appropriate for the SRKNN algorithm, since the SRKNN algorithm was an extension of the KNN algorithm. From which, we ran the grid search on all combinations of hyper-parameter settings with a

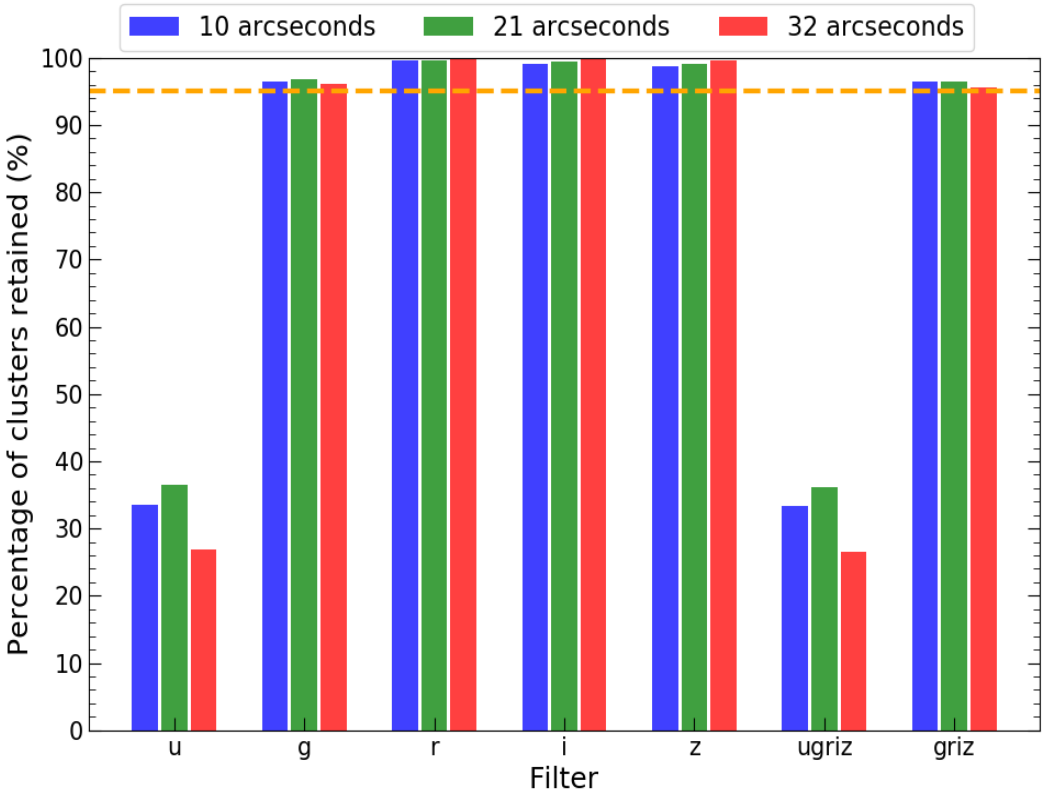


**Figure 3.5:** This figure displays the percentage of clusters retained in the MWAR training set after applying the optimal filter magnitude-cuts for each search radius to the *u*, *g*, *r*, *i*, *z*, *ugriz* and *griz* filters. The orange dashed line highlights the 95 per cent cluster retainment threshold.

specified range of values to evaluate how each hyper-parameter setting affected model generalisation and predictive performance. The following hyper-parameter setting values were used in the grid search:

- The number of internal KNN models - 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000.
- The number of initialised random features - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
- The number of nearest neighbours - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25.

We utilised validation curves (VanderPlas, 2016) to analyse the response from different hyper-parameter setting combinations of the SRKNN algorithm. This involved fixing each hyper-parameter setting as a constant with respect to the other hyper-parameter settings to compute the median of photometric redshift prediction errors across all tested clusters with that fixed hyper-parameter setting. We focused on minimising the photometric redshift prediction error on the MWAR validation set rather than the MWAR training set. Since the MWAR training set had already been seen by the model, the results from the MWAR training set would be biased whilst the MWAR validation set remained unseen by the model. However, running the model on both the MWAR training and validation sets was still beneficial to check the generalisation of the hyper-parameter settings, as the model could overfit and underfit when applied on its own training data.

Firstly, we evaluated the model performance based on the number of nearest neighbours for each search radius. In Figure 3.6, we found that for a small number of nearest neighbours the model had high predictive variance as we observed a large difference between the training and validation errors. Although, we noticed that the overall photometric redshift prediction error decreased as the number of nearest neighbours increased for the MWAR validation set, whereas the overall photometric redshift prediction error increased as the number of nearest neighbours increased for the MWAR training set. It can be seen that the number of

nearest neighbours was a very important hyper-parameter setting to tune since the model performance varied a lot depending on the value used. From which, we determined the optimal values for the number of nearest neighbours of each search radius to be 19 for 10 arcseconds, 19 for 21 arcseconds and 16 for 32 arcseconds. It should be noted that the number of nearest neighbours value with the lowest photometric redshift prediction error was actually 25 for each search radius. We purposely avoided selecting this value since the number of nearest neighbours value had a large impact on the model performance, such that selecting the hyper-parameter value with the lowest photometric prediction error could likely overfit the model on the MWAR validation set itself. Instead, we preferred to choose more conservative values for the optimal number of nearest neighbours to balance model generalisation and performance.

Secondly, we examined the model performance based on the number of initialised random features for each search radius. In Figure 3.7, we found that for both the MWAR training and validation sets, the change in the photometric redshift prediction errors quickly decreased for a small number of initialised random features but then slowly decreased when a medium to large number of initialised random features was used. From which, we observed that the overall redshift prediction error decreased as the number of initialised random features increased. This implied that the number of initialised random features was also an important hyper-parameter setting to tune, since the model performance on the MWAR training and validation sets was somewhat reliant on the value selected. We determined the optimal values for the number of initialised random features of each search radius to be 9 for 10 arcseconds, 8 for 21 arcseconds and 7 for 32 arcseconds. Although, it can be seen that having no initialised random features (using all features for the input features) at times had lower photometric redshift prediction errors. However, this could also worsen model generalisation since strongly correlated features would not be restricted during SFS. Therefore, we again decided to select more conservative values for the optimal number of initialised random features.

Thirdly, we assessed the model performance and behaviour based on the number of bootstrap resamples used for each search radius. In Figure 3.8, we show that for the MWAR training and validation sets the change in the photometric

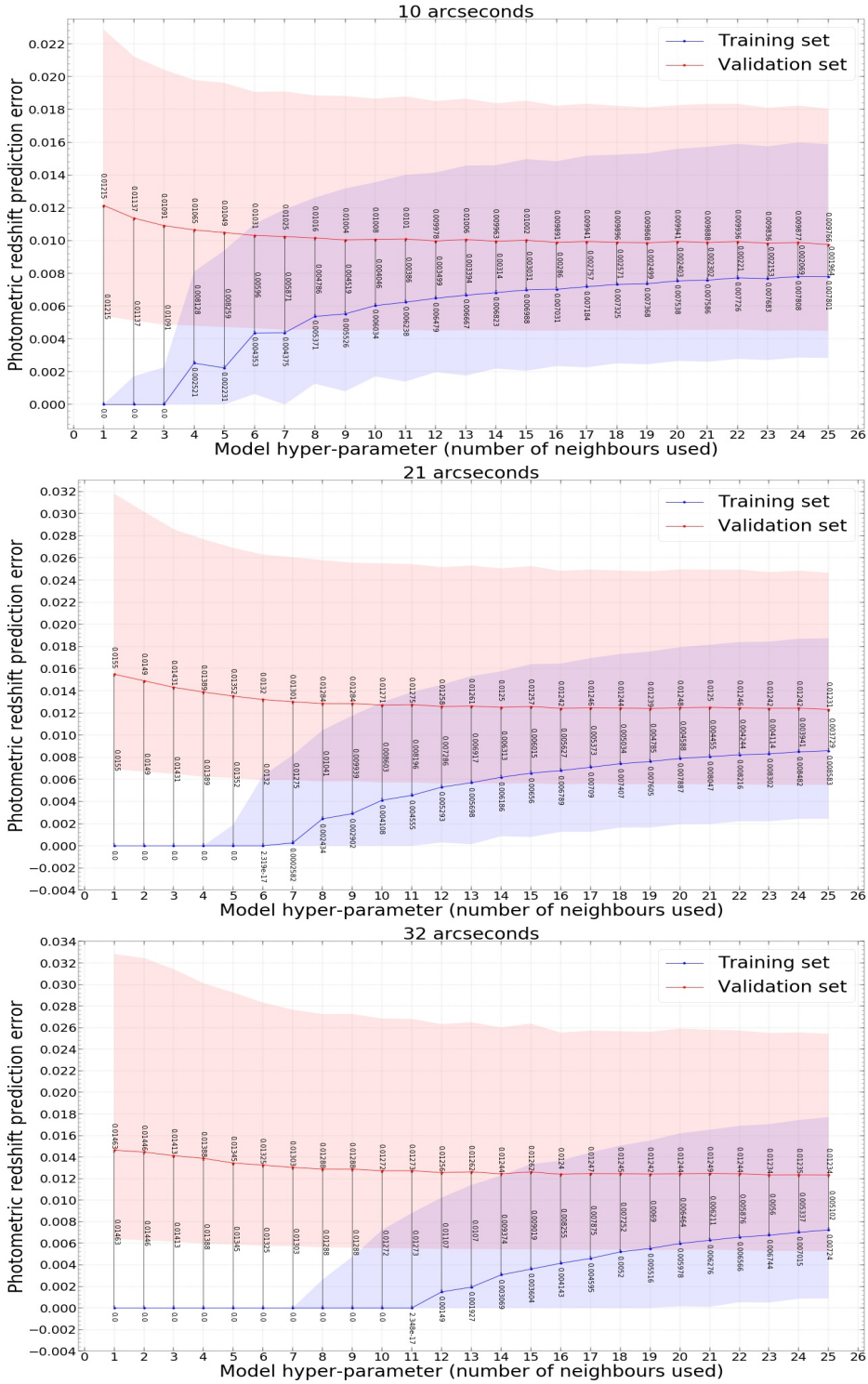


**Figure 3.6:** This figure displays validation curves from tuning the number of nearest neighbours hyper-parameter setting, where the photometric redshift prediction errors of the MWAR training (blue) and validation (red) sets are shown for each search radius (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the median of photometric redshift prediction errors across all tested clusters and the shaded regions represent the 25th and 75th percentiles of the photometric redshift prediction errors for a fixed number of nearest neighbours with respect to the other hyper-parameter settings of the SRKNN algorithm. We also labelled the difference between the individual points of the training and validation errors.

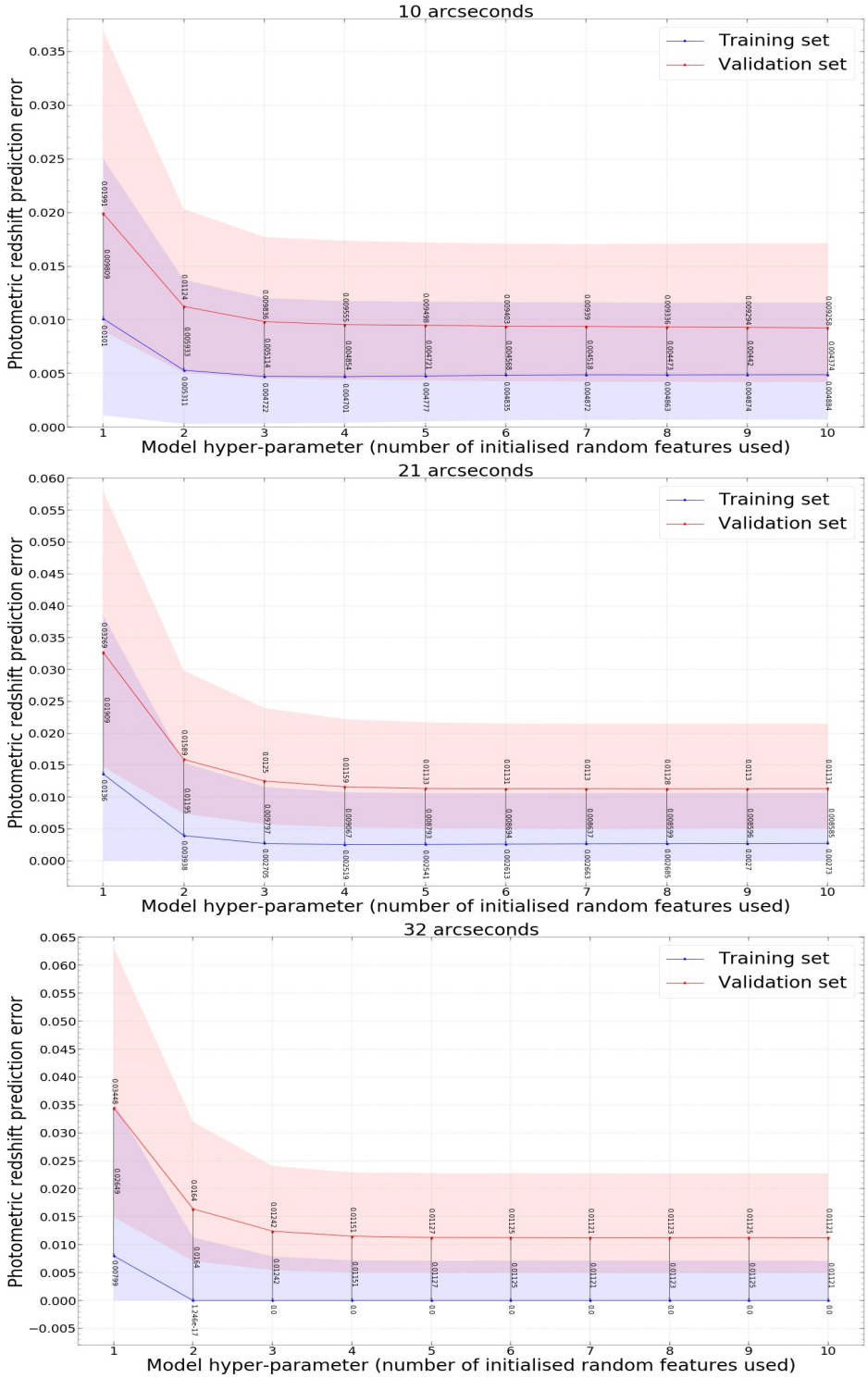


**Figure 3.7:** This figure is equivalent to Figure 3.6 except we tuned the number of initialised random features hyper-parameter setting.

redshift prediction error steeply decreased when a very small number of bootstrap resamples used but then remains flat as the number of bootstrap resamples increased. This tells us that the number of bootstrap resamples used was not a particularly important hyper-parameter setting to tune as the impact on the model performance for the MWAR training and validation sets was minimal. Efron & Tibshirani (1994) suggested that using fifty to two hundred bootstrap resamples was sufficient to calculate standard errors whereas bootstrap confidence interval estimates required at least one order of magnitude higher computational cost. From which, we decided that using one thousand bootstrap resamples for each search radius was enough to benefit from bootstrap confidence intervals. We also considered that since SFS would have selected different features for each bootstrap sample, we would not expect all internal KNN models to return predictions after filter magnitude-cuts were applied. In Figure 3.9, we show the percentage of clusters returned with full, partial and no bootstrap resamples returned for estimating photometric redshift at each search radius. We found that employing a large number of bootstrap resamples reduced the percentage of clusters returned with no bootstrap resamples. Whilst for clusters with a full set of bootstrap resamples returned, the percentage of clusters returned initially dropped but then remained flat as the number of bootstrap resamples increased. Whereas for clusters with partial bootstrap resamples returned, the percentage of clusters returned gradually increased as the number of bootstrap resamples increased. For this work, we preferred to minimise the percentage of clusters returned with no bootstrap resamples, since we wanted as many clusters as possible to have photometric redshift estimates. In Figure 3.10, we calculated the relative frequency of features selected by SFS with respect to the number of bootstrap resamples used at each search radius. It can be seen that as the number of bootstrap resamples increased, the spread of the relative frequency amongst the features decreased. From which, we also observed that the features with the highest relative frequency appeared to be colours whilst features with the lowest relative frequency were filters. The model had learned that colours were more significant than filters for estimating photometric redshifts of clusters.

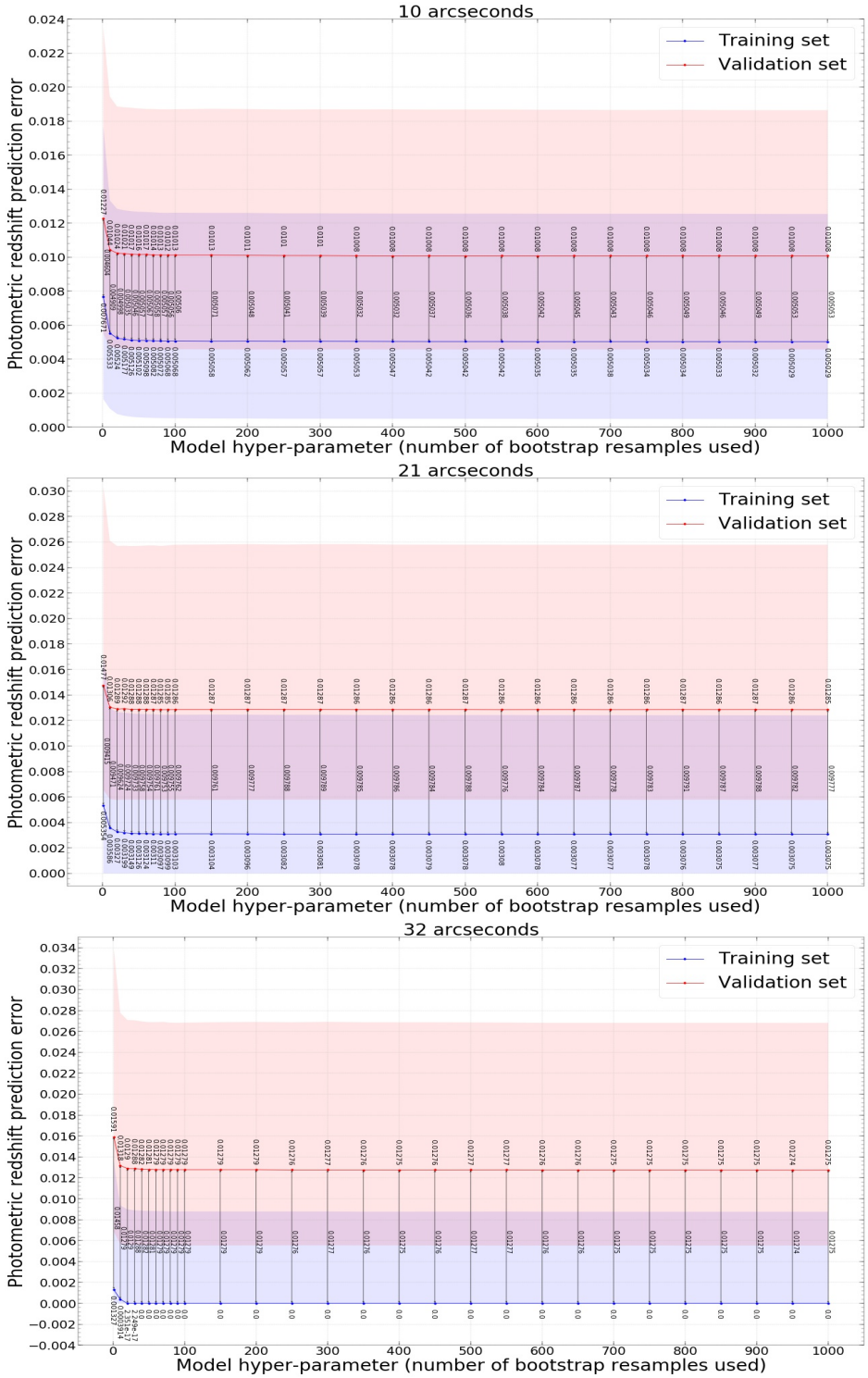


**Figure 3.8:** This figure is equivalent to Figure 3.6 except we tuned the number of bootstrap resamples hyper-parameter setting.

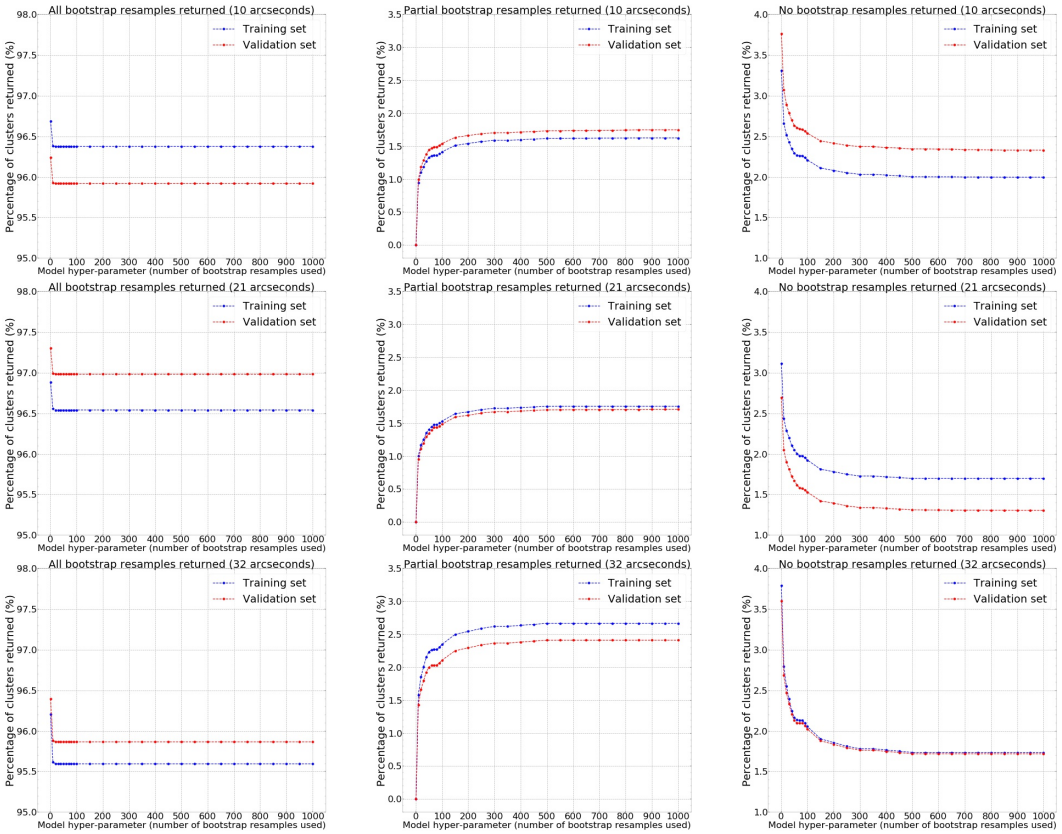


**Figure 3.9:** This figure displays validation curves from tuning the number of bootstrap resamples hyper-parameter setting, where the percentage of clusters returned with full, partial and no bootstrap resamples are from the MWAR training (blue) and validation (red) sets at each search radius (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the percentage of clusters returned across a fixed number of bootstrap resamples with respect to the other hyper-parameter settings of the SRKNN algorithm.

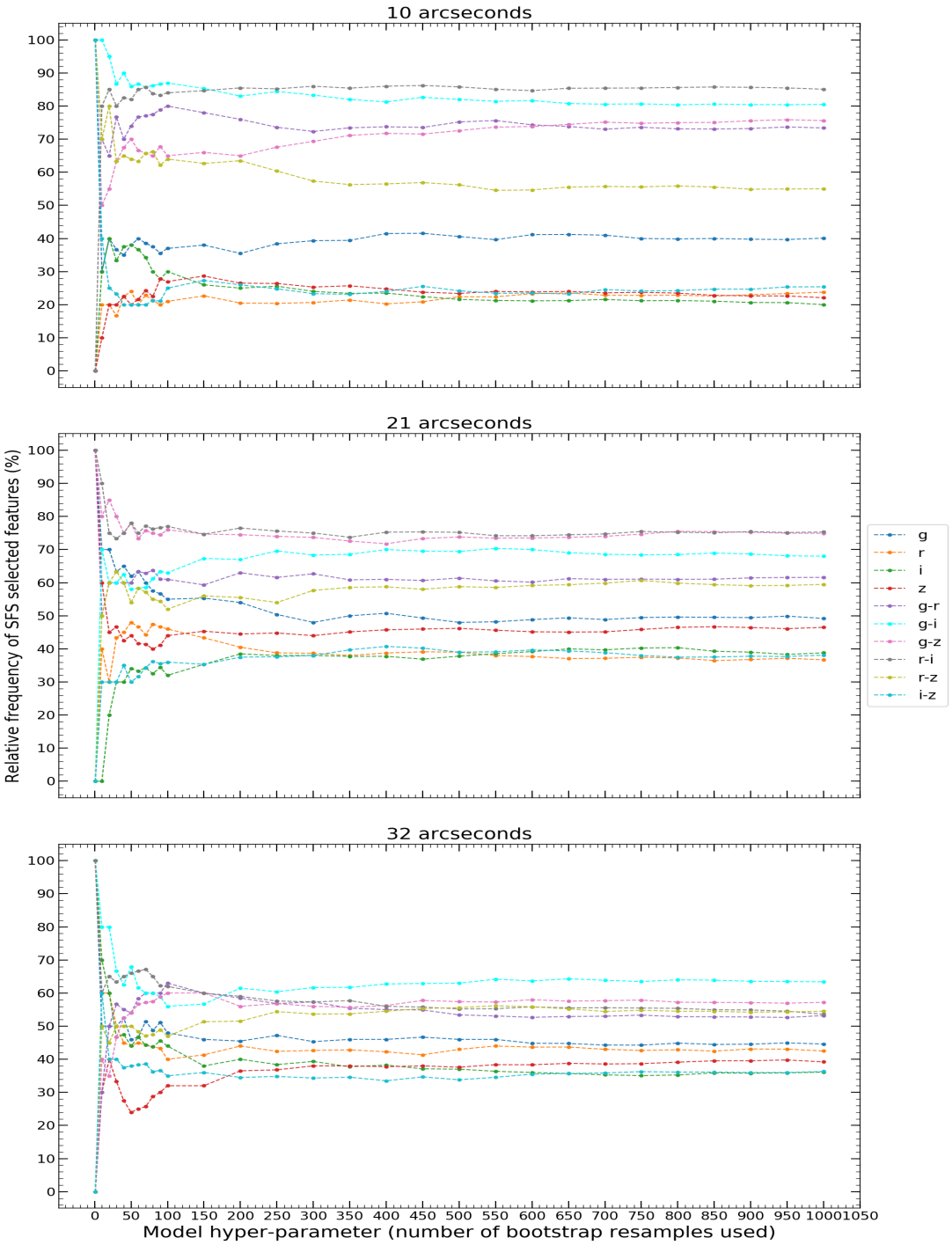


**Figure 3.10:** This figure displays validation curves from tuning the number of bootstrap resamples hyper-parameter setting, where the relative frequency of features selected by SFS with the MWAR training set is shown for each search radius (10 arcseconds on the top row, 21 arcseconds in the middle row and 32 arcseconds on the bottom row). The individual points display the relative frequency of features selected by SFS across a fixed number of bootstrap resamples with respect to the other hyper-parameter settings of the SRKNN algorithm.

| Test Set | Search Radius [arcseconds] | Optimal Filter Magnitude-Cut [mag] | # Clusters (total) | # Clusters (radius) | # Clusters (tested) | $\widetilde{E_z}$ |
|---|---|---|---|---|---|---|
| WNMR | 10 | LM | 9723 | 8844 | 8442 | 0.0106 |
| WNMR | 21 | LM | 9723 | 9564 | 9057 | 0.013 |
| WNMR | 32 | LM-0.5 | 9723 | 9691 | 9057 | 0.014 |
| RNMW | 10 | LM | 8646 | 8131 | 7319 | 0.0123 |
| RNMW | 21 | LM | 8646 | 8577 | 7870 | 0.0156 |
| RNMW | 32 | LM-0.5 | 8646 | 8635 | 7416 | 0.0181 |

**Table 3.3:** This table displays the median of photometric redshift prediction errors ($\widetilde{E_z}$, where $E_z = |\Delta z|/(1+z)$) across all tested clusters for each test set, search radius and optimal filter magnitude-cut. We also show the total number of clusters in the original full dataset (total), the number of clusters that had galaxies within the specified search radius (radius) and the number of clusters that had galaxies within the specified search radius after filter magnitude-cuts (tested). The values in this table summarise the test results in Figures 3.11, 3.12, 3.13, 3.14, 3.15 and 3.16.

### 3.3.3 Model performance analysis with test sets

We used the WNMR/RNMW test sets to assess the performance of the SRKNN algorithm with the optimal hyper-parameters learned in §§3.3.2 for each search radius. As described earlier in §§3.2.1, the test sets contained clusters from the WHL12 and redMaPPer cluster catalogues with no corresponding cross-match. A summarised version of the test results can be found in Table 3.3.

In Figures 3.11, 3.12, 3.13, 3.14, 3.15 and 3.16 we compared the known photometric redshifts with the predicted photometric redshifts for clusters in the WNMW/RNMW test sets that had full bootstrap resamples returned by the tuned model. We found that as the search radius increased the median of photometric redshift prediction errors across all tested clusters in both test sets increased as well possibly due to line-of-sight interloping galaxies. From which, in Figures SA4, SA5, SA6, SA7, SA8 and SA9 (available online) we also examined the spatial distribution of several clusters with relatively large photometric redshift prediction errors. We repeatedly observed that if line-of-sight interloping

galaxies were present within the search radius of clusters, the resultant model predictions had relatively large photometric redshift prediction errors. Moreover in Figures 3.11, 3.12, 3.13, 3.14, 3.15 and 3.16 it can be seen that the width of the 95 per cent confidence intervals around predictions decreased as the search radius increased, as shown by wider intervals. This meant there was lower precision of the predicted photometric redshift value. Despite this, we found that the tuned model seemed to perform well at all redshifts since the majority of cases had relatively low photometric redshift prediction errors for each search radius. Although, we noticed that an increasing number of cases had relatively large photometric redshift prediction errors near to the redshift training boundaries of the MWAR training set as the search radius increased. Furthermore, we also examined the performance of the tuned model on clusters in the WNMW/RNMW test sets with only partial bootstrap resamples returned for each search radius. From Figures S2, S3, S4, S5, S6 and S7 (available online) we found that in almost all cases the photometric redshift prediction error was poorly constrained when partial bootstrap resamples were used.

In Figures 3.17 and 3.18 we determined the number of galaxies used in photometric redshift predictions of clusters from the WNMR/RNMW test sets that had full bootstrap resamples returned by the tuned model for each search radius. This examined how the tuned model performed with respect to different numbers of galaxies. It can be seen that as the search radius increased the number of galaxies used in photometric redshift predictions increased too. From which, we found that the median of photometric redshift prediction errors across all tested clusters was similar regardless of the number of galaxies used by the tuned model. Although, we noticed that clusters with larger numbers of galaxies used for photometric redshift predictions were frequently seen between low and intermediate redshifts with relatively low photometric redshift prediction errors. Whereas clusters at considerably lower and higher redshifts rarely had large numbers of galaxies used for photometric redshift predictions and also had relatively large photometric redshift prediction errors.

In Figures 3.19 and 3.20 we examined the redshift distribution of clusters from the WNMR/RNMW test sets with no bootstrap resamples returned by the tuned model for each search radius. We observed that the redshift distributions

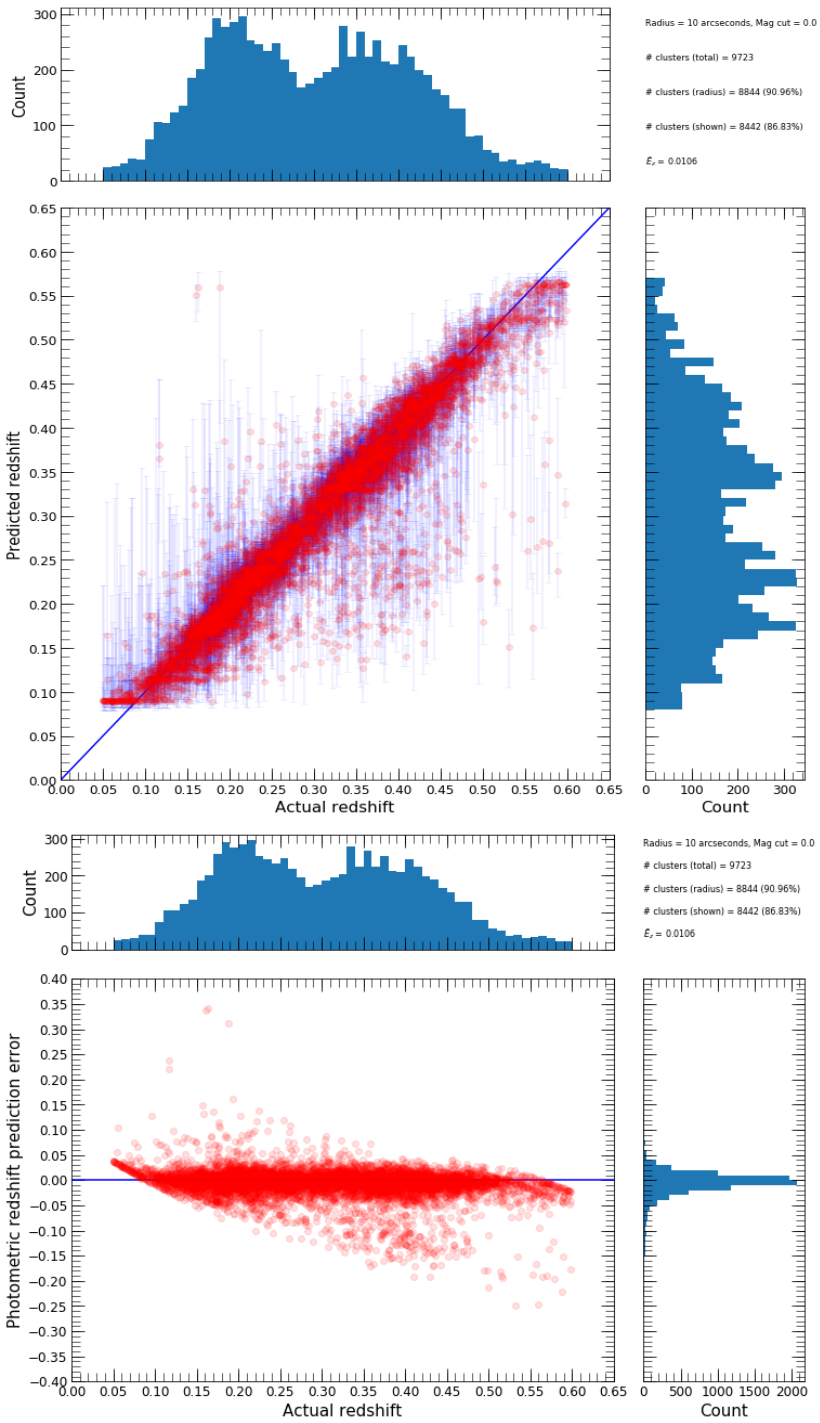


**Figure 3.11:** This figure displays the performance of photometric redshift predictions of clusters for the WNMR test set that had full bootstrap resamples returned within a 10 arcseconds search radius. Top row: Predicted versus 'actual' photometric redshift of tested clusters with frequency histograms of the distributions. Bottom row: Non-absolute photometric redshift prediction error versus 'actual' redshift of tested clusters with frequency histograms of the distributions. Other: '# clusters (total)' represents the total number of clusters in the WNMR dataset, '# clusters (radius)' represents the number of clusters in the WNMR test set that had observed galaxies within a 10 arcseconds search radius, '# clusters (shown)' represents the number of clusters in the WNMR test set that had observed galaxies within a 10 arcseconds search radius with full bootstrap resamples returned, $\widetilde{E_z}$ represents the median of photometric redshift prediction errors across all tested clusters within a 10 arcseconds search radius with partial bootstrap resamples returned.

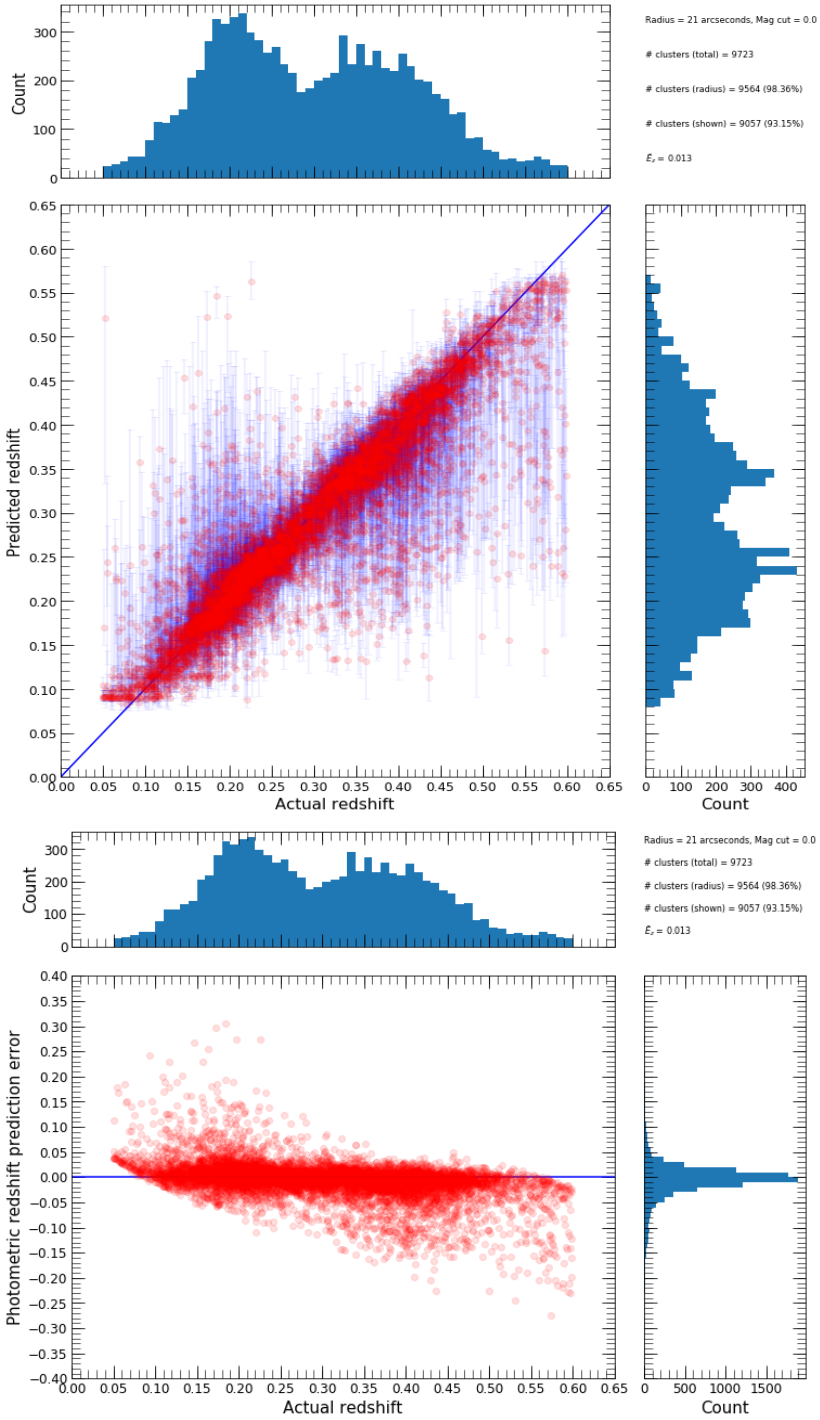


**Figure 3.12:** This figure is equivalent to Figure 3.11 except we examined the performance of photometric redshift predictions of clusters within a 21 arcseconds search radius.

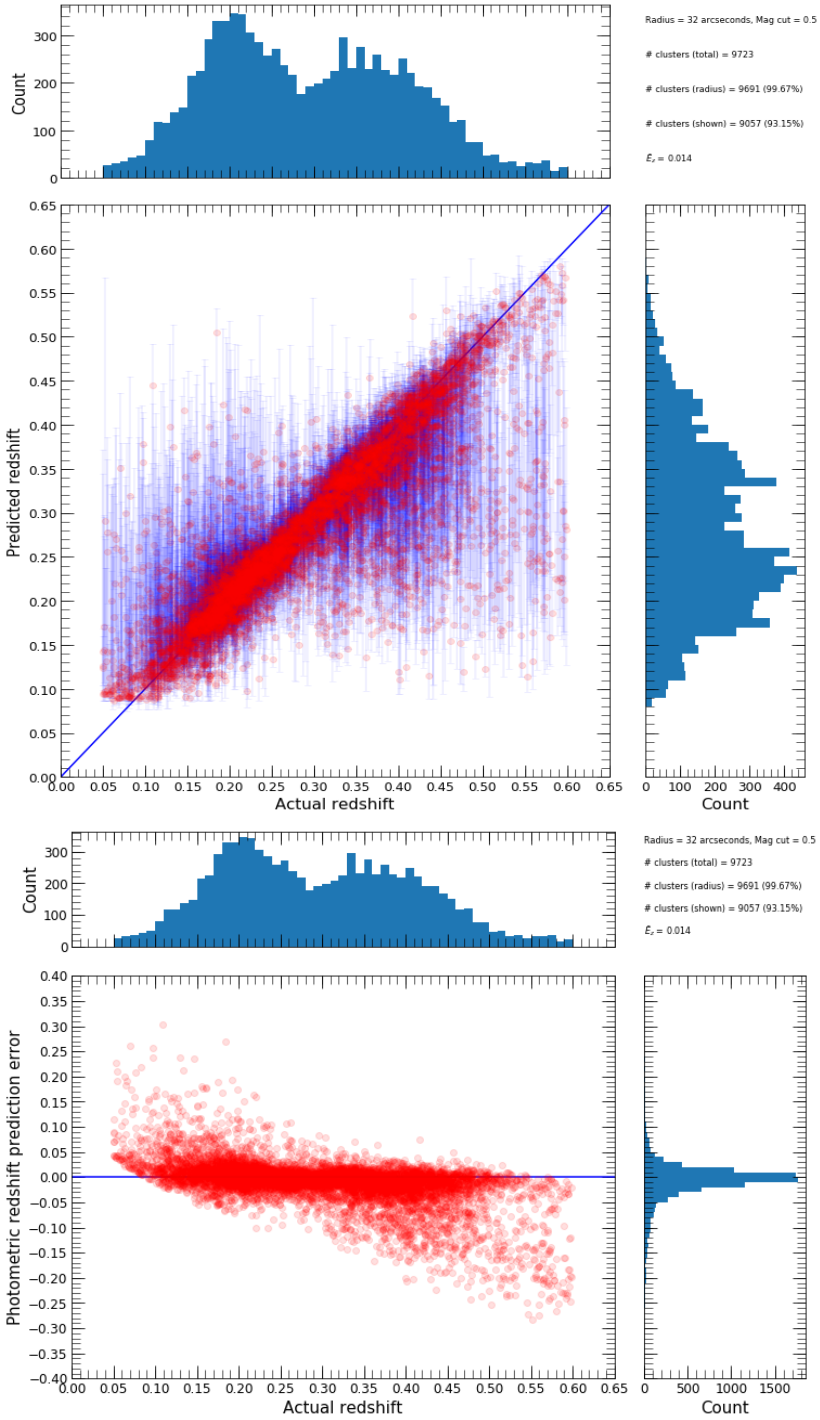


**Figure 3.13:** This figure is equivalent to Figure 3.11 except we examined the performance of photometric redshift predictions of clusters within a 32 arcseconds search radius.

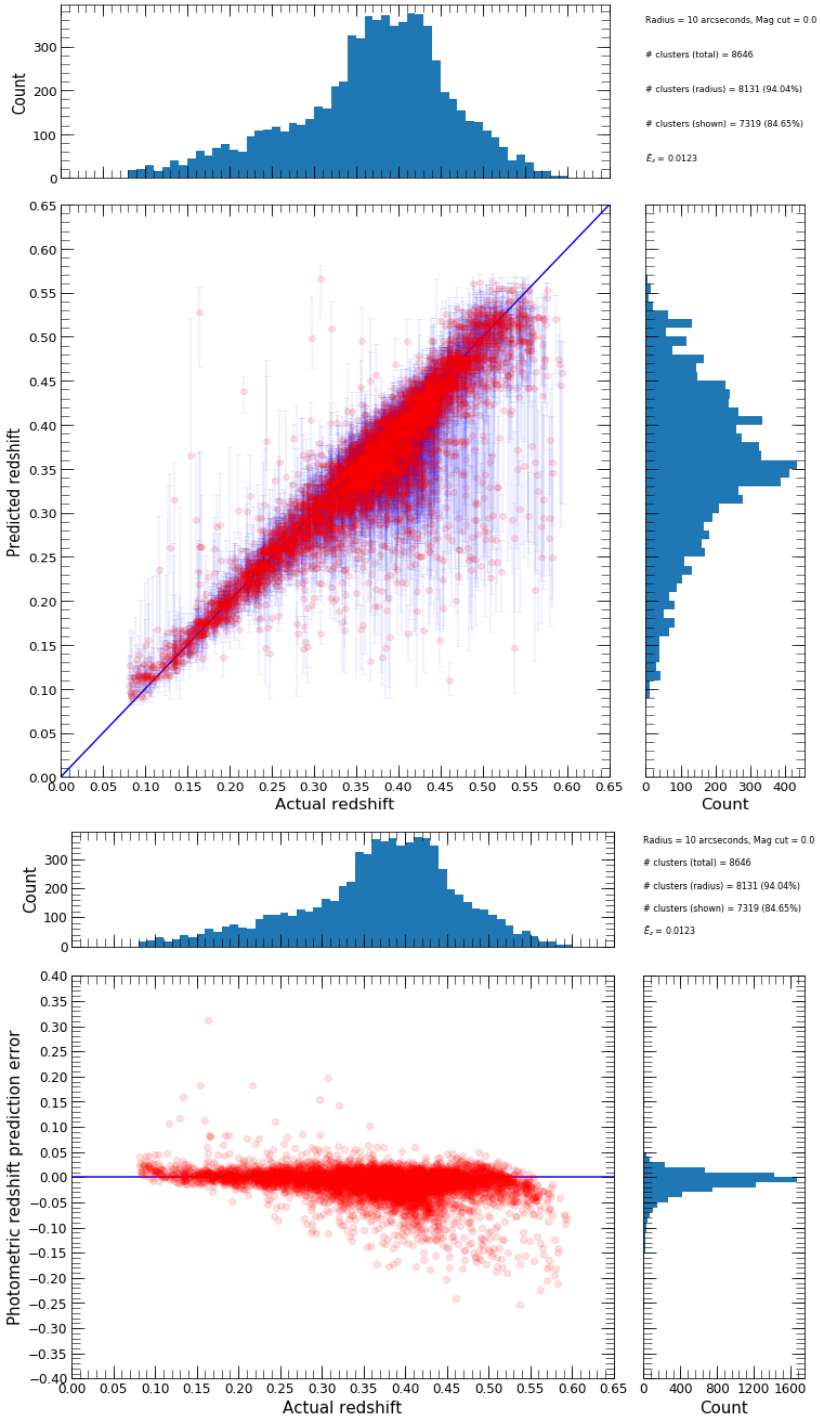


**Figure 3.14:** This figure is equivalent to Figure 3.11 except we examined the performance of photometric redshift predictions of clusters within a 10 arcseconds search radius for the RNMW test set.

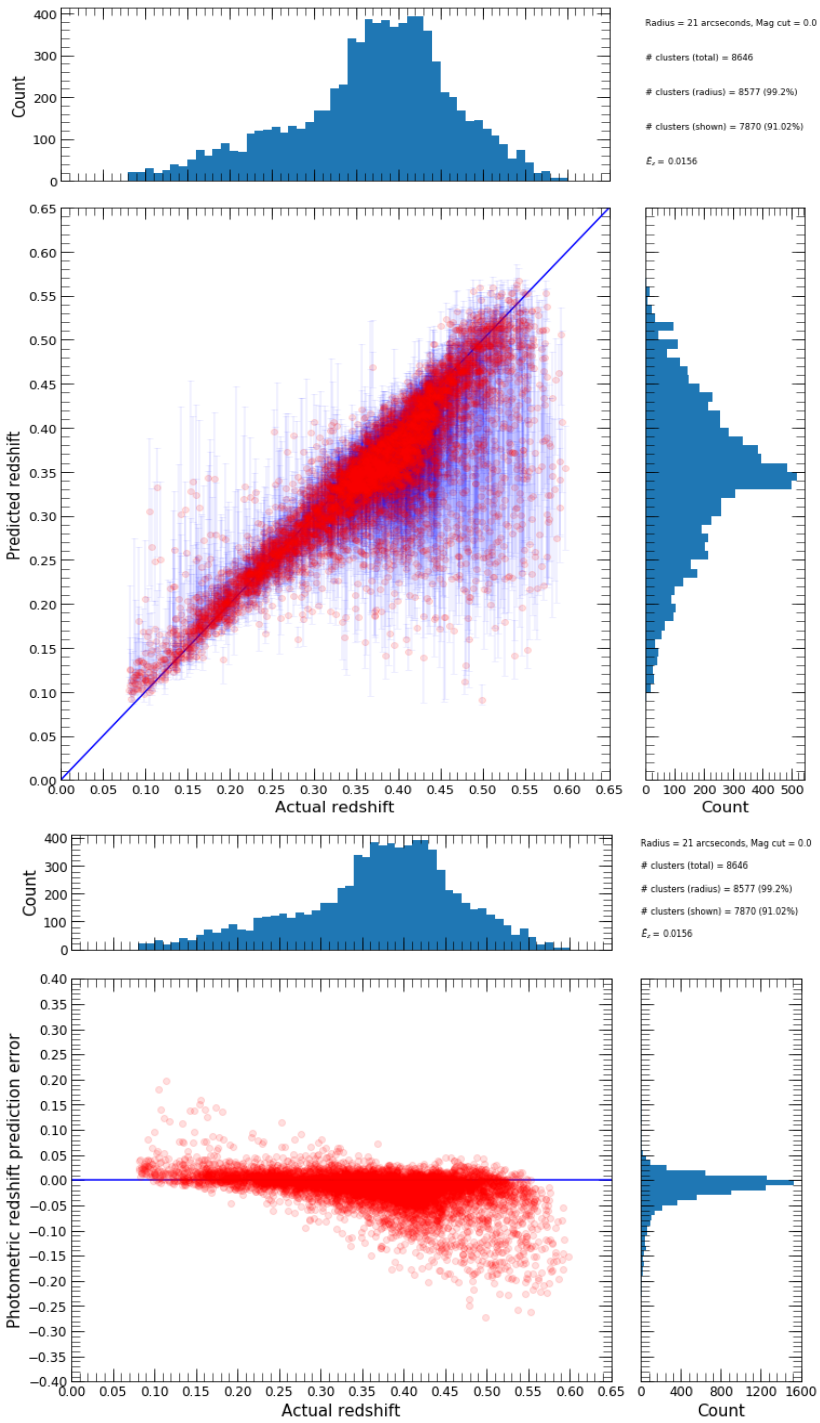


**Figure 3.15:** This figure is equivalent to Figure 3.11 except we examined the performance of photometric redshift predictions of clusters within a 21 arcseconds search radius for the RNMW test set.

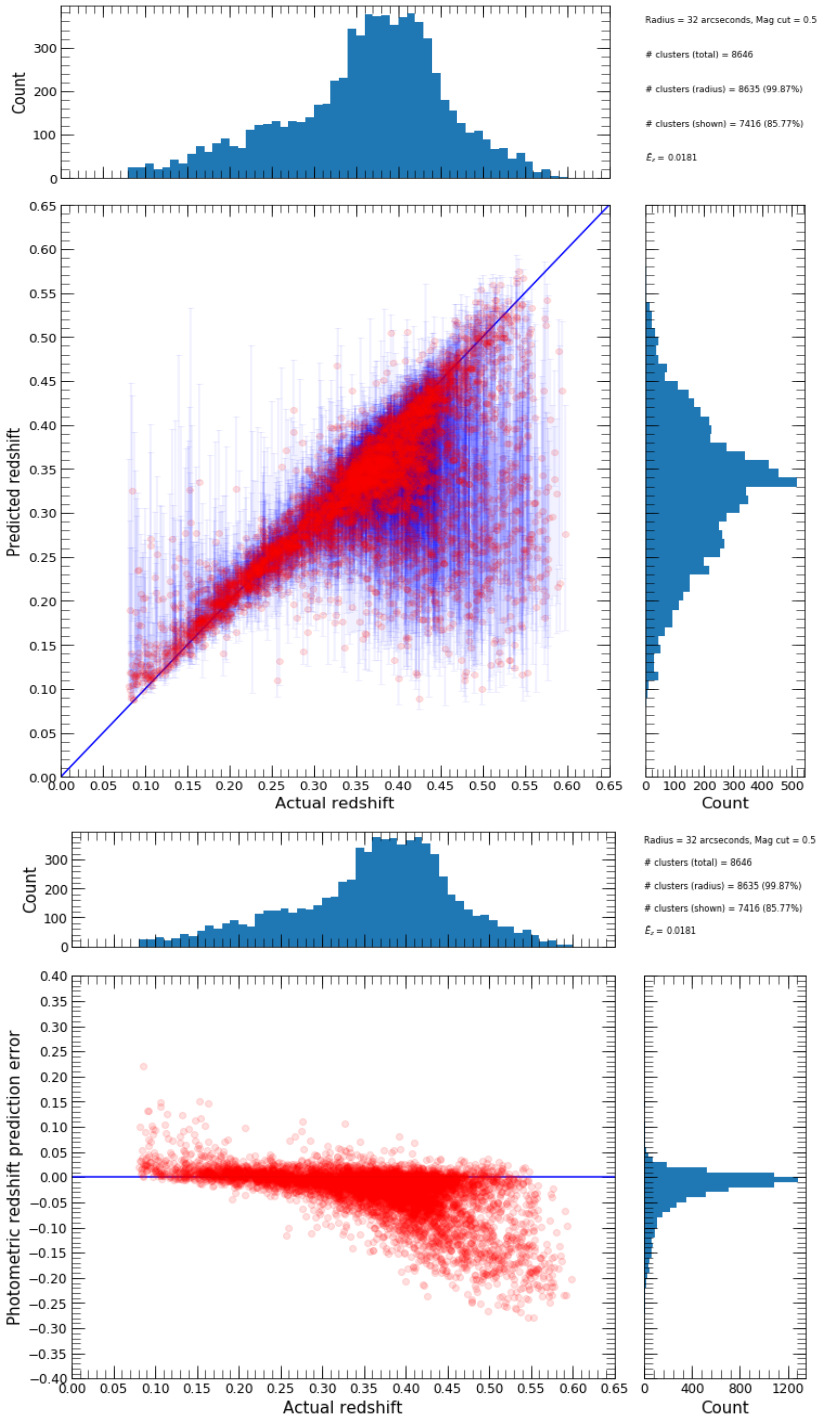


**Figure 3.16:** This figure is equivalent to Figure 3.11 except we examined the performance of photometric redshift predictions of clusters within a 32 arcseconds search radius for the RNMW test set.

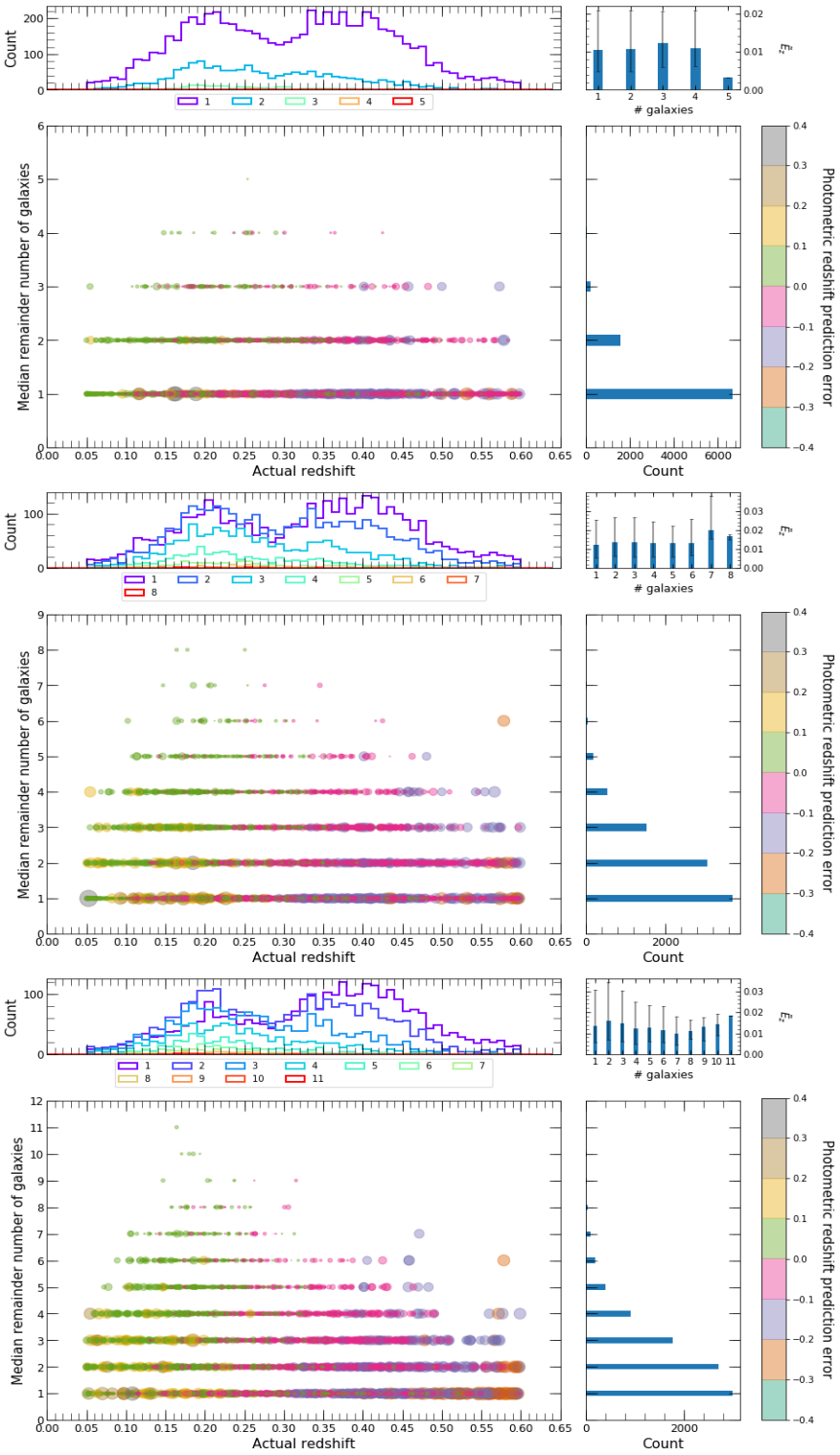


**Figure 3.17:** This figure displays the number of galaxies used in photometric redshift predictions versus 'actual' redshift of tested clusters for the WNMR test set, where predictions had full bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius. It should be noted that the size of individual points change in relation to the value of the non-absolute photometric redshift prediction error. Frequency histograms of the distributions are also shown. $\widetilde{E_z}$ represents the median of photometric redshift prediction errors across all tested clusters for each number of galaxies bin.

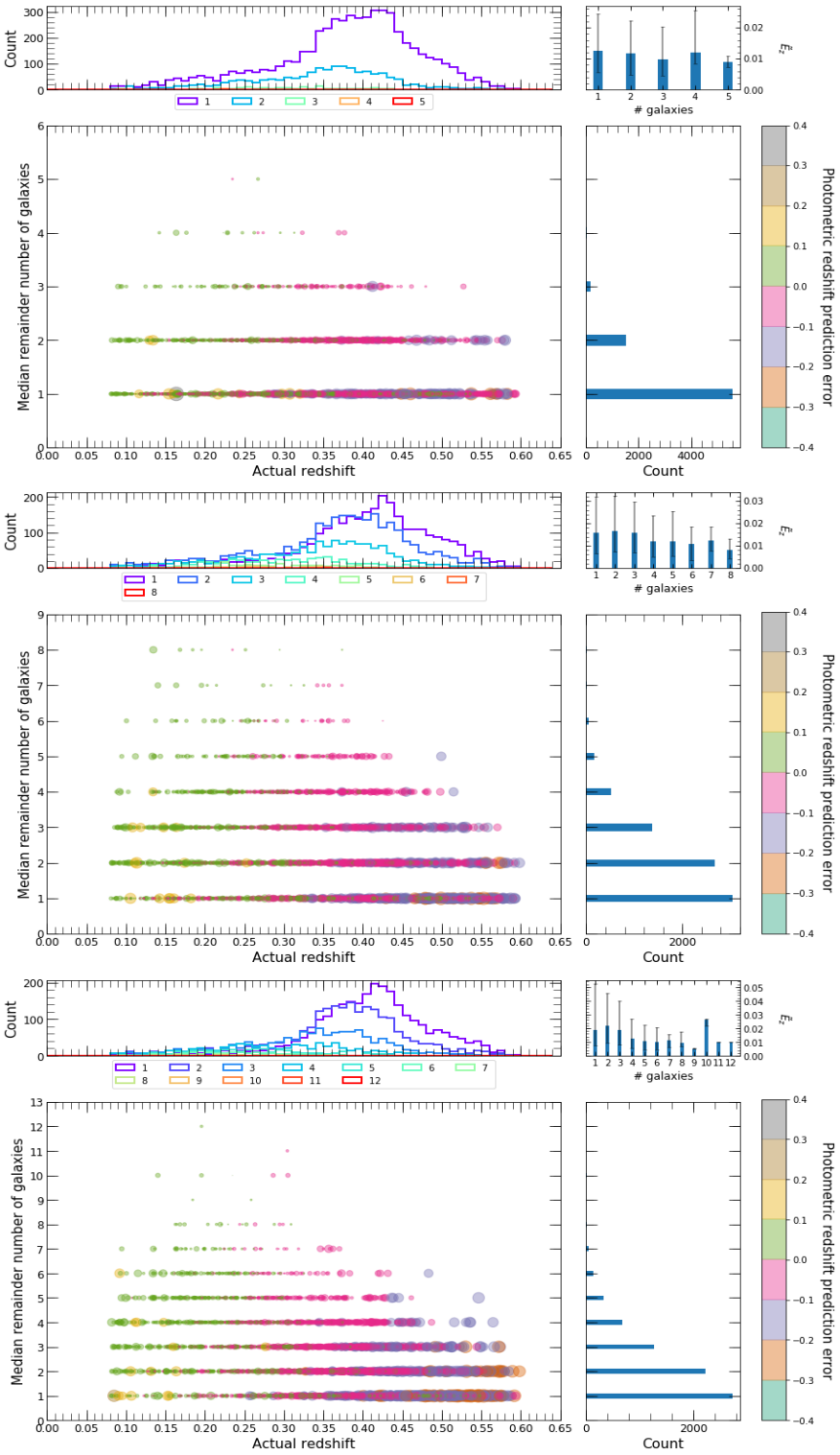


**Figure 3.18:** This figure is equivalent to Figure 3.17 except we examined the number of galaxies used in photometric redshift predictions for the RNMW test set.

were predominantly skewed towards higher redshifts. This could be due to the galaxies in clusters at higher redshifts having poorer photometric measurements in comparison to the galaxies in clusters at lower redshifts. Although, it should be noted that the redshift distribution for the RNMW dataset itself was also heavily skewed towards higher redshifts.

### 3.3.4 Further model testing

We also tested the tuned model on additional clusters that resided in unseen parameter space, such as clusters with low richness and clusters at redshift equal or greater than 0.6. This was to analyse the generalisation of the tuned model, by running it on clusters with properties that it had not been trained for, which were also likely to be encountered in surveys. We applied the same analysis procedure performed in §§3.3.3 and provide the full results in the online supplementary material. For this subsection, we will only describe the response of the tuned model with respect to different cluster properties.

In Figures S8, S9 and S10 (available online) we ran the tuned model on clusters with low richness, which had a richness of twenty or fewer observed member galaxies such that they did not qualify for the MWAR dataset, to obtain photometric redshift predictions that had full bootstrap resamples returned at each search radius. We found that the number of cases with relatively large photometric redshift prediction errors increased as the search radius increased, particularly at higher redshifts. However, we also noticed that the median of photometric redshift prediction errors for each search radius remained relatively low when compared to the median of photometric redshift prediction errors for the WNMR/RNMW test sets. Moreover, we observed that the precision of the 95 per cent confidence intervals became worse towards the redshift training boundaries when the search radius increased.

In Figures S16, S17 and S18 (available online) we ran the tuned model on clusters at high redshift, which had a redshift beyond the redshift training boundaries such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that had full bootstrap resamples returned at each search radius.

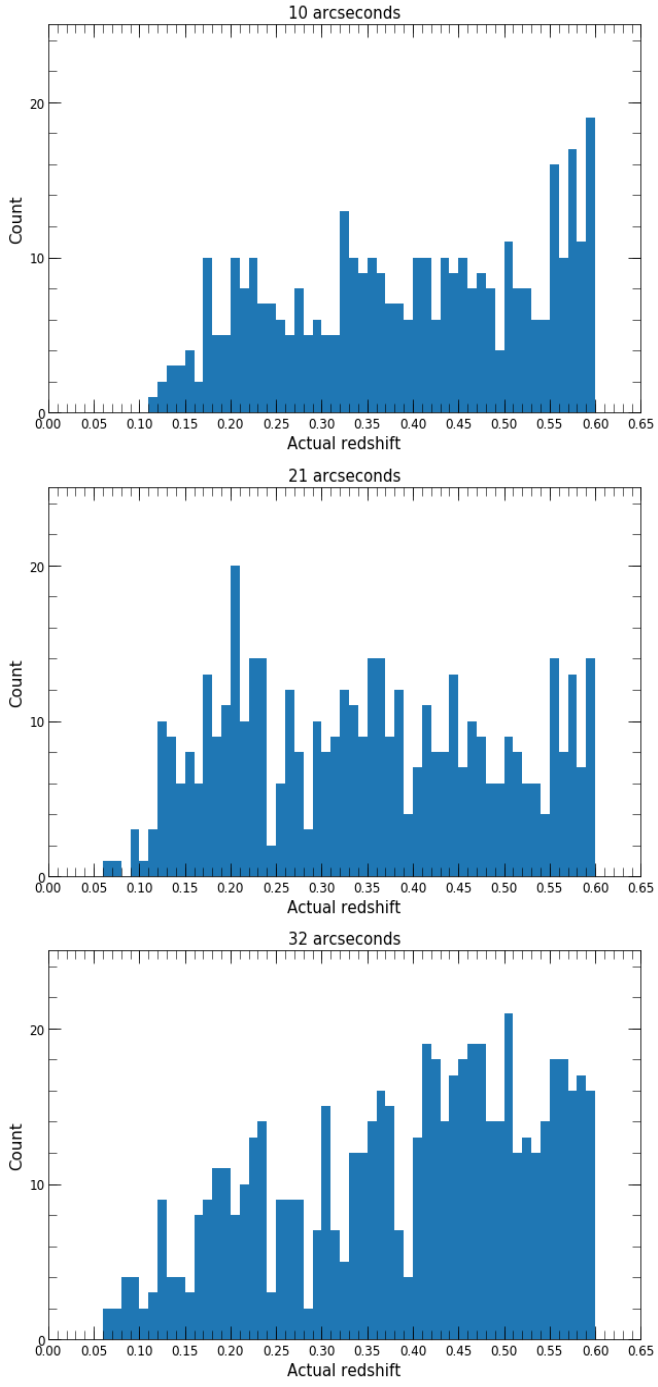


**Figure 3.19:** This figure displays frequency histograms of the 'actual' redshift distributions of clusters from the WNMR test set that had no bootstrap resamples returned within a 10 (top row), 21 (middle row) and 32 (bottom row) arcseconds search radius.

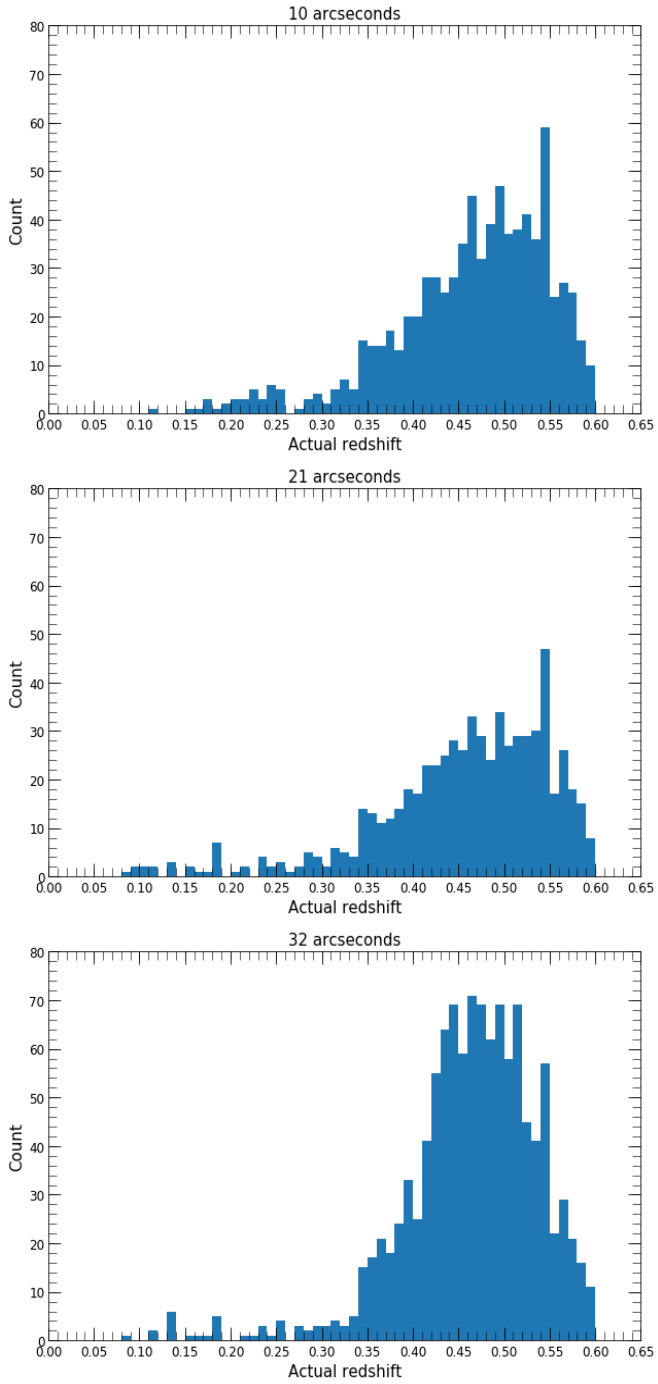


**Figure 3.20:** This figure is equivalent to Figure 3.19 except we examined the 'actual' redshift distributions of clusters that had no bootstrap resamples returned for the RNMW test set.

We immediately noticed that the overall accuracy of photometric redshift predictions was low when compared to the other test sets, as the tuned model constantly underestimated the photometric redshifts regardless of the search radius used. We also observed that the precision of the 95 per cent confidence intervals around predictions was poorly constrained, such that it would be difficult to distinguish clusters at high redshift from poorly constrained clusters at intermediate redshift.

In Figures S24, S25 and S26 (available online) we ran the tuned model on clusters at high redshift with low richness, which had a richness of twenty or fewer observed member galaxies and a redshift beyond the redshift training boundaries such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that had full bootstrap resamples returned at each search radius. Similar to the results in Figures S16, S17 and S18 for clusters at high redshift, we found that the overall accuracy of photometric redshift predictions was also low, as the tuned model constantly underestimated the photometric redshifts. In addition, the 95 per cent confidence intervals around predictions was also poorly constrained regardless of the search radius used.

In Figures S32, S33 and S34 (available online) we ran the tuned model on clusters with low richness, which had a richness of twenty or fewer observed member galaxies such that they did not qualify for the WNMR dataset, to obtain photometric redshift predictions that had full bootstrap resamples returned at each search radius. Similar to the results in Figures S8, S9 and S10 for clusters with low richness, we found that the overall accuracy of the photometric redshift predictions was high, as only a minority of cases had relatively large photometric redshift prediction errors. Although, we also observed that the precision of the 95 per cent confidence intervals became worse towards the redshift training boundaries when the search radius increased.

Lastly, we also evaluated the effectiveness from increasing the search radius on the performance of photometric redshift predictions and the number of clusters with full bootstrap resamples returned. For example, if a cluster did not have a photometric redshift estimate with full bootstrap resamples returned within a 10 arcseconds search radius, we would try using a 21 arcseconds search radius instead. From which, if a 21 arcseconds search radius was not sufficient, we would then try using a 32 arcseconds search instead. In Figures S40, S43, S46, S49, S52

and S55 we found that as the search radius increased the overall accuracy of photometric redshift estimates decreased. Although, this was still be beneficial rather than having clusters with no photometric redshift estimates at all. We also observed that as the search radius increased the number of photometric redshift estimates with full bootstrap resamples returned decreased as well. These trends can be seen repeating for all of the test sets.

## 3.4 Discussion

### 3.4.1 Effectiveness of Z-Sequence for photometric redshift estimation

In §§3.3.3 we employed samples from the WHL12 and redMaPPer cluster catalogues to examine the performance of the tuned model. From Figures 3.11, 3.12 and 3.13 it can be seen that majority of clusters in the WNMR test set were observed at low to intermediate redshifts, whereas from Figures 3.14, 3.15 and 3.16 it can be seen that majority of clusters in the RNMW test set were observed at intermediate redshift. This tells us that the methods used to estimate photometric redshifts in WHL12 and redMaPPer can significantly influence the resultant redshift distributions. Although, we found that the tuned model did not have much difficulty in working with either of these redshift distributions, as the overall performance of photometric redshift prediction errors for both test sets were similar. From which, we can infer that Z-Sequence can be effectively utilised across a wide range of redshifts if the appropriate training data is available.

For this work, we assigned the photometric redshifts of the WHL12 and redMaPPer cluster catalogues as 'actual' redshifts to examine the model performance on a large sample of clusters. Since we aimed to minimise data wastage, it was important to try to utilise all available clusters even though not all clusters had spectroscopic redshifts. We were aware that the 'actual' photometric redshifts for clusters in WHL12 and redMaPPer had a scatter of $\sim 0.01$ from spectroscopic redshifts. This was similar to the scatter in our photometric redshift prediction errors of $\sim 0.01$ from the 'actual' photometric redshifts, which

suggested that our model was as accurate as it could be based on the data used for training and testing. We expect that our photometric redshift prediction error would decrease if we trained on a large, entirely spectroscopic sample instead as the scatter associated with the photometric redshifts in the WHL12 and redMaPPer catalogues would be removed. In addition, it should be noted that the flaring seen in Figures 3.14, 3.15 and 3.16 lowered the predicted redshift values between 'actual' redshifts of $0.35 \geq z \geq 0.45$ for the RNMW test set. This was due to the flaring originating from redMaPPer itself and not from our algorithm, as it also occurs in Figures 7 and 9 of Rykoff et al. (2014).

In §§3.3.4 we tested the tuned model on clusters with unseen properties. We found that the tuned model performed well on clusters which had a similar parameter space to the MWAR training set and it also performed well on clusters of all richnesses within the redshift training boundaries. However, the tuned model performed poorly on clusters beyond the redshift training boundaries. This tells us that the performance of the tuned model was more dependent on the redshift of the cluster than the richness of the cluster. The tuned model was only effective on clusters at the redshift range it was trained for since we were limited to the redshift range of the majority of clusters available in SDSS. In addition, we observed an apparent feature seen at the lower and upper boundaries for predicted photometric redshifts in Figures 3.11, 3.12 and 3.13. We believe the cause of the apparent feature was due the nature of the machine learning algorithm itself. This was because the k-nearest neighbours algorithm calculates its prediction from the labels of the nearest neighbour examples in the training set when given an input data point, where the photometric redshift limits of the MWAR dataset was $0.0698 \leq z \leq 0.5986$ whilst the WNMR dataset was $0.05 \leq z \leq 0.599$. This meant that all photometric redshift predictions were bounded within the photometric redshift training range, such that clusters with 'actual' redshifts outside the boundaries could end up as part of the apparent feature. This explained why we did not observe the apparent feature in Figures 3.14, 3.15 and 3.16 as the photometric redshift limits of the RNMW dataset was $0.0811 \leq z \leq 0.5983$. As a further demonstration of the success of our algorithm, we note that the WNMR and RNMW test sets consisted of clusters found in one catalogue and not the other. This could mean that these clusters were more difficult to detect

and therefore potentially harder to assign a redshift value via other photometric redshift prediction methods, whereas our algorithm could estimate redshifts for the majority of these clusters. It should also be noted that the observed magnitude errors for all SDSS filters increased with redshift, as seen in Figure SA10 (available online). This meant it would be difficult for any empirical algorithm to make accurate photometric estimates in the high redshift regime. However, we expect our model would be successful at estimating photometric redshifts for high redshift clusters if trained on photometry data from imaging surveys such as LSST or *Euclid*, which will have greater photometric depths to increase the redshift limits of cluster detection when compared with SDSS.

We noticed in Table 3.3 that the median value of $|\Delta z|/(1+z)$ increased for the WNMR and RNMW test sets by 32 and 47 per cent when the search radius was enlarged from 10 arcseconds to 21 and 32 arcseconds respectively. This can also be seen in §§3.3.4 where the number of cases with accurate photometric redshift estimates decreased as the search radius increased, as a larger search radius was more likely to include interlopers. From Figures SA6, SA7 and SA9 (available online), we found that interlopers were evident in contaminating estimates with relatively large photometric redshift prediction errors if they appeared in the test set. Whilst Figures SA4, SA5 and SA8 (available online) indicated that interlopers were also somewhat present within the training set itself, as we found that some model predictions for clusters with no obvious interlopers in the test set still had relatively large photometric redshift prediction errors. Subsequently, we aimed to further improve the accuracy of the Z-Sequence model in future work by developing new strategies to constrain interlopers, such as with unsupervised machine learning techniques that can identify the presence of line-of-sight interloping galaxies and multiple projected line-of-sight clusters. This new method would be employed as an additional pre-processing tool to accompany the Z-Sequence model. From which, we could increase the size of the search radius once the obvious interlopers are removed and thus examine whether the photometric redshift prediction accuracy significantly improves if more cluster members are included. In addition, Figures SA6 and SA9 (available online) show that filter magnitude-cuts were also partially responsible for estimates with relatively large photometric redshift prediction errors, as we found that all of the galaxy members

in some cluster cores were removed from model predictions due to poor photometry measurements. Furthermore, we noticed in Figure SA7 (available online) that the 95 per cent confidence interval for the photometric redshift estimate involving the interloper became considerably wider in comparison to the photometric redshift estimates without the interloper. This suggests that the bootstrap confidence intervals could indicate whether interlopers were involved in the model prediction. Although, it should be noted that Figures 3.17 and 3.18 show that the majority of the model predictions seemed to employ relatively few galaxies for each search radius, such that it would be difficult to constrain interlopers in most instances. Moreover, by comparing the number of clusters that had photometric redshift estimates with full bootstrap resamples returned exclusively within each of the 10, 21 and 32 arcseconds search radii (see Figures S40, S43, S46, S49, S52 and S55 [available online]), we discovered that the majority of cases were actually within the 10 arcseconds search radius whereas only a minority of cases required an increase of the search radii to 21 and 32 arcseconds. This suggests that if we were to retrain the model on different surveys, we could consider not needing to employ multiple large search radii as the computational cost for training the model could outweigh the benefits gained.

It is worth noting that our approach resulted in photometric redshift predictions with full, partial and no bootstrap resamples returned. This was primarily due to the use of filter magnitude-cuts in each internal KNN model, which excluded galaxies with poorer photometric measurements from the cluster before any redshift predictions were made. Although, we observed in §§3.3.1 that applying filter magnitude-cuts could improve the overall accuracy of photometric redshift estimates. From which, we found that photometric redshift predictions with full bootstrap resamples returned were fairly accurate, as seen in §§3.3.3. However, it can also be seen that photometric redshift predictions with partial bootstrap resamples returned had low accuracy. This could be caused by the remaining bootstrap resamples not utilising strong predictive features. Subsequently, we advise that future photometric redshift estimates with partial bootstrap resamples returned should be flagged and used cautiously.

### 3.4.2 Practicality of the machine learning techniques used in this work

For this work, we were aware that the KNN algorithm can suffer from a dimensionality effect known as the 'curse of dimensionality' (Bellman, 1961). This can cause training samples to be disproportionately represented and sparsely distributed in high dimensional feature space, especially when the number of input features is greater than the number of training samples. As a consequence, this restricts the performance of machine learning algorithms due to the high complexity learning required. There are several approaches that can be used to limit the impact of this dimensionality effect, which include feature selection techniques (e.g. Sequential Feature Selection, Aha & Bankert 1995; Chi-Squared Test, Pearson 1900; Fisher's Score, Duda et al. 2001) and feature extraction techniques (e.g. Principal Component Analysis, Pearson 1901; Independent Component Analysis, Comon 1994; Partial Least Squares Regression, Wold 1983). These techniques promote useful features and ignore redundant features to subsequently constrain the dimensionality of a feature space. For a simple classification scenario, Raudys & Jain (1991) suggested that if the number of input features is not too large, such as between five to ten, then at least between fifty to one hundred corresponding training samples would be required per class to minimise the 'curse of dimensionality'. In our case, we ensured that the MWAR training set had a sufficient number of observations in the majority of redshift bins, as seen in Figure SA2 (available online). In addition, we preferred to use a feature selection method that employed features which maximised prediction accuracy rather than a feature extraction method that projected statistically significant features into a reduced feature space.

The most commonly used sequential feature selection strategies are SFS and Sequential Backward Selection (SBS, Aha & Bankert 1995). These methods are designed to be computationally efficient by searching through fewer combinations of feature space to provide a quasi-optimal solution rather than a global optimal solution. As described earlier in §§3.2.2.1, SFS iteratively adds features to an empty feature subset in a forward manner whilst SBS iteratively eliminates features from a full feature subset in a backward manner. This means that SBS will examine more high dimensional combinations of features when compared with

SFS, which could increase prediction accuracy but at a much higher computational cost. Nonetheless, we decided that SBS was not compatible for this work since the 95 per cent cluster retainment threshold would be immediately bypassed if all features were used at the same time, as seen in Figure 3.5. Although, we could consider SBS as an alternative feature selection strategy in imaging surveys that have greater filter sensitivity than SDSS. We also compared the performance between SFS and manual feature selection. From comparing Figures 3.4 and S1 (available online), we found that SFS selected features consistently performed better than the manually selected features for the KNN algorithm. This meant that SFS was more precise than manual feature selection at taking into account minor details in the datasets. From which, we decided that SFS had better synergy for working with bootstrap resamples in the SRKNN algorithm. It should be noted that we also randomly initialised the input features to the SRKNN algorithm as an additional starting step to SFS to reduce the impact from strong collinear features (see Figures SA11 and SA12 [available online]) during the feature selection process. Furthermore, in Figure 3.10 it can be seen that using a large number of bootstrap resamples for the SRKNN algorithm improved the stability for the relative frequency of SFS selected features. This was in contrast to using an individual algorithm such as the KNN algorithm (see Table 3.2) or using just one bootstrap resample in the SRKNN algorithm. This tells us that the SRKNN algorithm with a large number of bootstrap resamples was able to cope with minor changes to the training set, which would otherwise result in completely different features being used by the model.

The bias-variance tradeoff describes the generalisation performance of a supervised machine learning algorithm from fitting training data as a function of algorithmic complexity (Briscoe & Feldman, 2011). For instance, if an algorithm has tightly fit to the training data during learning, it will perform poorly on testing data. This results in many predictions with high variance and low bias. On the other hand, if an algorithm makes a lot of assumptions of the training data during learning, it will reduce the predictive power of the algorithm. This results in many predictions with high bias and low variance. The bias-variance tradeoff for the KNN algorithm varies depending on the number of nearest neighbours used, where using low values for the number of nearest neighbours can induce

overfitting whilst using high values for the number of nearest neighbours can induce underfitting (Geman et al., 1992). For the SRKNN algorithm, we examined a wide range of number of nearest neighbours from 1 to 25 but this range could be extended with increased computation in future work to explore using a larger number of nearest neighbours. In §§3.3.2 we had chosen a value for the number of nearest neighbours that showed no obvious indications of overfitting or underfitting. It is also known that ensemble algorithms can reduce the overall variance of predictions for a model by averaging estimates from multiple models that individually have high variance predictions (Bühlmann, 2012). This effect can be observed in the random forest (RF, Breiman 2001) algorithm, which is an ensemble that averages the estimates from multiple decision trees (DT, Breiman et al. 1984).

The main difference between the SRKNN and RF algorithms is the choice of internal model, such that each ensemble is better suited for different applications. The KNN algorithm utilises instance-based learning (Aha et al., 1991), which means it has no learnable internal parameters. Whilst the DT algorithm utilises partition-based learning (Strobl et al., 2009), which means it learns optimal splitting parameters for segmenting data. It should be noted that the KNN algorithm can support a similar partition strategy to the DT algorithm by utilising K-Dimensional Tree (Bentley, 1975) or Ball Tree search (Omohundro, 1989). Generally, the KNN algorithm provides higher flexibility for evaluating complex patterns whereas the DT algorithm has greater interpretability for understanding underlying decisions (Mohanapriya & Jayabalan, 2018). In Figures SA13, SA14, SA15, SA16, SA17 and SA18 (available online) we used the t-Distributed Stochastic Neighbour Embedding (t-SNE, van der Maaten & Hinton 2008) algorithm to visualise how the feature space of the MWAR training set appeared in two-dimensional space with and without feature scaling applied for each search radius. We observed that galaxies with similar photometric redshifts were somewhat clustered to form smooth transitions from low to intermediate redshifts when feature scaling was applied. Moreover, we also observed that galaxies with similar photometric redshifts were considerably dispersed across feature space when feature scaling was not applied. Nevertheless, the structure of these feature spaces would be difficult for the DT algorithm to apply partitions, whilst the

KNN algorithm is better suited to work with these smooth transitions, regardless of whether feature scaling is applied. This indicated that the SRKNN algorithm was more applicable at handling photometry data for estimating photometric redshifts than the RF algorithm.

We were also aware that the accuracy of photometric redshift estimates had a dependency on the accuracy of the cluster finder used to locate the cluster. For this work, we treated all input data points in CMS with uniform distance weighting. This meant that all input data points were not influenced by the distance to the training set data points. However, this may have reduced the accuracy of photometric redshift estimates in regions of the sky that had many line-of-sight interloping galaxies since the cluster finder would be unable to cleanly define the cluster core, where the red-sequence is most well-defined. To limit the dependency on the cluster finder, we could consider employing simple non-uniform weighting strategies for the SRKNN algorithm such as inverse distance weighting (Dudani, 1976). This computes weights based on the distance of the input data points to the training set data points, where the significance of the training set data points decreases as the distance increases. The reason we did not utilise this approach was due to the fact that it is also highly susceptible to noise in the training set. Although, in future work we could consider inverse distance weighting as an alternative, if we can further constrain line-of-sight interloping galaxies within the training set. In addition, the reason we did not utilise photometric redshift estimates of individual galaxies determined by SDSS itself is due to the fact that our method allows us to operate in situations where no photometric redshifts of individual galaxies are available.

In this work, we decided that ten-fold cross-validation was appropriate for feature selection and filter magnitude-cut analysis of the KNN algorithm, as the KNN algorithm had moderate computational training cost requirements. On the other hand, the SRKNN algorithm had higher computational training cost requirements especially when a large number of bootstrap resamples was used. From which, we decided that hold-out validation was more preferable for hyperparameter tuning of the SRKNN algorithm. However, with increased computation we could consider using k-fold or Monte Carlo cross-validation for hyperparameter tuning of the SRKNN algorithm in future work. In addition, we de-

cided that it was appropriate to utilise a grid search strategy to determine the optimal hyper-parameter settings for the SRKNN algorithm, since the number of hyper-parameter settings to examine for the SRKNN algorithm was relatively small.

## 3.5 Conclusion

We present Z-Sequence, an empirical model that is composed of an ensemble of the k-nearest neighbours algorithm, known as the sequential random k-nearest neighbours algorithm. The model makes use of photometry data from observed galaxies within a specified search radius to estimate photometric redshifts of clusters. In this proof-of-concept study, we assembled training sets with cross-matched clusters detected in SDSS by the WHL12 and redMaPPer cluster catalogues, where using cross-matched clusters reduced the likelihood of having false detections in the training set. Whilst clusters that were not cross-matched were used to test the performance of the model. We demonstrated that employing an automated feature selection strategy, known as sequential forward selection, was effective at identifying predictive features from an initial set of photometric features (i.e. filters and colours). We have shown that applying filter magnitude-cuts to the photometry data improved the overall accuracy of photometric redshift estimates, as this excluded galaxies with poor photometric measurements from model predictions. We examined the behaviour of each hyper-parameter setting for the SRKNN algorithm to understand how varying them affected model performance and generalisation. From which, we found that the choice of the number of nearest neighbours had the biggest impact, the choice of the number of initialised random features had moderate impact and the choice of the number of bootstrap resamples used had the least impact. The optimal values for each hyper-parameter setting were subsequently chosen for model testing. Our results showed that the tuned model performed well on clusters that were within the same redshift range (i.e. low and intermediate redshift) as the clusters in the training set and we also demonstrated that the tuned model was effective on clusters of all richnesses that were within the redshift training boundaries. We have shown the photometric

redshift prediction error of Z-Sequence via the median value of $|\Delta z|/(1+z)$ on the WHL12 test samples (across a photometric redshift range of $0.05 \leq z \leq 0.599$) to be 0.0106 and on the redMaPPer test samples (across a photometric redshift range of $0.081 \leq z \leq 0.598$) to be 0.0123 within a 10 arcseconds search radius, where the photometric redshift prediction error for both test samples increased by 32 and 47 per cent when the search radii was enlarged to 21 and 32 arcseconds respectively. In future work, we aim to apply our technique to imaging surveys as a tool to approximate redshifts for many clusters. It should be noted that our approach has no prerequisites which means that it is fully data driven. This is beneficial for photometric redshift estimation since Z-Sequence can be adapted to any imaging survey and trained on galaxy photometry data from known cluster positions in existing cluster catalogues. To prepare for upcoming surveys, we intend to run Z-Sequence as a complementary tool to our own Deep-CEE cluster finder to examine the entirety of the SDSS sky coverage in a preliminary data pipeline, where clusters detected directly from wide-field colour images would be accompanied with estimated photometric redshifts.

# Chapter 4

# AutoEnRichness: A hybrid empirical and analytical approach for estimating the richness of galaxy clusters

# Abstract

We introduce AutoEnRichness, a hybrid approach that combines empirical and analytical strategies to determine the richness of galaxy clusters (in the redshift range of $0.1 \leq z \leq 0.35$) using photometry data from the Sloan Digital Sky Survey Data Release 16, where cluster richness can be used as a proxy for cluster mass. In order to reliably estimate cluster richness, it is vital that the background subtraction is as accurate as possible when distinguishing cluster and field galaxies to mitigate severe contamination. AutoEnRichness is comprised of a multi-stage machine learning algorithm that performs background subtraction of interloping field galaxies along the cluster line-of-sight and a conventional luminosity distribution fitting approach that estimates cluster richness based only on the number of galaxies within a magnitude range and search area. In this proof-of-concept study, we obtain a balanced accuracy of 83.20 per cent when distinguishing between cluster and field galaxies as well as a median absolute percentage error of 33.50 per cent between our estimated cluster richnesses and known cluster richnesses within $r_{200}$. In the future, we aim for AutoEnRichness to be applied on upcoming large-scale optical surveys, such as the Legacy Survey of Space and Time and *Euclid*, to estimate the richness of a large sample of galaxy groups and clusters from across the halo mass function. This would advance our overall understanding of galaxy evolution within overdense environments as well as enable cosmological parameters to be further constrained.

**Supplementary material for this chapter can be found in 2022chanphdchapter4supplementary.pdf**

## 4.1 Introduction

Historically, in order to estimate the mass of clusters, researchers have regularly turned to optical surveys for determining cluster richness, where cluster richness can provide a proxy of cluster mass such that the number of galaxies within a cluster is expected to scale with cluster mass. For example, the Abell catalogue was the first comprehensive large scale catalogue to establish a measurement system for cluster richness, where cluster richness was defined as the number of galaxies counted within a specific radius and between two magnitude limits (i.e. the bright limit is the magnitude of the third brightest cluster galaxy whilst the faint limit is two magnitudes dimmer than the magnitude of the third brightest cluster galaxy). Similarly, the Zwicky catalogue established its own measurement system for cluster richness, where cluster richness was defined as the number of galaxies counted within an isopleth (i.e. the apparent boundary where the cluster density is twice that of the field density) and also between two magnitude limits (i.e. the bright end limit is the magnitude of the brightest cluster galaxy whilst the faint end limit is three magnitudes dimmer than the magnitude of the brightest cluster galaxy). We note that our definition of richness in this chapter is the number of cluster galaxies up to an absolute magnitude faint-end $r$ filter limit of $-20.5$ and within an $r_{200}$ radius.

In more recent times, a variety of automated methods have been developed that enable cluster mass or richness to be estimated without the need for extensive manual processing, such as utilising linking algorithms within redshift space (e.g. Huchra & Geller 1982; Yang et al. 2005; Calvi et al. 2011; Farrens et al. 2011; Wen et al. 2012; Tempel et al. 2016; Rodriguez & Merchán 2020), employing template fitting algorithms within colour-magnitude space (e.g. Postman et al. 1996; Kepner et al. 1999; Koester et al. 2007; Dong et al. 2008; Szabo et al. 2011; Rykoff et al. 2014) or training machine learning algorithms on observational/simulated measurements to indirectly estimate cluster mass (e.g. Ntampaka et al. 2019; Cohn & Battaglia 2019; Ho et al. 2019; Gupta & Reichardt 2020; Yan et al. 2020; de Andres et al. 2022; Lin et al. 2022).

Alternative approaches to determine cluster mass commonly include X-ray, caustic and weak gravitational lensing methods. From which, X-ray methods

assume that the intracluster gas within a cluster is under hydrostatic equilibrium in order to calculate the cluster mass required to produce the observed X-ray emissions, based on X-ray temperature and surface brightness measurements (e.g. Balland & Blanchard 1995; Ettori et al. 2013; Amodeo et al. 2016); caustic methods assume a cluster has spherical symmetry in order to calculate the cluster mass required to generate an estimated average escape velocity for cluster galaxies, based on galaxy position and velocity measurements (e.g. Diaferio & Geller 1997; Diaferio et al. 2005; Alpaslan et al. 2012); whilst weak gravitational lensing methods make no physical assumptions about a cluster to estimate the cluster mass required to produce the observed gravitational lensing of light from background objects, based on light distortion and magnification measurements (e.g. Hoekstra et al. 2013; van Uitert et al. 2016; McClintock et al. 2019). Although, these methods have somewhat time-consuming and expensive prerequisites (e.g. conducting deep X-ray observations, requiring complete spectroscopic analysis, obtaining high quality image data for performing weak gravitational lensing analysis), whereas methods involving optical photometry are typically quicker and cheaper to obtain and analyse the resultant data.

We note that determining cluster richness from the direct counting galaxies within a cluster is limited by the projection effect (Frenk et al. 1990; van Haarlem et al. 1997; Reblinsky & Bartelmann 1999; Costanzi et al. 2018; Sunayama et al. 2020), where van Haarlem (1996) estimated that approximately one third of the clusters in the Abell catalogue may have had their richnesses severely misestimated due to contamination from the projection effect. This effect arises when foreground or background galaxies are in the same line-of-sight as the cluster itself, which means it is difficult to accurately associate galaxies to a cluster unless spectroscopic redshifts for each galaxy are known. However, this is time-consuming especially when working with large sample sizes, as it is dependent on the preciseness of the distance measurement required.

In the literature, various statistical and non-statistical background subtraction methods have been utilised to address the projection effect when obtaining counts of cluster galaxies without the need for distance measurements. One typical way is to count the number of field galaxies within a known control field sample, which can be used as a direct reference to subtract a proportional number of

galaxies from a cluster's overall population to account for field galaxies (e.g. Kodama et al. 2001; Stott et al. 2007; Wylezalek et al. 2014). Another way is to define an annuli around the apparent outer perimeter of a cluster, which assumes that the annuli is far enough away to likely not contain cluster galaxies, such that a proportional number of galaxies can be subtracted from a cluster's overall population to account for field galaxies (e.g. Paolillo et al. 2001; Goto et al. 2003; Popesso et al. 2004). A further approach is to establish colour cuts for differentiating between cluster and field galaxies, where most of the galaxies within a cluster will appear to have similar colours especially if they are at the same redshift (i.e. red-sequence galaxies), whilst galaxies in the field will appear more randomised in terms of colour, especially if they are at different redshifts (e.g. Boué et al. 2008; Owers et al. 2017; Strazzullo et al. 2019). However, the approaches described here may not provide a robust or precise enough background subtraction, which is essential for accurately estimating cluster richnesses, due to these methods either being statistical or not assessing the true membership status of each cluster galaxy.

For this work, we describe in detail a novel hybrid method, nominally known as AutoEnRichness, to perform background subtraction and estimate cluster richnesses by employing a multi-stage machine learning algorithm and a conventional luminosity distribution fitting approach respectively. The first key stage of our hybrid method involves training the multi-stage machine learning algorithm to differentiate between cluster and field galaxies. This approach is completely data-driven to automatically capture underlying relationships for maximising the accuracy of cluster galaxy identification. The second key stage of our hybrid method involves learning the best fit parameters for a luminosity distribution fitting function to enable the estimation of cluster richness from the luminosity distribution of individual clusters. This approach has a strong theoretical basis that depends only on the brightness of the cluster galaxy population within a given search radius of a cluster. Our proposed strategy will be beneficial to provide researchers in the field with well-founded estimates of cluster richness as well as consistency and robustness against line-of-sight effects to mitigate severe contamination.

We present this chapter with the following structure. Firstly, in §4.2 we divide our methodology into five subsections, where §§4.2.1 describes the preparation of

a photometric dataset to train a background subtraction model; §§4.2.2 describes the mechanisms of a multi-stage machine learning algorithm that is used as our background subtraction model; §§4.2.3 describes our strategy for establishing a scaling relation to estimate $r_{200}$ of clusters; §§4.2.4 describes the preparation of a photometric dataset to train a luminosity distribution fitting function and §§4.2.5 describes the mechanisms of a luminosity distribution fitting function to estimate cluster richness. In §4.3 we outline our results across three subsections, where §§4.3.1 describes the model tuning analyses of our learned background subtraction model, scaling relation and luminosity distribution fitting function; §§4.3.2 describes the overall performance of our methodology on unseen clusters in various test sets and §§4.3.3 describes the importance of input features to our background subtraction model. Lastly, §4.4 discusses our findings and §4.5 summarises this work.

## 4.2 Methodology

A brief outline of our multi-stage method to estimate the richness of a cluster can be seen in Figure 4.1. From which, the following subsections will describe our workflow in more detail.

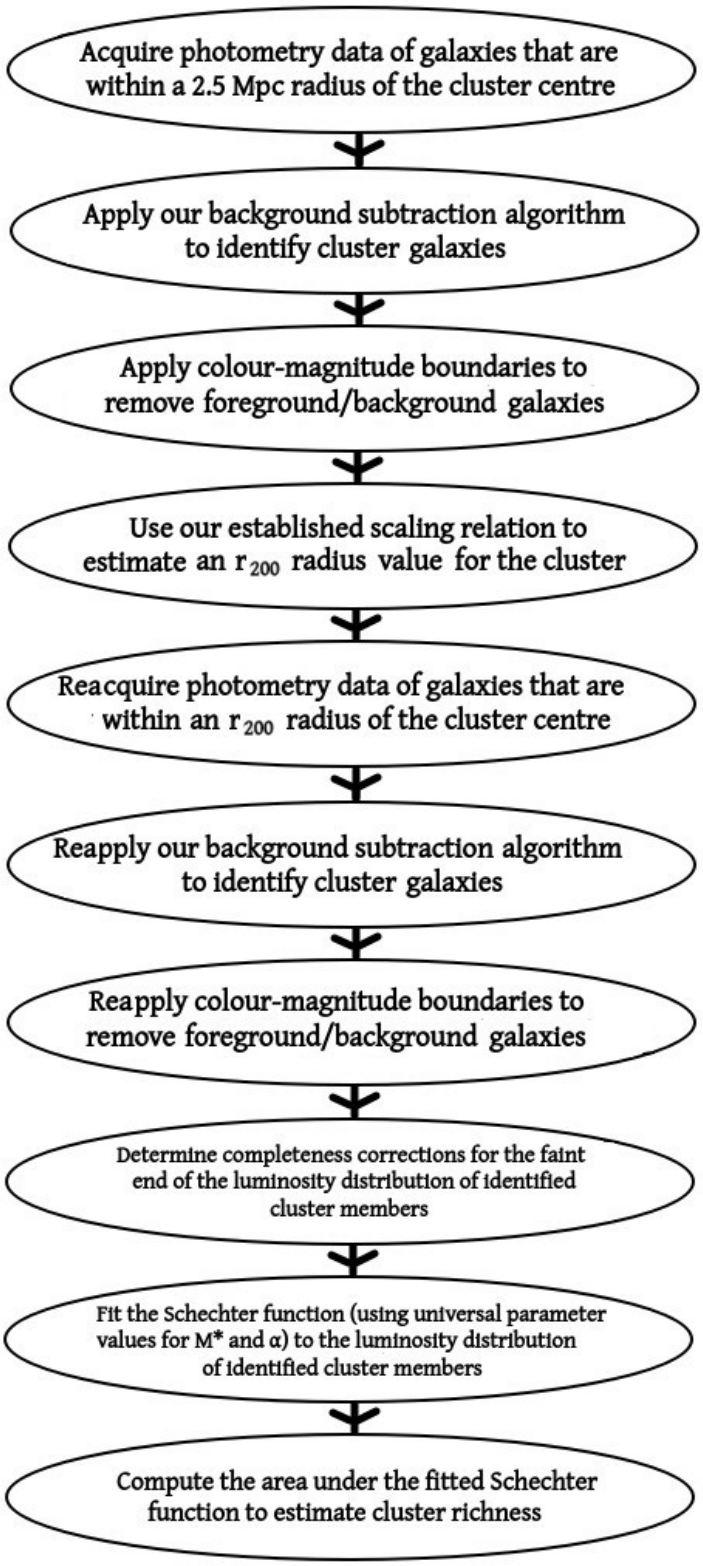


**Figure 4.1:** This figure shows a flowchart of the various steps in our multi-stage method to estimate the richness a cluster, where the start of the flowchart is the first step whilst the end of the flowchart is the last step.

### 4.2.1 Preparation of a photometric dataset to train a background subtraction model

To train our background subtraction model, we employed cluster galaxies that were identified by the AMF11 catalogue with an estimated photometric redshift between $0.1 \leq z \leq 0.35$. We note that the AMF11 catalogue applied matched filters[26] to galaxies observed in the Sloan Digital Sky Survey Data Release 6 (SDSS-II DR6, Adelman-McCarthy et al. 2008), where clusters were detected from maximising the likelihood of the matched filters whilst cluster galaxy membership identification was based on the proximity of the galaxy from the cluster center within $r_{200}$ as well as whether the likelihood difference (i.e. the difference in likelihood of detecting a cluster with and without the presence of the galaxy) was above a specified threshold. The reason we decided to use the cluster galaxies from the AMF11 catalogue was because they assessed the cluster membership status of each galaxy based on their contribution to a combination of various cluster profiles (i.e. radial surface density, luminosity and redshift). In addition, their selection method does not discriminate between 'blue' and 'red' cluster galaxies, which means it is representative of different galaxy types in clusters.

We cross-matched these cluster galaxies with galaxies observed in the Sloan Digital Sky Survey Data Release 16 (SDSS-IV DR16, Ahumada et al. 2020) to obtain the following fifteen features that are based on SDSS-IV DR16 photometry[27]: $u$, $g$, $r$, $i$, $z$, $u-g$, $g-r$, $r-i$, $i-z$, $u-r$, $g-i$, $r-z$, $u-i$, $g-z$ and $u-z$. For a cluster galaxy to be successfully cross-matched, the input astronomical coordinates must be within 1 arcsecond from the astronomical coordinates of a galaxy within SDSS-IV DR16 as well as satisfying additional observing flags. These flags are as follows: the observed object should be a 'primary' observation and must be classified as a galaxy object type by the SDSS photometric pipeline. We also ensured that our cluster galaxy sample only contained galaxies with a unique SDSS object identifier to prevent accidentally including galaxies that may have been selected multiple times within our search radius due to very

[26]The matched filters were constructed from modeling positional, brightness and redshift information of cluster and field galaxy distributions.

[27]We employed SDSS 'modelMag' measurements as well as full-sky dust reddening maps (Schlegel et al. 1998; Schlafly & Finkbeiner 2011) to account for galactic extinction.

small angular separation between overlapping line-of-sight galaxies and errors in the astrometry. In addition, we did not include cluster galaxies that were within 1646 arcseconds (i.e. 3 Mpc at $z = 0.1$) of a subsample (see §§4.2.3 for further details) of cross-matched[28] clusters from the WH15 catalogue and redMaPPer catalogue. This ensured that these clusters remained unseen for later usage in §§4.2.3. Furthermore, we applied a cut within colour-magnitude space (i.e. if greater than the $99.75th$ percentile in $r$ and $g - r$) to remove any cluster galaxies that still appeared to have spurious photometry. It should be noted that throughout this work, we used $r$ and $g - r$ to visualise cluster and field galaxies in colour-magnitude diagrams due to $g - r$ straddling the $4000\mathring{A}$ break of cluster galaxies in our working redshift range.

Correspondingly, we also required a field galaxy[29] sample to train our background subtraction model to differentiate between cluster and field galaxies. However, we were unable to find a sizable catalogue containing identified field galaxies. This meant that we had to manually search for 'field' regions that did not visually appear to contain clusters from the full WH15 and redMaPPer catalogues. This resulted in the identification of forty different 'field' regions, where the resultant astronomical sky map displaying the position of clusters and our proposed 'field' regions can be seen in Figure S1 (available online). We sampled galaxies from SDSS-IV DR16 that were within these 'field' regions. This involved applying a 1372 arcseconds (i.e. 2.5 Mpc at $z = 0.1$) search radius on each of the 'field' regions as well as reusing the same observing flags mentioned earlier within this section to obtain our field galaxy sample. The astronomical coordinates and number of observed field galaxies for each 'field' region are provided in Table 4.1. We did not include field galaxies that were within 10 arcseconds from the cluster galaxies in the AMF11 catalogue to remove cluster galaxies that may have accidentally been included as part of the field regions. We also did not include field galaxies that were within 1646 arcseconds (i.e. 3 Mpc at $z = 0.1$) of the

[28] This involved identifying clusters that were within 70 arcseconds (i.e. 250 kpc at $z = 0.225$) of each other in astronomical coordinate space and within $\pm 0.04(1 + z)$ (see Wen et al. (2009) for further explanation) of each other in redshift space. In addition, the clusters had to be observed within SDSS-IV DR16 between a redshift range of $0.1 \leq z \leq 0.35$.

[29] We refer to interloping galaxies along a clusters line-of-sight as field galaxies.

same subsample of cross-matched WH15 and redMaPPer clusters mentioned earlier within this section to ensure that the clusters remained unseen for later usage in §§4.2.3. Furthermore, we removed any field galaxies that were not within the same region of colour-magnitude space as our cluster galaxy sample, based on the observed minimum and maximum values for the cluster galaxies in $r$ and $g - r$. This was intended to encourage our background subtraction model to learn to be more proficient at classifying galaxies with similar photometric properties. Subsequently, this yielded a total of 83315 field galaxies that had the same fifteen photometry features as our cluster galaxy sample. For this work, we assumed that these field galaxies can be considered as 'actual' field galaxies.

We decided to set the redshift values of galaxies in our cluster and field galaxy samples to be based only on the photometric redshifts estimated by SDSS-IV DR16. This would enable a more straightforward comparison between the redshift distributions of both samples if they were measured via the same approach. We note that SDSS-IV DR16 applied the kd-tree nearest neighbor fit algorithm (see Csabai et al. (2007) for further details) to estimate the photometric redshifts of individual galaxies. We also used their estimated photometric redshifts to further constrain galaxies within our cluster galaxy sample to only be between a redshift range of $0.1 \leq z \leq 0.35$, whereas galaxies within our field galaxy sample were not redshift restricted to mimic field galaxies appearing along the line-of-sight of clusters. Although, we note that galaxies were not required to have photometric redshifts available to be included in our field galaxy sample. In addition, we computed the $r$ filter absolute magnitudes for the cluster and field galaxies based on their photometric redshifts and corresponding K corrections[30].

We note that our background subtraction model will learn to identify all cluster galaxies between a redshift range of $0.1 \leq z \leq 0.35$, which may result in overcounting of cluster galaxies within a cluster if there are other clusters along the line-of-sight. To limit this effect, we decided to establish colour-magnitude boundaries within colour-magnitude space when applying our background subtraction model. These boundaries are designed to capture the majority of the

[30]In order to estimate the amount of K correction required, we performed linear interpolation between redshift and $r$ filter K corrected values determined from a simple stellar population model (see Bruzual & Charlot (2003) for further details).

| **Right ascension (degrees)** | **Declination (degrees)** | **Number of observed galaxies** |
|---|---|---|
| 5.10408 | 21.0611 | 4488 |
| 5.77875 | 2.30955 | 7049 |
| 9.14699 | 33.7121 | 4030 |
| 10.4395 | -3.84934 | 5377 |
| 21.3533 | -3.02485 | 5595 |
| 23.0452 | 30.3202 | 5664 |
| 26.0012 | 23.8818 | 5024 |
| 28.3683 | 20.8955 | 2748 |
| 29.8579 | 11.0895 | 5108 |
| 38.6421 | -8.10943 | 5168 |
| 113.919 | 28.0332 | 5223 |
| 115.323 | 15.458 | 6310 |
| 121.916 | 0.109439 | 5518 |
| 129.971 | 51.2851 | 4753 |
| 143.518 | 63.0839 | 4564 |
| 144.097 | 47.3992 | 4929 |
| 158.633 | 56.6154 | 5982 |
| 166.196 | 20.0867 | 5861 |
| 177.951 | 65.8863 | 5733 |
| 196.86 | 15.2355 | 5885 |
| 207.127 | 65.0462 | 5711 |
| 214.632 | 3.77235 | 6109 |
| 219.239 | 62.0499 | 6009 |
| 226.958 | 54.9023 | 5346 |
| 231.26 | -0.0179 | 6696 |
| 235.973 | 18.662 | 7669 |
| 236.823 | 39.221 | 6421 |
| 238.422 | 58.5216 | 6760 |
| 255.233 | 18.8401 | 6079 |
| 263.834 | 28.0521 | 3565 |
| 316.524 | -6.36219 | 6273 |
| 316.54 | -1.56952 | 5201 |
| 326.258 | -6.56724 | 6597 |
| 332.263 | 28.8328 | 4580 |
| 332.498 | 19.5978 | 4043 |
| 333.873 | 24.14 | 4214 |
| 340.098 | 4.71022 | 4252 |
| 353.561 | 33.6502 | 4239 |
| 357.0789 | -4.69765 | 5103 |
| 359.369 | 17.4744 | 5809 |

**Table 4.1:** This table contains the astronomical coordinates (J2000) and number of observed galaxies that were sampled from our forty different proposed 'field' regions using a 1372 arcseconds (i.e. 2.5 Mpc at $z = 0.1$) search radius. We note that the number of observed galaxies does not include field galaxies that were within 10 arcseconds of the galaxies in our cluster galaxy sample nor did we include field galaxies that were within 1646 arcseconds (i.e. 3 Mpc at $z = 0.1$) of a subsample of cross-matched WHL and redMaPPer clusters.

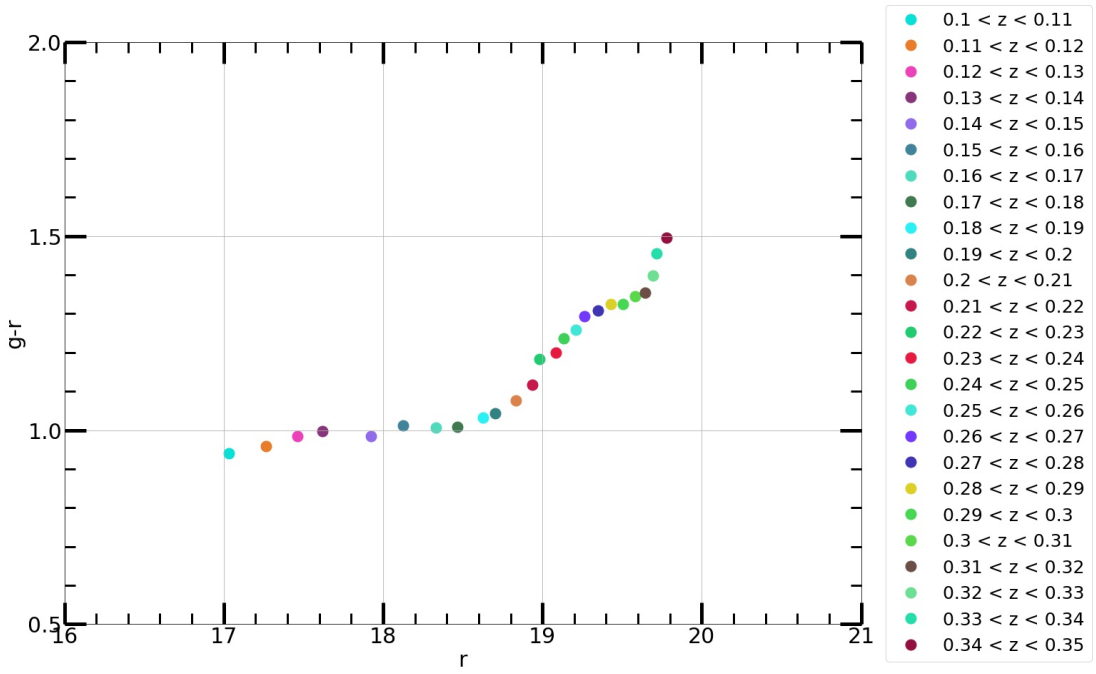


**Figure 4.2:** This figure shows a colour-magnitude diagram (using apparent magnitudes) of the median $r$ and $g-r$ for cluster galaxies at different redshift intervals from our cluster galaxy sample.

population of cluster galaxies at specific redshifts. We first computed the median values of $r$ and $g-r$ for cluster galaxies in our cluster galaxy sample across redshift intervals of $\pm 0.005$ that are centered in redshift bins from 0.105 to 0.345 with step sizes of 0.01, as shown in Figure 4.2. We then manually determined appropriate lower and upper boundaries of $r_{median} - 0.01 \leq r_{median} \leq r + 0.4$ and $g - r_{median} - 0.05 \leq g - r \leq g - r_{median} + 0.4$ for each redshift bin. This would result in 'L-shaped' boundaries around the cluster galaxies at a given redshift, where an example of the 'L-shaped' boundaries for cluster galaxies at $z = 0.225$ is shown in Figure 4.3. We then applied these colour-magnitude boundaries to our cluster galaxy sample across redshift bin sizes of 0.01 to remove any cluster galaxies that were not within the colour-magnitude boundaries at their respective redshift. Subsequently, this yielded a total of 60663 cluster galaxies that were available to train our background subtraction model. For this work, we assumed that these cluster galaxies can be considered as 'actual' cluster galaxies.

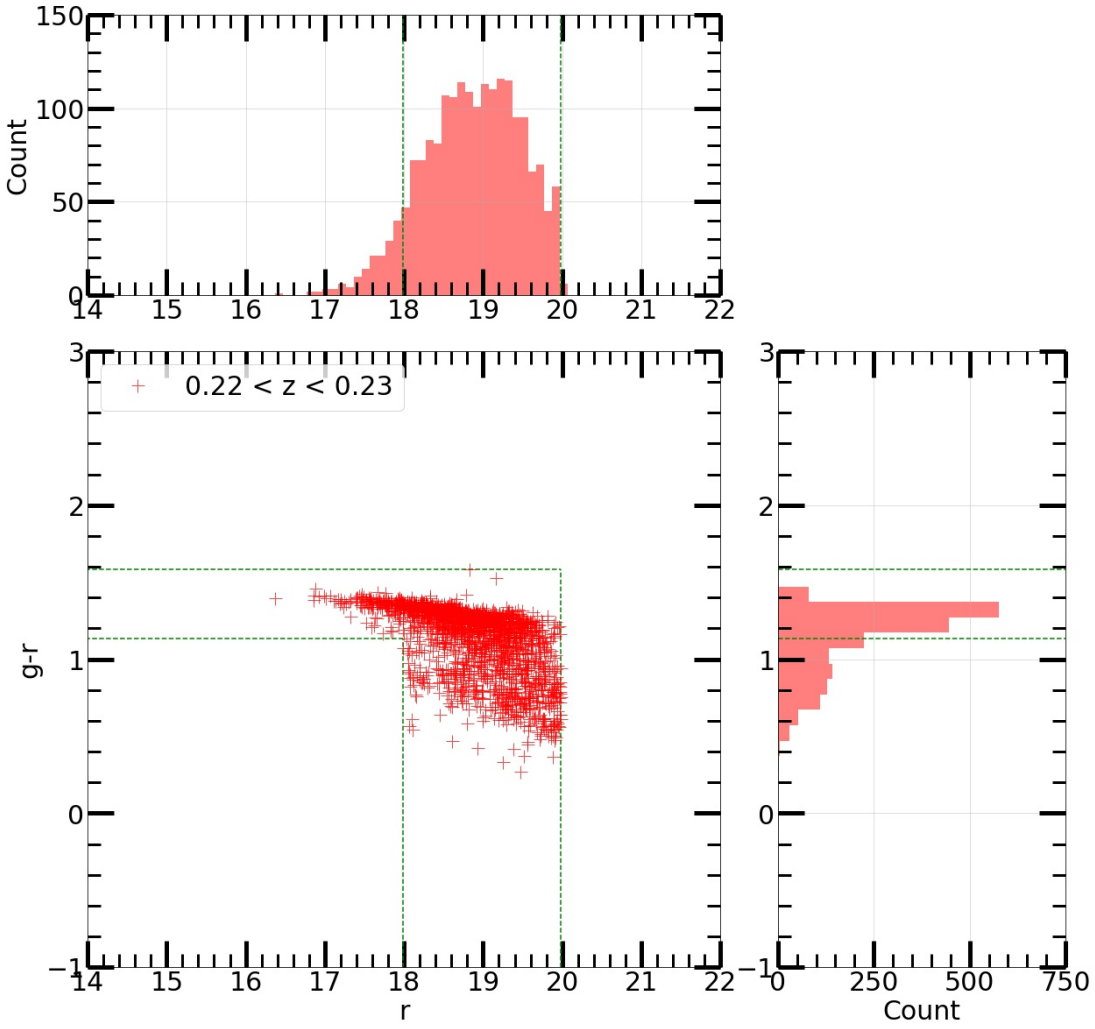


**Figure 4.3:** This figure shows an example of the colour-magnitude boundaries (green dotted lines) for cluster galaxies (red cross) between $0.22 < z < 0.23$ from our cluster galaxy sample, where only galaxies that are between the colour-magnitude boundaries will be considered as part of a cluster at that redshift.

In Figure S2 (available online), it can be seen that our cluster and field galaxies were taken from different areas across SDSS-IV DR16. This meant that our cluster and field galaxy samples were likely to be representative of the whole population of cluster (between a redshift range of $0.1 \leq z \leq 0.35$) and field galaxies. Moreover, in Figures 4.4, S3 and S4 (available online), it can be seen that our field galaxy sample had an overall noticeable disparity to our cluster galaxy sample within colour-magnitude space. This somewhat validated our approach for obtaining the field galaxies given the underlying differences in photometry between the majority of the cluster and field galaxies. Although, we also observed some overlap of the 'blue' and faint cluster galaxies with bright field galaxies. We expect that it may be more difficult for our background subtraction model to differentiate between the galaxy classes within these overlap regions of colour-magnitude space. Furthermore, in Figure 4.5 we display the photometric redshift, $r$ filter apparent magnitude and $r$ filter absolute magnitude distributions of galaxies in our cluster and field galaxy samples. It can be seen that we had fewer cluster and field galaxies at lower redshifts when compared to those at higher redshifts. This indicated that we would need to sample equally across different redshifts to prevent our background subtraction model from being biased towards any particular redshift. We note that the number of field galaxies decreased significantly after $z = 0.4$ due to the observing limitations of SDSS-IV DR16 at higher redshifts, where SDSS-IV DR16 had a $r$ filter limiting magnitude of 22.2. We also noticed that there was a gradual drop in the number of cluster and field galaxies towards fainter magnitudes due to the incompleteness of cluster galaxies in the AMF11 catalogue and observing limitations of SDSS-IV DR16 respectively.

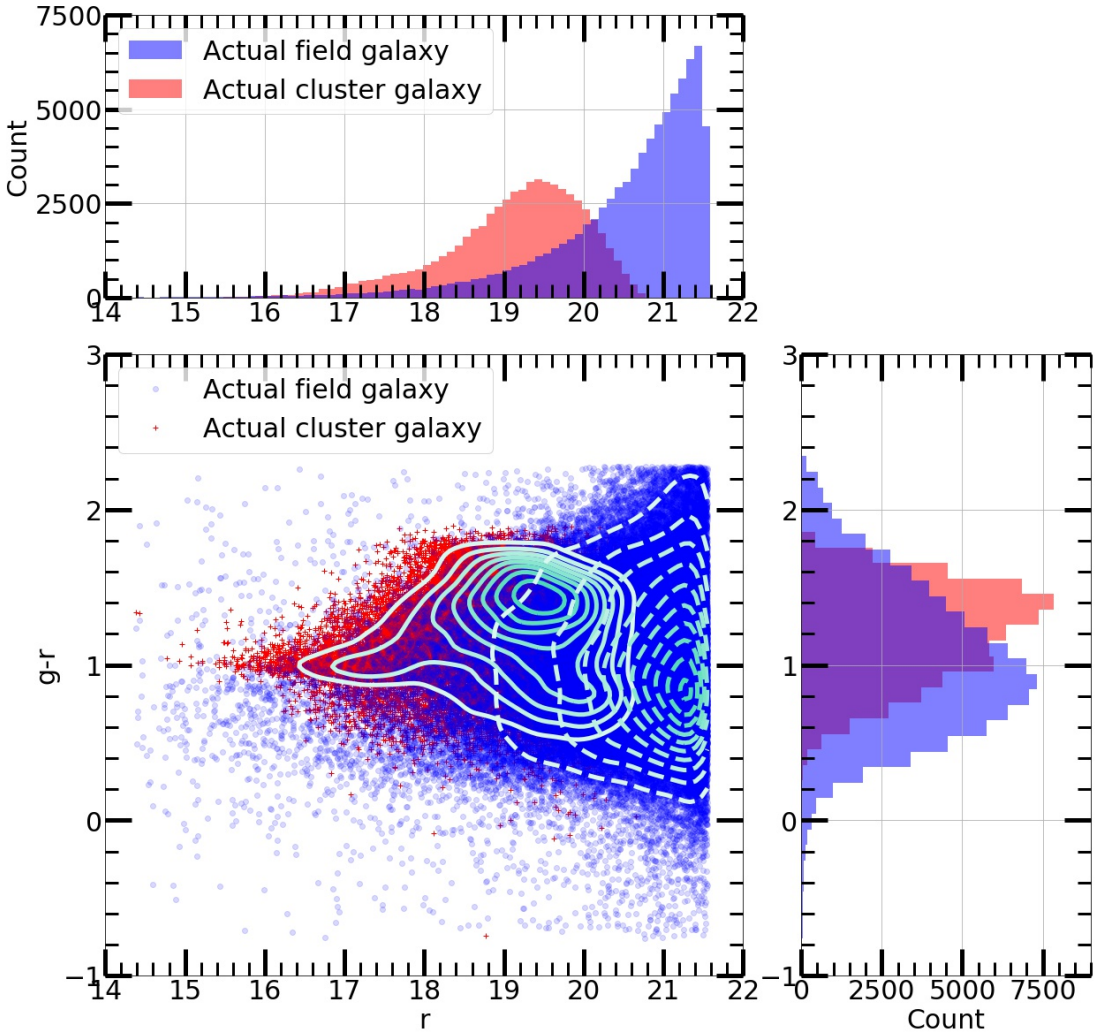


**Figure 4.4:** This figure shows colour-magnitude diagrams (using apparent magnitudes) of the cluster (red cross) and field (blue circle) galaxies in our cluster and field galaxy samples that were observed within SDSS-IV DR16. The non-dashed contour lines represent the density of data points for cluster galaxies whilst the dashed contour lines represent the density of data points for field galaxies.

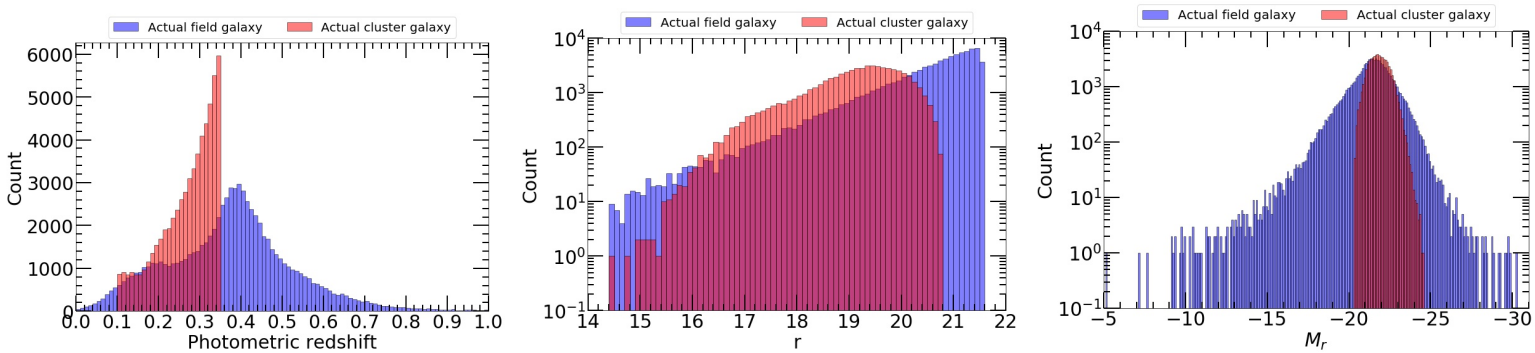


**Figure 4.5:** This figure shows histograms of the photometric redshift (left image), $r$ filter apparent magnitude (middle image) and $r$ filter absolute magnitude (right image) of galaxies in our cluster (red) and field (blue) galaxy samples after being cross-matched with galaxies observed within SDSS-IV DR16, where the cluster galaxies had to be between a redshift range of $0.1 \leq z \leq 0.35$. It should be noted that we only display field galaxies that had an available photometric redshift in the top and bottom images.

Finally, we partitioned our cluster and field galaxy samples into three different subsets, known as the training, validation and test sets. In particular, the training set would be used to train our background subtraction model, the validation set would be used to tune its hyper-parameters and the test set would be used to obtain an unbiased estimate of the predictive performance of our background subtraction model. This involved randomly selecting 450, 150 and 150 cluster galaxies within fixed redshift bin sizes of 0.01 across a redshift range of $0.1 \leq z \leq 0.35$ to be within our training, validation and test sets, which resulted in a total of 11250, 3750 and 3750 cluster galaxies respectively. We also randomly selected 33750, 11250 and 11250 field galaxies to be within our training, validation and test sets respectively. It should be noted that we applied sampling weights[31] when selecting field galaxies to be within our training and validation sets to ensure that the $r$ filter apparent magnitudes of the field galaxies overlapped with the $r$ filter apparent magnitudes of the cluster galaxies. This would expose our background subtraction model to a larger proportion of the more difficult instances (i.e. cluster and field galaxies that had very similar photometry) during its training. Furthermore, we wanted our training, validation and test sets to remain as realistic as possible. As such, we permitted the number of field galaxies to outnumber (i.e. we assumed that having three field galaxies for every cluster galaxy was appropriate) the number of cluster galaxies with these sets. Although, we only permitted random sampling (i.e. equal sampling weightage) of field galaxies in our test set. These properties can be seen in Figure 4.6.

[31] The amount of sampling weightage applied to each field galaxy in our training and validation sets was based on the resultant likelihood of the $r$ filter apparent magnitude for the field galaxy under a normal distribution that was constructed from the mean and standard deviation of the $r$ filter apparent magnitudes of cluster galaxies in our training and validation sets. We also shifted the computed means by $-1$ in our training and validation sets to ensure that the cluster and field galaxy distributions overlapped at all $r$ filter apparent magnitudes. In addition, we note that sampling with replacement was used when selecting field galaxies to be within our training and validation sets.

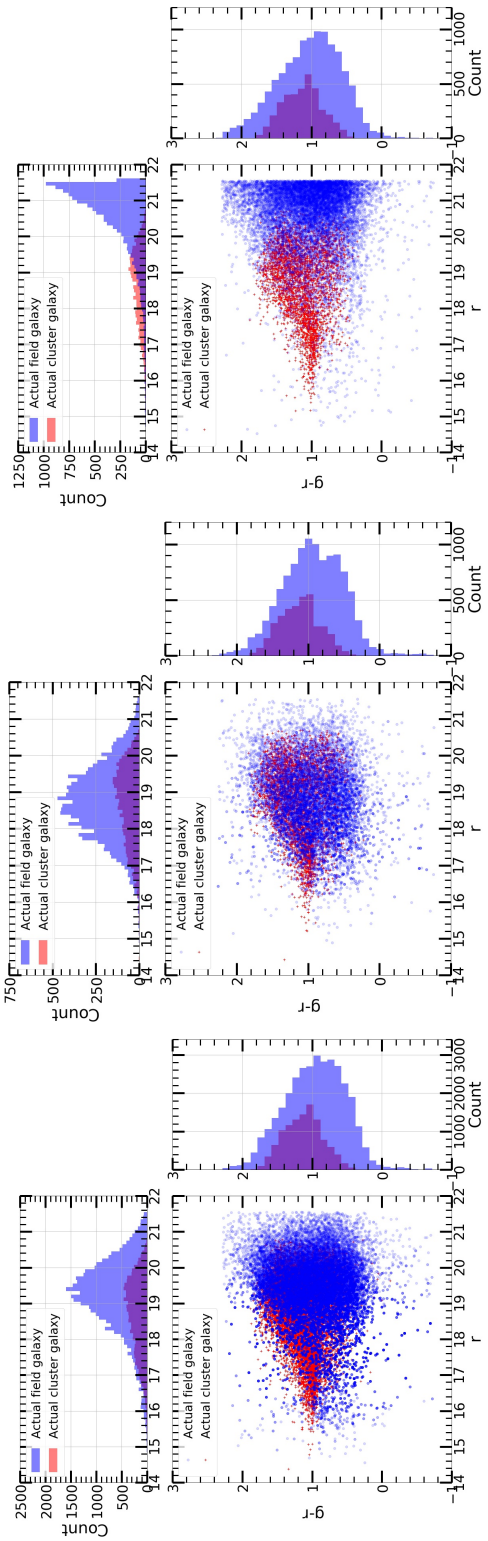


**Figure 4.6:** This figure shows colour-magnitude diagrams (using apparent magnitudes) of the cluster (red cross) and field (blue circle) galaxies in our training (left image), validation (middle image) and test (right image) sets that were observed within SDSS-IV DR16.

### 4.2.2 Using a multi-stage machine learning algorithm to perform background subtraction

We employed an unsupervised deep learning algorithm, known as an autoencoder (AE, Rumelhart et al. 1985), as the first stage of our background subtraction model. Our overall objective for using an AE is to train it to learn to accurately reconstruct input data. The mechanism behind the AE can be separated into three main stages, that are known as the encoder network, bottleneck and decoder network. The overall architecture for a typical AE is shown in Figure 4.7.

The encoder network is composed of fully-connected layers that are responsible for processing an input dataset by performing nonlinear transformations of the input data into a compressed representation. This is achieved by decreasing the number of nodes in the fully-connected layers as the size of the encoder network increases. The compression is maximised within the bottleneck, where the number of nodes in the bottleneck determines the amount of compression. The underlying objective of the bottleneck is to obtain the lowest dimensional representation of the data that captures the most generalisable aspects about the data. From which, the compressed data is then passed to the decoder network for reconstruction. This involves decompressing the compressed data back into its original input dimensionality by increasing the number of nodes in the fully-connected layers as the size of the decoder network increases. If the AE is properly trained, the reconstructed feature values should closely resemble the feature values of the input data. Overall, an AE can be considered as a type of dimensionality reduction-based algorithm since it focuses on reducing the dimensionality of the input data. However, we reconfigure its functionality from a dimensionality reduction-based algorithm into an outlier detection algorithm by also examining the differences between the reconstructed outputs and the input data. We note that the decoder network has the same but reversed architecture to the encoder network, where the number of nodes in the fully-connected layers increases rather than decreases as the size increases.

In order to train the AE to generate accurate reconstructions, we used the mean squared error as our loss function. This measured the similarity between

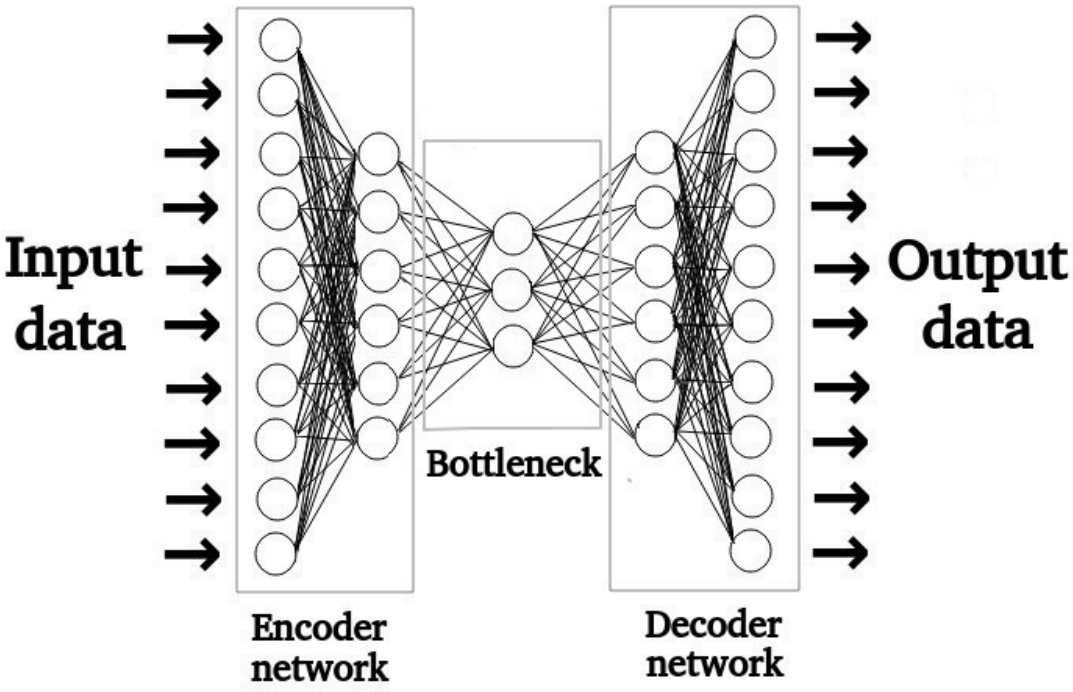


**Figure 4.7:** This figure shows an example of the architecture layout for a typical AE. The AE is composed of three main stages that are known as the encoder network, bottleneck and decoder network, where the nodes in each hidden layer are fully-connected to the nodes of the adjacent hidden layers. We also employed a ReLU activation function with 'He uniform' (He et al., 2015) weight initialisation for each hidden layer in the encoder network, bottleneck and decoder network, whilst a linear activation function with 'Glorot uniform' (Glorot & Bengio, 2010) weight initialisation was used for the output layer of the decoder network. In addition, we initialised all biases to zeros. It should be noted that we utilised the Keras deep learning framework (Chollet et al., 2015) to construct the AE.

all of the input and reconstructed feature values of galaxies via the following equation:

$$\text{Mean Squared Error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ , \tag{4.1}$$

where $n$ is the number of input features, $y$ is the input feature values and $\hat{y}$ is the reconstructed feature values.

We set the batch size, learning rate, optimiser algorithm[32] and architecture layout[33] to be tunable hyper-parameters, where the full hyper-parameter search space is shown in Table 4.2.

We employed a separate machine learning algorithm, known as logistic regression (see Morgan & Teachman (1988) for further details), as the second stage of our background subtraction model. This served to convert the outputs of the AE into class predictions. In particular, we used the known class labels as the target variable and the mean squared error between the input and reconstructed feature values as the input variable, where if an input was poorly reconstructed by the AE then the corresponding mean squared error will be large too. From which, the logistic regression algorithm determines whether a galaxy should be classified as a cluster or field galaxy (i.e. the galaxy class with the higher predicted probability) when given the mean squared error of each galaxy. In this work, we decided to use the defaulted hyper-parameter values for the logistic regression algorithm (N.B. without regularisation) in the Scikit-Learn machine learning library since we primarily wanted to examine the influence of the AE in our background subtraction model. We expect that tuning the hyper-parameters for the logistic regression algorithm may slightly improve the overall predictive performance of our background subtraction model but this can be explored further in future work. It should be noted that the logistic regression algorithm minimised the following loss function during its training:

[32] We recommend the reader to refer to Ruder (2016) for an overview of different optimiser algorithms.

[33] We considered the number of nodes in the bottleneck to be the most significant component of an AE's architecture since it is influential in the amount of generalization learned. We decided that the number of nodes in the bottleneck should be a tunable hyper-parameter whilst the number of nodes for the hidden layers in the encoder and decoder networks would remain fixed.

| **Tunable hyper-parameter name** | **Hyper-parameter search space** |
|---|---|
| Batch size | 256 or 512 or 1024 or 2048 |
| Learning rate | 0.0001 or 0.001 or 0.01 or 0.1 |
| Optimiser algorithm | Adaptive Moment Estimation (Adam) or Adaptive Delta (Adadelta) or Adaptive Gradient Optimiser (Adagrad) or Adam Based On The Infinity Norm (Adamax) or Adam With Nesterov Momentum (Nadam) or Stochastic Gradient Descent (SGD) or Root Mean Squared Propagation (RMSprop) |
| Architecture layout (number of nodes and hidden layers in the encoder network and bottleneck) | **1** (13 nodes in first hidden layer, 11 nodes in second hidden layer, 9 nodes in third hidden layer, 7 nodes in the fourth hidden layer and 1 node in the bottleneck) or **2** (13 nodes in first hidden layer, 11 nodes in second hidden layer, 9 nodes in third hidden layer, 7 nodes in the fourth hidden layer and 3 nodes in the bottleneck) or **3** (13 nodes in first hidden layer, 11 nodes in second hidden layer, 9 nodes in third hidden layer, 7 nodes in the fourth hidden layer and 5 nodes in the bottleneck) |

**Table 4.2:** This table contains a list of tunable hyper-parameters for the AE as well as the range of values that were explorable in the hyper-parameter space via random search. We also set a maximum of ten thousand trainable epochs as well as enabling early stopping of the model training if the validation loss had not decreased by 0.001 over fifty epochs from the best observed validation loss. Furthermore, we again remind the reader that the encoder and decoder networks had reversed symmetrical designs, so we did not specify the number of nodes or hidden layers for the decoder network within this table.

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n} p_i^* log(p_i) + (1 - p_i^*) log(1 - p_i) \ , \tag{4.2}$$

where $n$ is the number of inputs, $p^*$ is the true class value (i.e. either 0 or 1) and $p$ is the predicted probability (i.e. between 0 and 1) of being a galaxy class. This loss function measured the difference between predicted probability and true class value of galaxies.

We utilised a random search strategy to examine the predictive performance of our background subtraction model with different hyper-parameter combinations, where random search is a computationally efficient approach that does not need to examine every hyper-parameter combination. Instead, it considers that hyper-parameter optimization can be characterised by a Gaussian process, such that only a minority of hyper-parameter combinations are actually important. For example, we can assume that a randomly selected hyper-parameter combination has a ninety-five per cent probability of being situated within the top five per cent of all possible hyper-parameter combinations from the optimum after conducting only sixty iterations of random search. At the same time, we employed a Monte Carlo cross-validation strategy to examine the variability of the predictive performance of our background subtraction model with different weight initialisations and dataset compositions. This involved repeated random sampling of new training, validation and test sets over ten iterations to measure the average predictive performance across the ten iterations. Ideally, we aimed to select a hyper-parameter combination that offered consistency and good predictive performance.

To determine the optimal hyper-parameter combination of our background subtraction model, we computed the area under a precision-recall curve (AUCPR, Boyd et al. 2013) for each hyper-parameter combination by applying the trapezium rule to the following equation:

$$\text{AUCPR} = \int \text{Precision} \, d\,(\text{Recall}) \ , \tag{4.3}$$

where the formulae for Precision and Recall can be found in Equations 2.5 and 2.6 respectively. Although, we note that in this case $TP$ is the number

of correctly classified 'actual' cluster galaxies, $FP$ is the number of incorrectly classified 'actual' cluster galaxies and $FN$ is the number of incorrectly classified 'actual' field galaxies. Briefly, this metric measured the proportion of predictions that were predicted as cluster galaxies as well as the proportion of 'actual' cluster galaxies that were recovered across all class probability thresholds. It is ideal for assessing the predictive performance of a model that focuses on correctly identifying 'rare' instances (i.e. when there is a class imbalance). The optimal hyper-parameter combination would maximise the AUCPR for galaxies in our validation set.

Next, we determined the corresponding optimal class probability threshold when using the optimal hyper-parameter combination. This involved comparing the F1 score yielded for each class probability threshold (i.e. from 0 to 1 with class probability threshold step sizes of 0.01) via Equation 2.7.

This metric was similar to AUCPR in functionality except it only considered the predictive performance at a specific class probability threshold. The optimal class probability threshold would maximise the F1 score for galaxies in our validation set.

Lastly, we determined the overall classification accuracy of our background subtraction model at distinguishing between cluster and field galaxies in our test set. This involved computing the balanced accuracy (Brodersen et al., 2010) when using the optimal class probability threshold and optimal hyper-parameter combination via the following equation:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) , \tag{4.4}$$

where $TP$ is the number of correctly classified 'actual' cluster galaxies, $TN$ is the number of correctly classified 'actual' field galaxies, $FP$ is the number of incorrectly classified 'actual' cluster galaxies and $FN$ is the number of incorrectly classified 'actual' field galaxies. The primary advantage of using balanced accuracy rather than conventional classification accuracy is that balanced accuracy takes into account class imbalance whereas conventional classification accuracy assumes equal class sizes when measuring the predictive performance of a binary classification model.

### 4.2.3 Establishing a scaling relation to estimate $r_{200}$

It is beneficial to measure the richness of clusters within a characteristic radius (e.g. $r_{200}$, $r_{500}$, $r_{2500}$) because it enables a more straightforward comparison of cluster richness across different catalogues. We decided to establish a scaling relation that predicted values for the characteristic radius of cross-matched WH15 and redMaPPer clusters from §§4.2.1. We note that WH15 used a friends-of-friends grouping algorithm on galaxies with known spectroscopic or photometric redshifts to identify clusters in the Sloan Digital Sky Survey Data Release 12 (SDSS-III DR12, Alam et al. 2015) whilst redMaPPer used a red-sequence fitting algorithm on galaxies within colour-magnitude space to identify clusters in SDSS-III DR8. Subsequently, we obtained a total of 6064 cross-matched WH15 and redMaPPer clusters between a redshift range of $0.1 \leq z \leq 0.35$. We decided to use these clusters because they were found via two conventional approaches for cluster detection. This enabled us to directly compare the consistency of richness estimates from using our novel cluster galaxy identification technique versus other cluster galaxy identification techniques. In this proof-of-concept study, we choose to employ only a subsample of 1000 cross-matched WH15 and redMaPPer clusters when creating our scaling relation for time efficiency.

We also decided to use $r_{200}$ as our characteristic radius since the cluster galaxies in our training and test sets from the AMF11 catalogue were originally sampled within $r_{200}$. In particular, we utilised $r_{200}$ values that were estimated by WH15 as the dependent variable in our scaling relation, where their $r_{200}$ estimates were computed via a scaling relation between $r_{200}$ measurements from X-ray/weak gravitational lensing observations and total luminosity in the $r$ band of all the identified cluster galaxies. For this work, we assumed that these $r_{200}$ values can be considered as 'actual' $r_{200}$ values. In Figure S5 (available online), we noticed that there was a strong linear relationship between WH15[34] and redMapper richness[35]. This means that we can directly compare our predicted richnesses with

[34]We refer to the $R_{L*}$ variable from the WH15 catalogue as WH15 richness, where they computed cluster richness by measuring the total luminosity of identified galaxy members as a function of the typical luminosity of galaxies in the $r$ filter.

[35]We refer to the $\lambda/S$ variable from the redMaPPer catalogue as redMaPPer richness, where they computed cluster richness by determining an expected richness which would yield the observed projected density, $i$ filter magnitudes and multiple colour indices of the identified

the richness estimates of redMaPPer. Furthermore, we noticed that there was a non-linear relationship between $r_{200}$ and both WH15 and redMaPPer richnesses which was in accordance with the empirical richness-size relation observed in Hansen et al. (2005), where our 'actual' $r_{200}$ values appeared to have greater variability at lower richnesses. As such, we expected that our scaling relation would have greater variability in $r_{200}$ at lower richnesses too.

We then partitioned the cross-matched WH15 and redMaPPer clusters into a training set and test set. We nominally referred to these sets as the CMWR (i.e. cross-matched WH15 and redMaPPer) training and test sets to avoid confusion with the training and test sets created in §§4.2.1. The purpose of having the CMWR training set was to determine the best fit coefficients of our scaling relation whilst the purpose of having the CMWR test set was to measure the predictive performance of our learned scaling relation. Since we knew the spectroscopic redshift of the CMWR clusters, we segmented them into fixed redshift bin sizes of 0.01. This ensured that our training and test sets contained clusters from across the redshift scale via stratified sampling[36]. This involved randomly allocating approximately half of the clusters within each redshift bin into both sets, which resulted in our CMWR training and test sets containing 500 clusters each. The spectroscopic redshift and richness distributions of clusters in our CMWR training and test sets can be seen in Figure 4.8.

Next, we applied a search radius of 2.5 Mpc at each cluster's spectroscopic redshift as well as reapplying the same observing flags mentioned in §§4.2.1 to acquire galaxies from SDSS-IV DR16. This gave us a total of 2020690 galaxies, where our CMWR training set consisted of 1005167 galaxies and our CMWR test set consisted of 1015523 galaxies. We then applied our background subtraction model and colour-magnitude boundaries to count the number of cluster galaxies within each cluster. We established a linear scaling relation that was based on the number of identified cluster galaxies as an independent variable and $r_{200}$ as the dependent variable. This involved learning the best fit coefficients by minimising the residual sum of squares between the dependent and independent variables in

red-sequence galaxies.

[36]Stratified sampling is a strategy that minimises selection bias by splitting a dataset into new distributions that approximately resemble the original distribution.

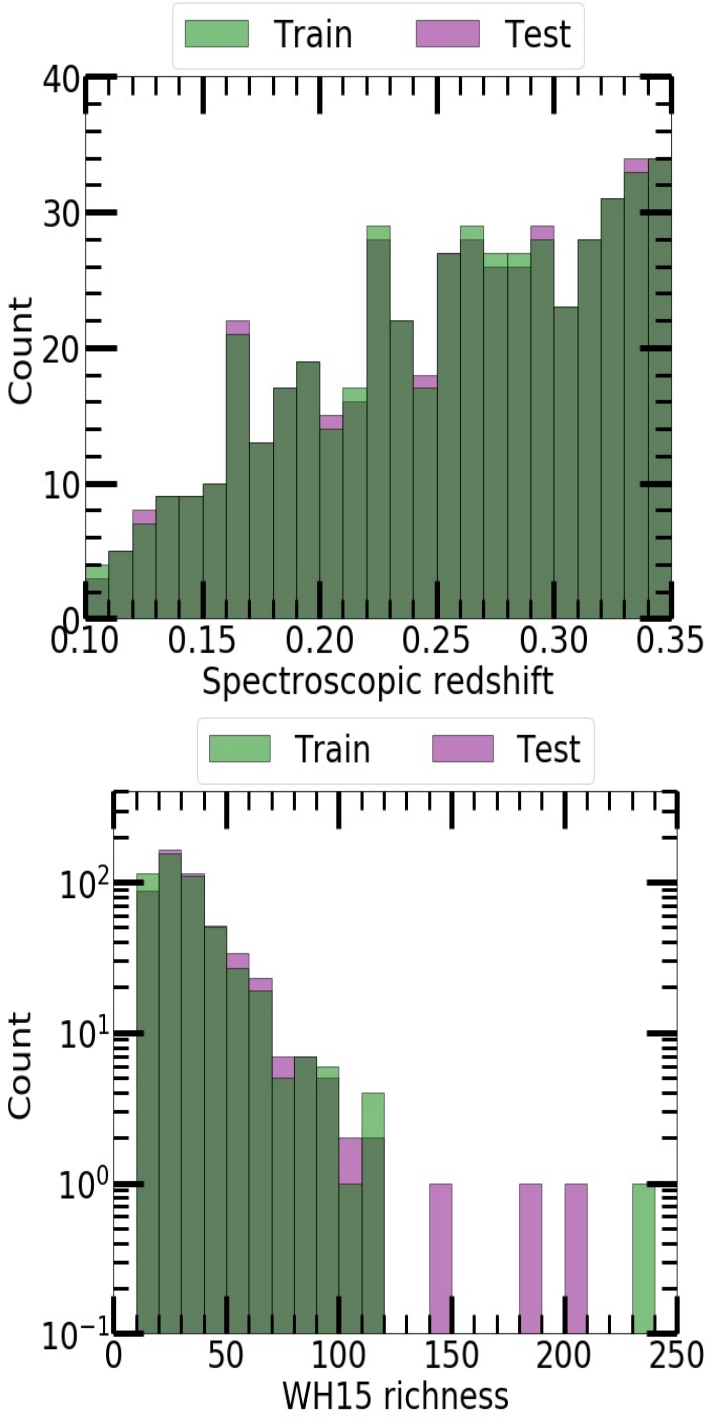


**Figure 4.8:** This figure shows histograms of the cluster spectroscopic redshift (top image) and WH15 richness (bottom image) distributions of clusters in our CMWR training (green) and test (purple) sets that were between a redshift range of $0.1 \leq z \leq 0.35$.

a linear regression algorithm from the Scikit-Learn machine learning library, where we again used the defaulted hyper-parameter values for the linear regression algorithm. We note that our cross-matched WH15 and redMaPPer cluster sample contained many clusters with low richness but only a few clusters with high richness. As such, we decided to assign the WH15 richness of each cluster as individual weights in the linear regression algorithm to minimise the effect of overfitting to potential outliers from clusters with low richness.

### 4.2.4 Preparation of a photometric dataset to estimate individual cluster richnesses

In order to measure richness within $r_{200}$ of individual clusters, we first approximated $r_{200}$ for clusters in our CMWR training and test sets using the learned scaling relation from §§4.2.3 to reacquire galaxies within $r_{200}$ from SDSS-IV DR16. We nominally referred to these new sets as the CMWR-$r_{200}$ training and test sets to avoid confusion with the CMWR training and test sets created in §§4.2.3. Similar to before, the purpose of having the CMWR-$r_{200}$ training set was to determine the best fit coefficients of a luminosity distribution fitting function whilst the purpose of having the CMWR-$r_{200}$ test set was to measure the predictive performance of the learned luminosity distribution fitting function. We obtained a total of 299807 galaxies in our CMWR-$r_{200}$ training set and 306953 galaxies in our CMWR-$r_{200}$ test set. The resultant color-magnitude diagrams of galaxies in our CMWR-$r_{200}$ training and test sets is shown in Figure 4.9.

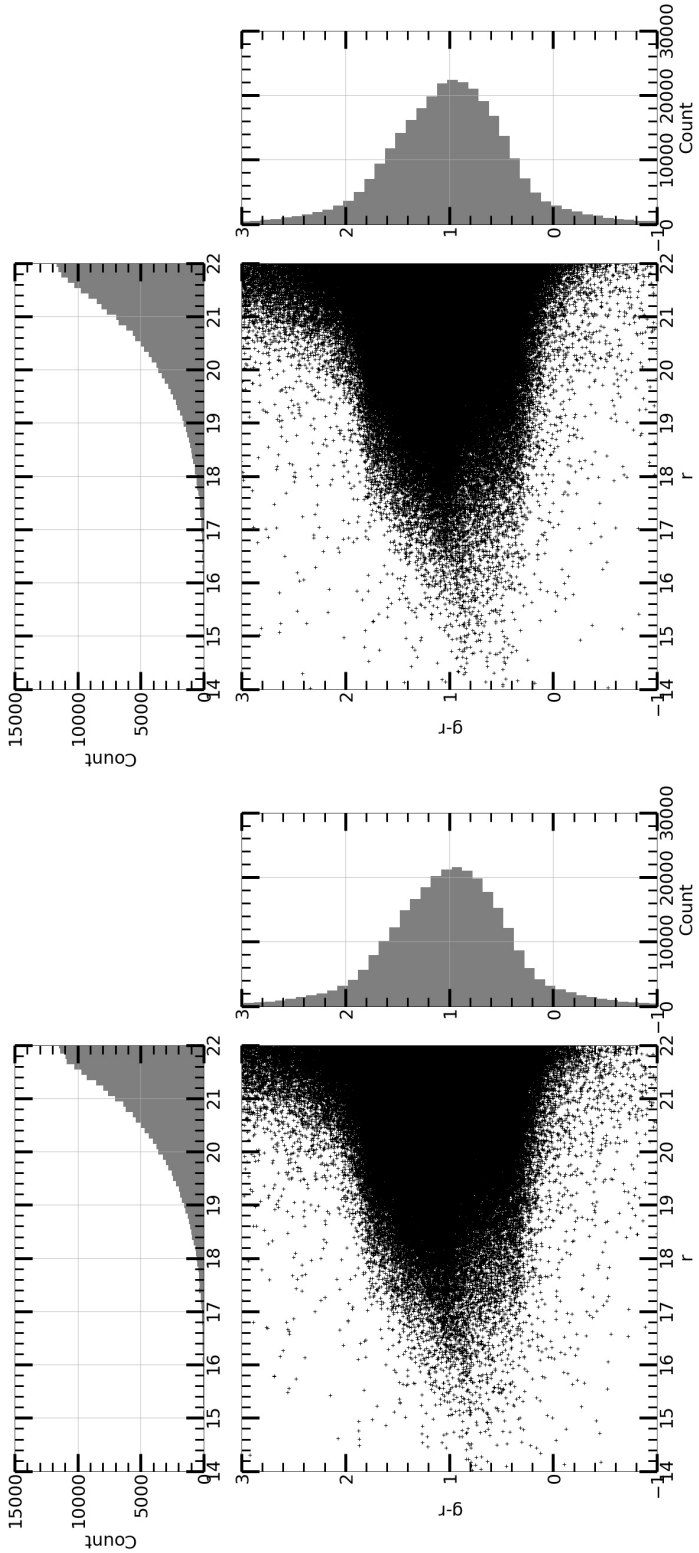


**Figure 4.9:** This figure shows colour-magnitude diagrams (using apparent magnitudes) of galaxies in our CMWR-$r_{200}$ training (left image) and test (right image) sets that were within an $r_{200}$ search radius and observed within SDSS-IV DR16.

### 4.2.5 Using a luminosity distribution fitting function to estimate individual cluster richnesses within $r_{200}$

We adopted a similar approach to the methodology described in Schechter (1976) to estimate the richness of individual clusters. Schechter (1976) showed that it was possible to use a luminosity distribution fitting function (i.e. the Schechter function) to do this. Briefly, this involved fitting the function to a composite luminosity distribution of cluster galaxies in order to determine best fit parameter values of the function. Then Schechter (1976) assumed that the best fit parameter values for $M^*$ and $\alpha$ can be applied universally to the luminosity distribution of individual clusters to locally fit for $n^*$ and thus estimate cluster richness. The Schechter function is expressed via the following equation:

$$n(M)dM = [0.4ln(10)]n^*[10^{0.4(M^*-M)}]^{\alpha+1}e^{-10^{0.4(M^*-M)}}dM \ , \tag{4.5}$$

where $M$ is absolute magnitude, $n^*$ is the number of galaxies per unit magnitude, $M^*$ is the 'characteristic' magnitude at which the distributions of faint and bright galaxies rapidly changes and $\alpha$ is the faint end slope parameter that describes the distribution of galaxies fainter than $M^*$. We note that $M^*$ and $\alpha$ directly influence the steepness of the bright and faint ends in the Schechter function whilst $n^*$ varies based on the observed number of galaxies within magnitude bins.

Firstly, we applied our background subtraction model and colour-magnitude boundaries to identify cluster galaxies from the CMWR-$r_{200}$ training set. We then performed Chi-squared fitting[37] with initialisation bounds for $M^*$ (i.e. between $-30$ and $-15$), $n^*$ (i.e. between 0 and positive infinity) and $\alpha$ (i.e. between $-2$ and $-1$) when fitting the Schechter function to a composite luminosity distribution that consisted of a subsample of identified cluster galaxies which appeared to have high completeness (i.e. greater than 90 per cent when using a base-10 logarithmic scale for the counts) between a restricted $r$ filter absolute magnitude

[37]We used the *curve fit* function from the SciPy Python library (Virtanen et al., 2020) to perform Chi-squared fitting of the Schechter function, which returned the best fit parameter values that minimised the Chi-squared fitting error and also returned an estimated covariance matrix of the best fit parameter values.

(i.e. between $-25$ and $-21.5$) and redshift (i.e. between 0.1 and 0.15) range. At the same time, we also explored various $r$ filter absolute magnitude bin sizes (i.e. from 0.01 to 3 with step sizes of 0.01) to obtain an optimal $r$ filter absolute magnitude bin size that minimised the Chi-squared fitting error and yielded galaxies across five or more $r$ filter absolute magnitude bins. Furthermore, we approximated the uncertainty in the number of identified cluster galaxies within each magnitude bin by assuming that the uncertainty followed a Poisson sampling hypothesis[38] when fitting the Schechter function. From which, we determined an optimal absolute magnitude bin size and best fit parameter values for $M^*$, $n^*$ and $\alpha$, where we also assumed that the best fit parameter values for $M^*$ and $\alpha$ can be applied universally to the luminosity distribution of individual clusters.

We remind the reader that our background subtraction model had not yet been corrected for the incompleteness of faint galaxies from observing limitations. This meant that we had to derive completeness corrections for the luminosity distribution (i.e. using $r$ filter absolute magnitudes) of individual clusters at different redshifts. Initially, we grouped the identified cluster galaxies from the CMWR-$r_{200}$ training set into redshift intervals of $\pm 0.04(1+z)$ that were centered in redshift bins from 0.105 to 0.345 with step sizes of 0.01, where identified cluster galaxies from different redshifts can go into multiple bins. Next, we fitted a 100 per cent completeness line across adjacent $r$ filter apparent magnitude bins[39] that were on the bright side of the peak and within the completeness limit of the AMF11 catalogue for each redshift interval. We then approximated the completeness fraction of the faintest $r$ filter apparent magnitude bin (N.B. we considered the two faintest magnitude bins on the bright side of the peak beyond $z > 0.14$ and three faintest magnitude bins on the bright side of the peak beyond $z > 0.33$ as the incompleteness of galaxies became more visibly noticeable for more magnitude bins at higher redshifts) by calculating the fraction in the

[38] A Poisson sampling hypothesis assumed that the distribution of galaxies is dictated by a Poisson process, such that the standard deviation of the counts within each magnitude bin was based on the square root of the count (Schechter, 1976).

[39] We used an $r$ filter apparent magnitude bin size that corresponded to the optimal $r$ filter absolute magnitude bin size. In addition, when working with the luminosity distribution of individual clusters we only considered cluster galaxies within an $r$ filter absolute magnitude range of $-25$ to $-20.5$, where $-20.5$ was the $r$ filter absolute magnitude limit that was used to determine WH15 richness in Wen & Han (2015).

expected number of cluster galaxies (i.e. based on the 100 per cent completeness line) to the observed number of cluster galaxies (i.e. identified by our background subtraction model).

We applied these completeness fractions to the luminosity distribution of individual clusters by multiplying the observed count of the faintest (N.B. we again considered the two faintest magnitude bins on the bright side of the peak beyond $z > 0.14$ and three faintest magnitude bins on the bright side of the peak beyond $z > 0.33$) $r$ filter absolute magnitude bin[40] on the bright side of the peak by the completeness fraction of the corresponding $r$ filter apparent magnitude bin within the nearest redshift interval. We then replaced the uncertainty range of the observed count in the magnitude bin with this computed completeness correction value as the new lower and upper uncertainty limits when performing Chi-square fitting. This ensured that the Schechter function did not fit to incomplete $r$ filter absolute magnitude bins.

Finally, we estimated cluster richnesses within $r_{200}$ by integrating[41] the locally fit Schechter function. This gave us the expected number of cluster galaxies within $r_{200}$ that had an $r$ filter absolute magnitude brighter than $-20.5$. We also compared our estimated cluster richnesses with WH15 richnesses, spectroscopic redshift, 'actual' $r_{200}$ and redMaPPer richnesses in order to examine the predictive performance of the optimal $r$ filter absolute magnitude bin size and best fit parameters for $M^*$ and $\alpha$ in the Schechter function. We note that WH15 richness was specific to $r_{200}$ whereas redMaPPer richness was specific to redMaPPer's own scaling radius rather than $r_{200}$. This meant that we could directly quantify the error between our estimated cluster richnesses and WH15 richnessses by using root mean squared error as a metric.

[40] Since our completeness fractions were measured in $r$ filter apparent magnitudes, we had to convert between $r$ filter apparent magnitudes and $r$ filter absolute magnitudes to determine the relevant completeness fraction.

[41] We utilised the incomplete Gamma function (see Equation 27 in Schechter (1976)) to compute the integral.

## 4.3 Results

### 4.3.1 Model tuning analyses

#### 4.3.1.1 Analysis of our trained background subtraction model

We conducted ten iterations of Monte Carlo cross-validation to measure the variability of the predictive performance of our background subtraction model, as well as conducting sixty iterations of random search on the tunable hyper-parameters of our background subtraction model, to determine an optimal hyper-parameter combination that maximised the AUCPR of galaxies in our validation set. It can be seen in Table S1 (available online) that the optimum hyper-parameter combination was as follows: optimal batch size = 2048; optimal learning rate = 0.0001; optimal optimiser algorithm = RMSprop and optimal architecture layout = 3. This optimum hyper-parameter combination yielded a mean AUCPR value of 40.24 per cent with a standard deviation of 1.85 per cent for galaxies in our validation set. Furthermore, it can be seen in Table S2 (available online) that the optimum class probability threshold was 0.29, when using the optimum hyper-parameter combination. This optimum class probability threshold yielded a F1 score of 48.92 per cent for galaxies in our validation set.

#### 4.3.1.2 Analysis of our established scaling relation to estimate $r_{200}$

We constructed a scaling relation using clusters in our CMWR training set to estimate the $r_{200}$ of each cluster when given the number of cluster galaxies identified by our background subtraction model as an input. The best fit coefficients of our scaling relation were determined by minimising the weighted residual sum of squares between the independent and dependent variables, where our scaling relation is defined via the following equation:

$$pred_{r_{200}} = (3.39 \pm 0.23) n_{gal} \ + \ (950.65 \pm 25.25) \ , \tag{4.6}$$

where $pred_{r_{200}}$ is the predicted $r_{200}$, $n_{gal}$ is the number of cluster galaxies identified within a 2.5 Mpc search radius at each cluster's spectroscopic redshift

and the uncertainty represents the standard error of the parameter estimates. In Figure 4.10, it can be seen that there was a larger drop in the number of cluster galaxies identified by our background subtraction model at higher redshifts (i.e. $z > 0.3$) when compared to the number of identified cluster galaxies at lower redshifts with the same 'actual' $r_{200}$ values. This was likely due to cluster galaxies at higher redshifts having larger observed photometric errors, which made it more difficult for our background subtraction model to identify these cluster galaxies. We note that we obtained a Pearson correlation coefficient value of 0.39 between the number of identified cluster galaxies and 'actual' $r_{200}$ variables. We also observed that both WH15 and redMaPPer richnesses appeared to somewhat linearly increase with 'actual' $r_{200}$ and the number of identified cluster galaxies. Furthermore, in Figure 4.11, we compared the predictive performance of our predicted $r_{200}$ with the 'actual' $r_{200}$, where we found that our predicted $r_{200}$ was quite comparable to the 'actual' $r_{200}$ across all cluster sizes. Although, we noticed there was greater variability in the predicted $r_{200}$ at lower cluster richnesses, where we obtained a root mean squared error of 218.14 and a median absolute percentage error of 11.89 per cent between our predicted and 'actual' $r_{200}$ values.

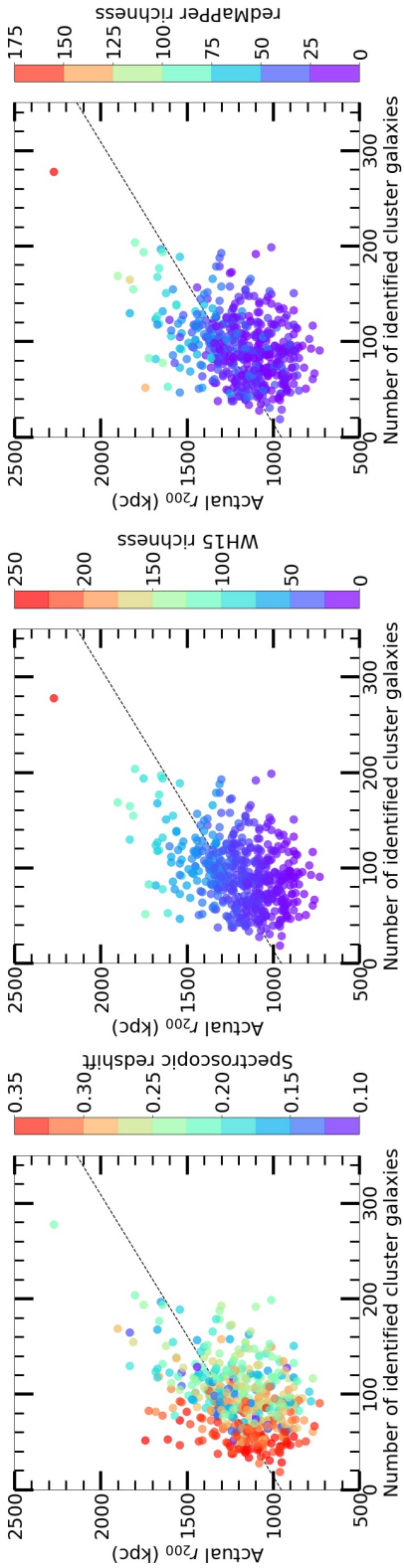


**Figure 4.10:** This figure shows our scaling relation (black dotted line) to estimate the $r_{200}$ of clusters. It used the 'actual' $r_{200}$ of clusters from our CMWR training set as a dependent variable and the number of cluster galaxies identified by our background subtraction model within a 2.5 Mpc search radius at each cluster's spectroscopic redshift as an independent variable. We also display the corresponding spectroscopic redshift (left image), WH15 richness (middle image) and redMaPPer richness (right image) of each cluster.

#### 4.3.1.3 Analysis of the best fit parameters for a luminosity distribution fitting function to estimate individual cluster richnesses Within $r_{200}$

We used a Chi-squared fitting approach to determine the best fit parameters of the Schechter function when fitting to a composite luminosity distribution that consisted of a subsample of identified cluster galaxies from our CMWR training set with high completeness. We also simultaneously determined an optimal $r$ filter absolute magnitude bin size that minimised the Chi-squared fitting error and yielded galaxies across five or more $r$ filter absolute magnitude bins. In Table S3 (available online), we identified an optimal $r$ filter absolute magnitude bin size of 0.52 that had corresponding best fit parameter values of $M^* = -22.81$ with a standard deviation of $\pm 0.5$; $n^* = 159.82$ with a standard deviation of $\pm 154.62$ and $\alpha = -1.99$ with a standard deviation of $\pm 0.37$. In Figure 4.12, we display the composite luminosity distribution and fitted Schechter function using the optimal $r$ filter absolute magnitude bin size and best fit parameter values.

We then used the optimal $r$ filter absolute magnitude bin size and best fit parameter values for $M^*$ and $\alpha$ to fit the Schechter function to the luminosity distribution of individual clusters from our CMWR-$r_{200}$ training set. This enabled us to estimate individual cluster richnesses by integrating the locally fit Schechter function. Subsequently, we obtained a root mean squared error of 18.06 and a median absolute percentage error of 34.33 per cent between our estimated cluster richnesses and WH15 richnesses within $r_{200}$. In Figure 4.13, we noticed that WH15 richnesses had a strong linear correlation with our estimated cluster richnesses. We also observed that spectroscopic redshifts seemed to have no distinguishable correlation with our estimated cluster richnesses. In addition, we noticed that there was a strong linear correlation between 'actual' $r_{200}$, redMaPPer richnesses and our estimated cluster richnesses. These results confirmed that our approach to estimate individual cluster richnesses was appropriate since we did not train any of our models to minimise cluster richness prediction error but we still obtained strong correlations with WH15 and redMaPPer richnesses. Furthermore, we were aware that our CMWR-$r_{200}$ training set contained clusters that

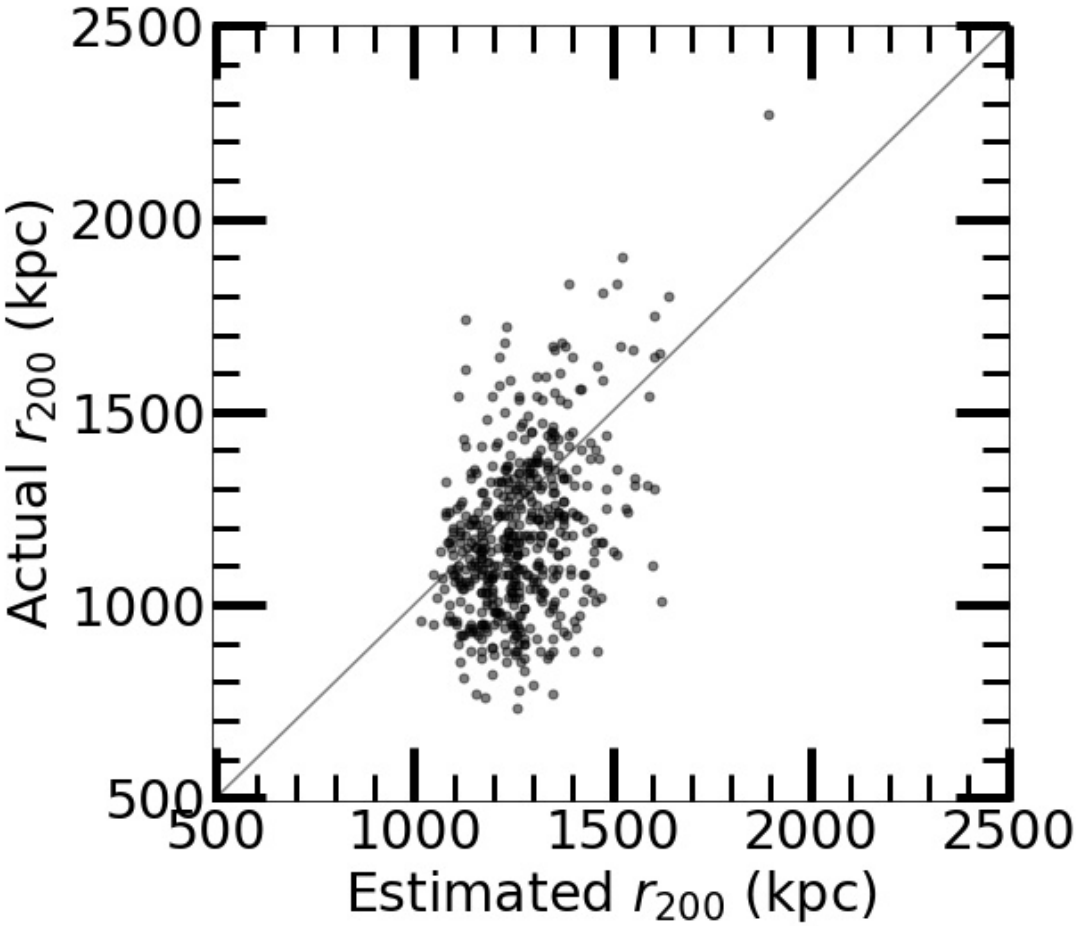


**Figure 4.11:** This figure shows a direct comparison of $r_{200}$ predicted by our scaling relation with the 'actual' $r_{200}$ of clusters from our CMWR training set.

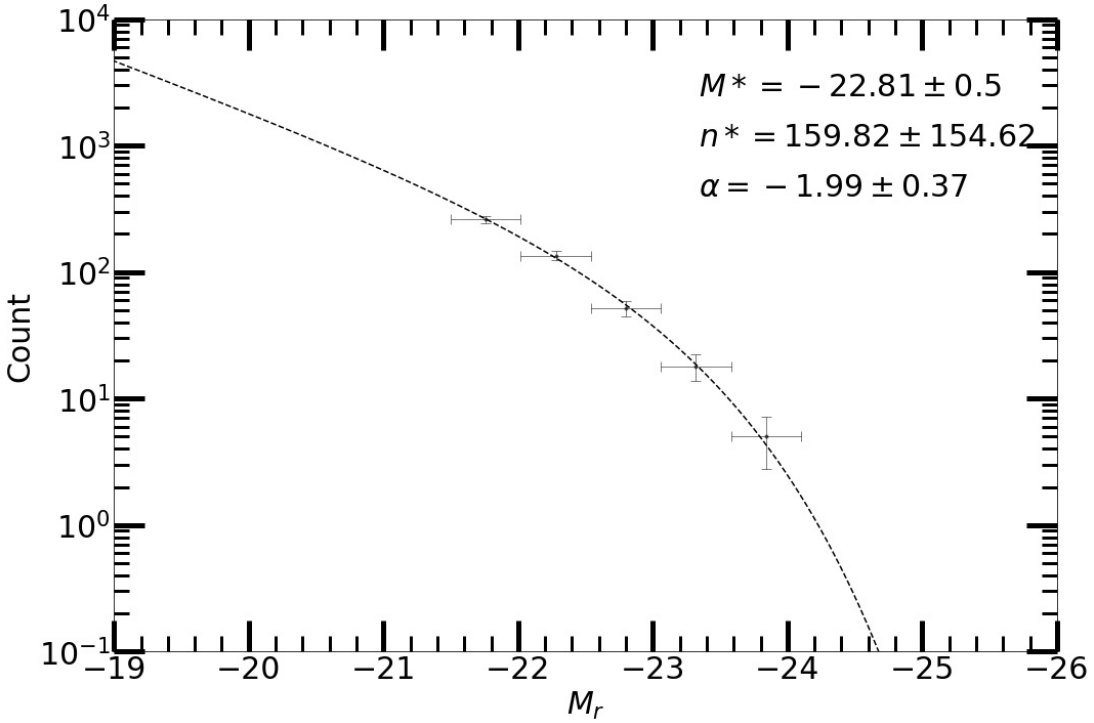


**Figure 4.12:** This figure shows the best fit Schechter function (black dotted line) overlaid on a composite luminosity distribution (using $r$ filter absolute magnitudes) that consisted of a subsample of identified cluster galaxies from our CMWR-$r_{200}$ training set with an optimal $r$ filter absolute magnitude bin size of 0.52. The best fit parameter values and their respective standard deviations are displayed in the top right corner of the figure. The x-axis error bars display the width of each $r$ filter absolute magnitude bin and the y-axis error bars display the standard deviation of the observed count within each $r$ filter absolute magnitude bin when assuming a Poisson sampling hypothesis.

were not truly unseen, as we had utilised these clusters before to create our scaling relation. Although, it was still interesting to test our methodology on clusters that were seen and unseen to compare differences in predictive performance.

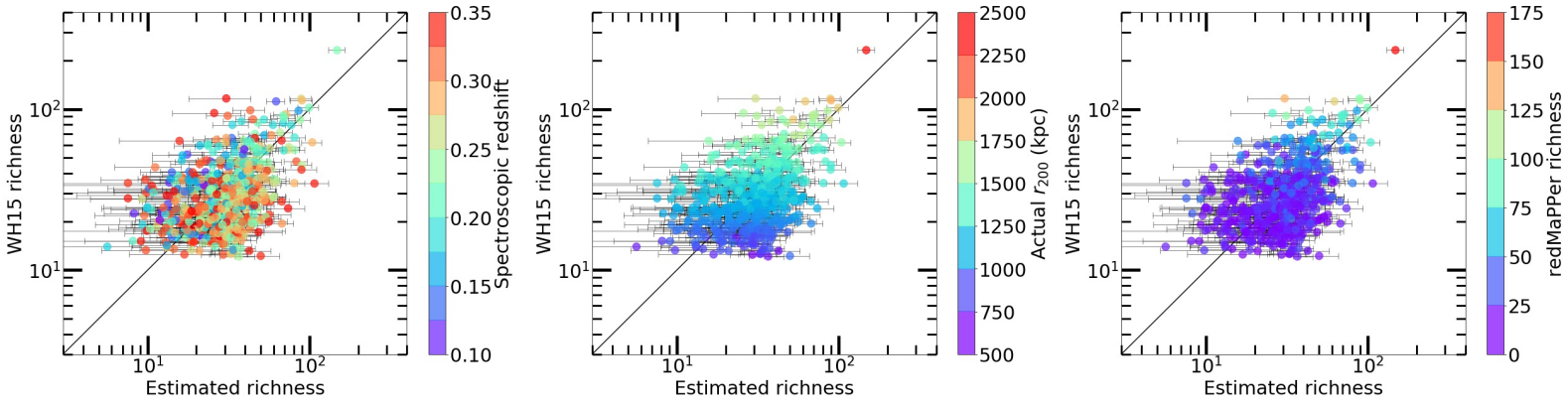


**Figure 4.13:** This figure shows a direct comparison between our estimated cluster richnesses and WH15 richnesses of clusters from our CMWR-$r_{200}$ training set when using the optimal $r$ filter absolute magnitude bin size and best fit parameter values for $M^*$ and $\alpha$. We also display the corresponding spectroscopic redshifts (left image), 'actual' $r_{200}$ (middle image) and redMaPPer richness (right image) for each cluster. The x-axis error bars display the standard deviation of the locally fit $n^*$ when computing the integral of the Schechter function to determine our estimated cluster richnesses.

### 4.3.2 Overall performance analyses with test sets

We further assessed our entire methodology on clusters belonging to our various test sets to obtain an unbiased evaluation of the true predictive performance of our models. Firstly, we applied our background subtraction model to cluster and field galaxies in our test set. This yielded a F1 score of 72.81 per cent and a balanced accuracy of 83.20 per cent when using the optimal hyper-parameter combination and optimal class probability threshold for our background subtraction model. In Figure 4.14, we display a direct comparison of the 'actual' and predicted cluster and field galaxies. It can be seen that our background subtraction model learned to correctly classify almost all of the field galaxies surrounding the cluster galaxies but it made more incorrect classifications in regions where the 'actual' cluster and field galaxies had greater overlap within colour-magnitude space. Meanwhile, in Figure 4.15, we compared the number of cluster and field galaxies identified by our subtraction model across redshift bin sizes of 0.01. At lower redshifts (i.e. $z \leq 0.45$), we noticed that our background subtraction model slightly underestimated (i.e. misclassified 'actual' field galaxies as cluster galaxies or misclassified 'actual' cluster galaxies as field galaxies) the overall number of 'actual' cluster and field galaxies. In particular, we noticed a larger drop in the number of identified 'actual' cluster galaxies between $0.3 \leq z \leq 0.35$, which was similar to our observation in Figure 4.10. Correspondingly, at higher redshifts (i.e. $z > 0.45$), we found that our background subtraction model correctly classified almost all of the galaxies. In Figures 4.16 and 4.17, we compared the number of cluster and field galaxies identified by our subtraction model across $r$ filter apparent and absolute magnitude bin sizes of 0.1 respectively. In both magnitude distributions, we noticed that our background subtraction model slightly underestimated the overall number of 'actual' cluster galaxies at all magnitudes. We also noticed that our background subtraction model slightly underestimated the overall number of 'actual' field galaxies at intermediate brightnesses (i.e. between 16.5 and 20.5 in $r$ filter apparent magnitude and between $-24$ and $-20$ in $r$ filter absolute magnitude) but correctly classified almost all of the other fainter and brighter 'actual' field galaxies. Furthermore, in Figure 4.18 we examined the proportion of 'red' and 'blue' 'actual' cluster galaxies that were identified

by our background subtraction model at different redshifts. We found that our background subtraction model identified 84.32 per cent of 'red' 'actual' cluster galaxies and recovered 73.11 per cent of 'blue' 'actual' cluster galaxies between a redshift range of $0.1 \leq z \leq 0.35$. This indicated that our background subtraction model was more confident at identifying 'red' 'actual' cluster galaxies than 'blue' 'actual' cluster galaxies, which was likely due to the 'blue' 'actual' cluster galaxies having greater overlap with field galaxies within colour-magnitude space.

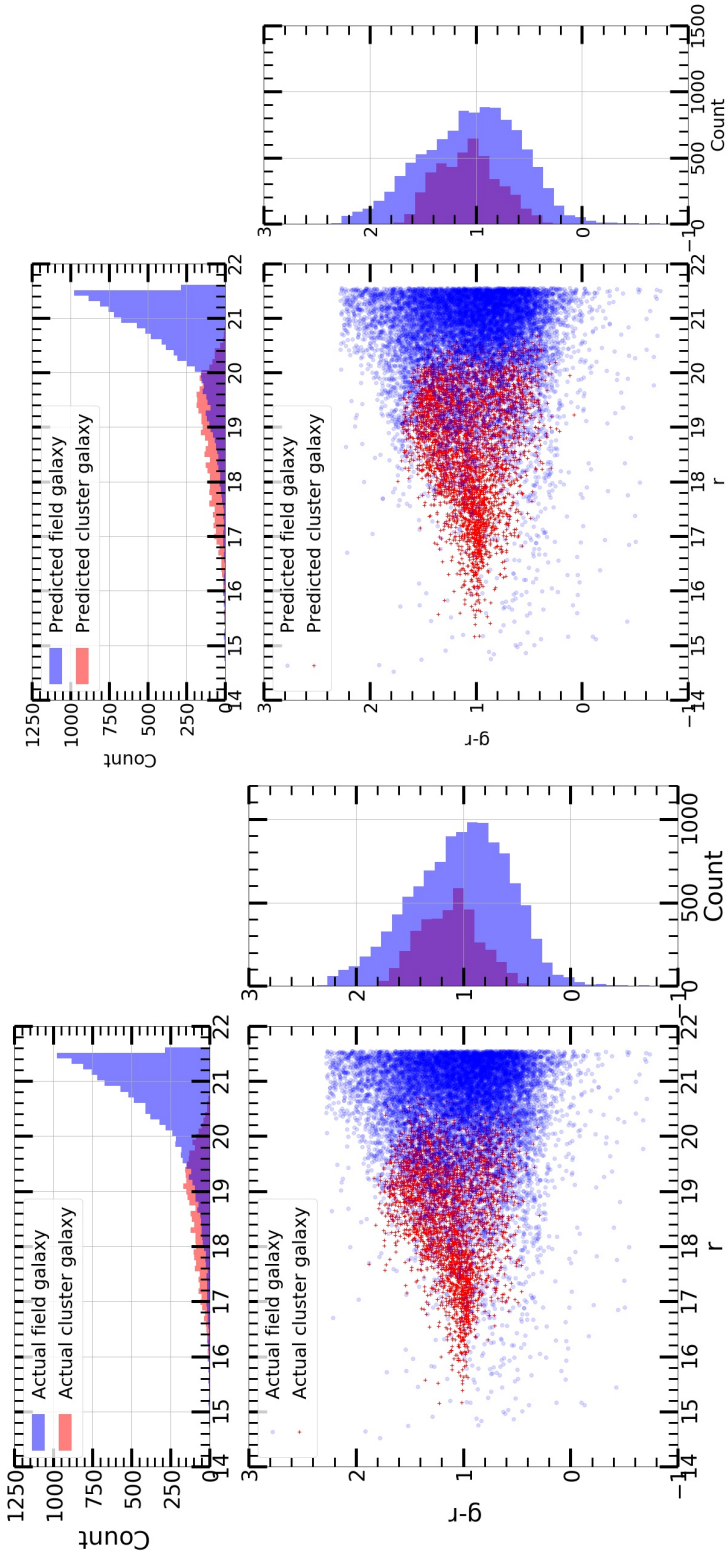


**Figure 4.14:** This figure shows a direct comparison of the colour-magnitude diagrams (using apparent magnitudes) for the 'actual' (left image) and predicted (right image) cluster (red cross) and field (blue circle) galaxies in our test set.

**Figure 4.15:** This figure shows histograms of the number of identified cluster (left image) and field (right image) galaxies in our test set when using fixed redshift bin sizes of 0.01. The blue fill with black dotted lines represents the original number of 'actual' cluster or field (N.B. we only display field galaxies that had an available photometric redshift) galaxies within each redshift bin. The red points represent the number of cluster or field galaxies identified by our background subtraction model within each redshift bin, the green crosses represent the number of 'actual' cluster or field galaxies identified by our background subtraction model within each redshift bin and the x-axis error bars display the width of each redshift bin.

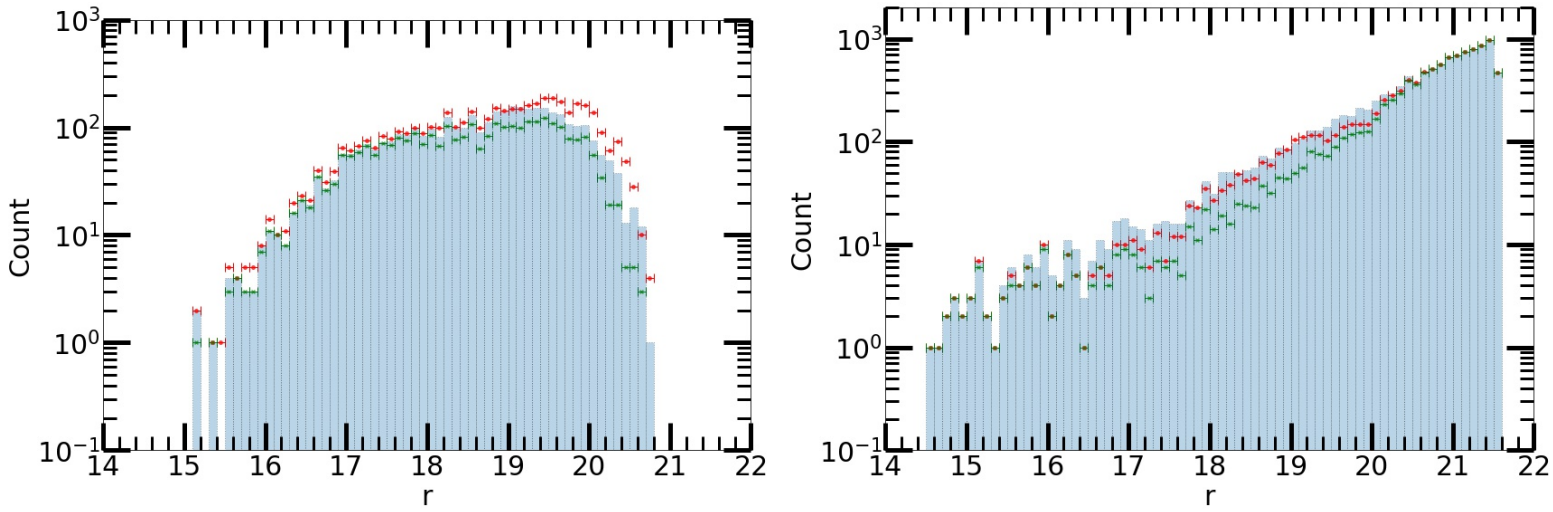


**Figure 4.16:** This figure shows histograms of the number of identified cluster (left image) and field (right image) galaxies in our test set when using fixed $r$ filter apparent magnitude bin sizes of 0.1. The blue fill with black dotted lines represents the original number of 'actual' cluster or field galaxies within each $r$ filter apparent magnitude bin. The red points represent the number of cluster or field galaxies identified by our background subtraction model within each $r$ filter apparent magnitude bin, the green crosses represent the number of 'actual' cluster or field galaxies identified by our background subtraction model within each $r$ filter apparent magnitude bin and the x-axis error bars display the width of each $r$ filter apparent magnitude bin.

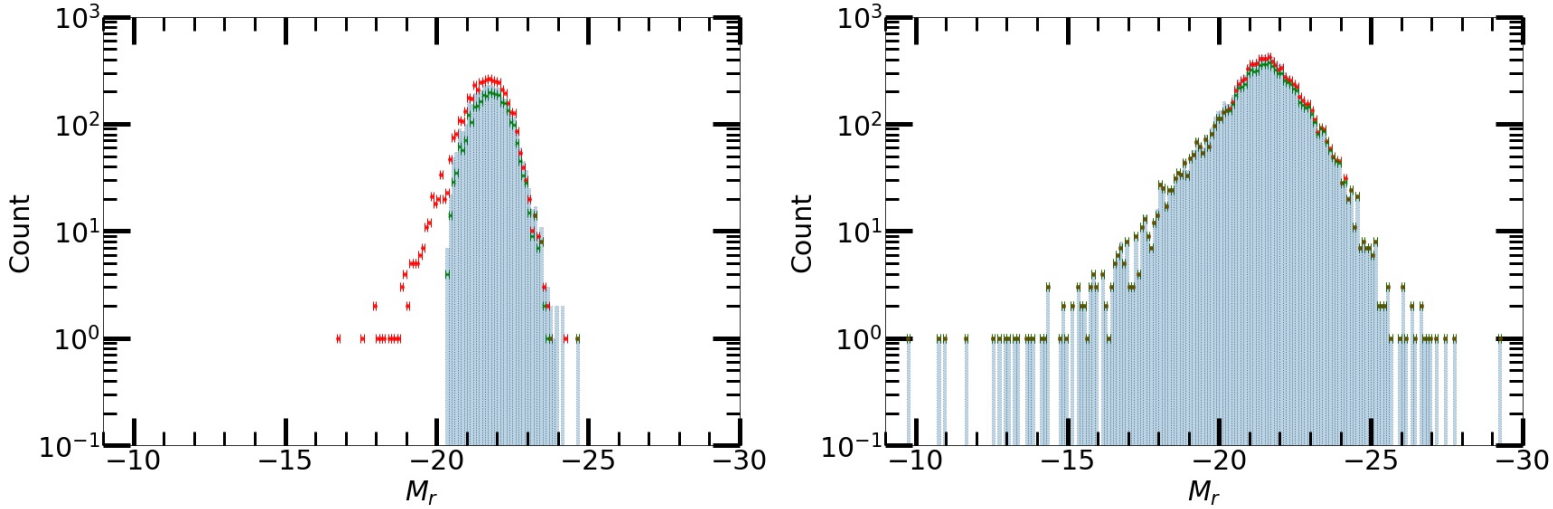


**Figure 4.17:** This figure shows histograms of the number of identified cluster (left image) and field (right image) galaxies in our test set when using fixed $r$ filter absolute magnitude bin sizes of 0.1. The blue fill with black dotted lines represents the original number of 'actual' cluster or field (N.B. we only display field galaxies that had an available photometric redshift) galaxies within each $r$ filter absolute magnitude bin. The red points represent the number of cluster or field galaxies identified by our background subtraction model within each $r$ filter absolute magnitude bin, the green crosses represent the number of 'actual' cluster or field galaxies identified by our background subtraction model within each $r$ filter absolute magnitude bin and the x-axis error bars display the width of each $r$ filter absolute magnitude bin.

We then applied our learned scaling relation and colour-magnitude boundaries to clusters in our CMWR test set to approximate $r_{200}$ for each cluster. In Figure 4.19, we noticed that the number of identified cluster galaxies and 'actual' $r_{200}$ was relatively consistent with our learned scaling relation from Figure 4.10. From which, we obtained a Pearson correlation coefficient value of 0.50 between the number of identified cluster galaxies and 'actual' $r_{200}$ variables in Figure 4.19. We also noticed that our predicted and 'actual' $r_{200}$ values in Figure 4.20 was similar to the overall trend observed in Figure 4.11, where we obtained a root mean squared error of 200.86 and a median absolute percentage error of 11.66 per cent between our predicted and 'actual' $r_{200}$ values in Figure 4.20.

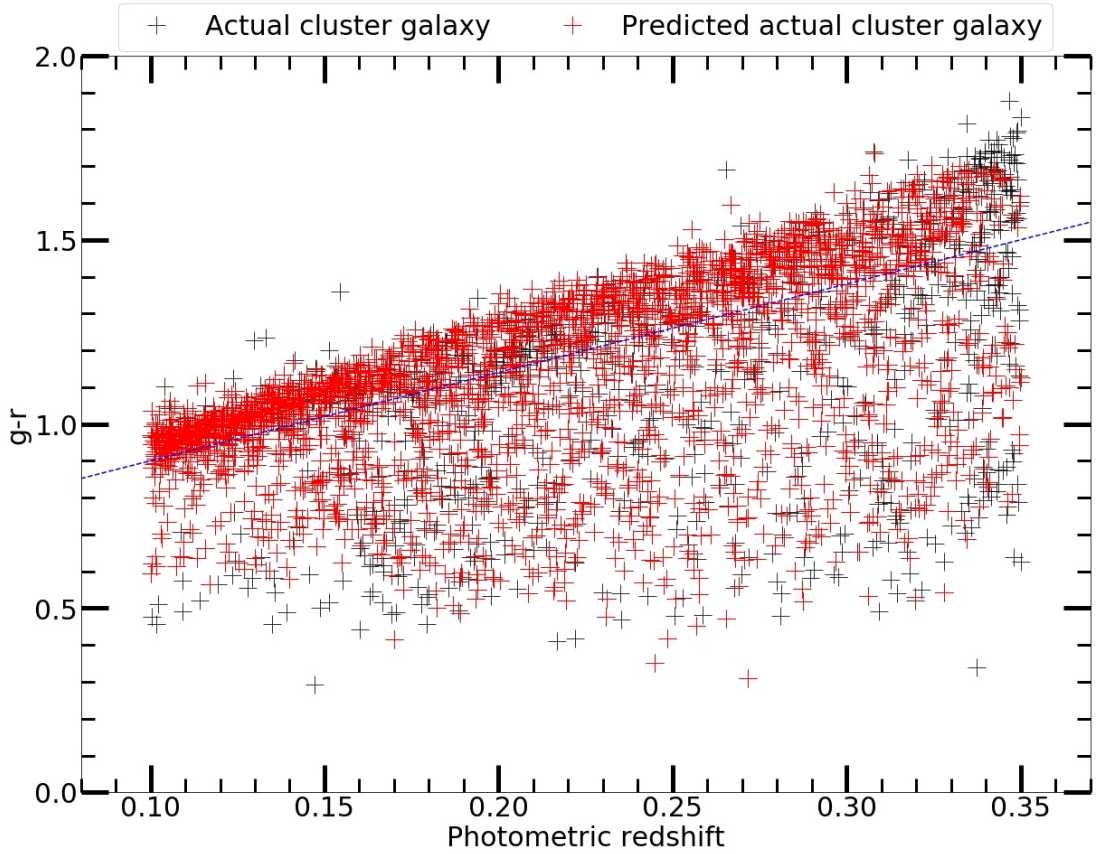


**Figure 4.18:** This figure shows a comparison of the 'red' and 'blue' 'actual' cluster galaxies (black cross) in our test set that were identified (red cross) by our background subtraction model at different redshifts, where we assumed that galaxies above the blue dashed line were 'red' and galaxies below the blue dashed line were 'blue'.

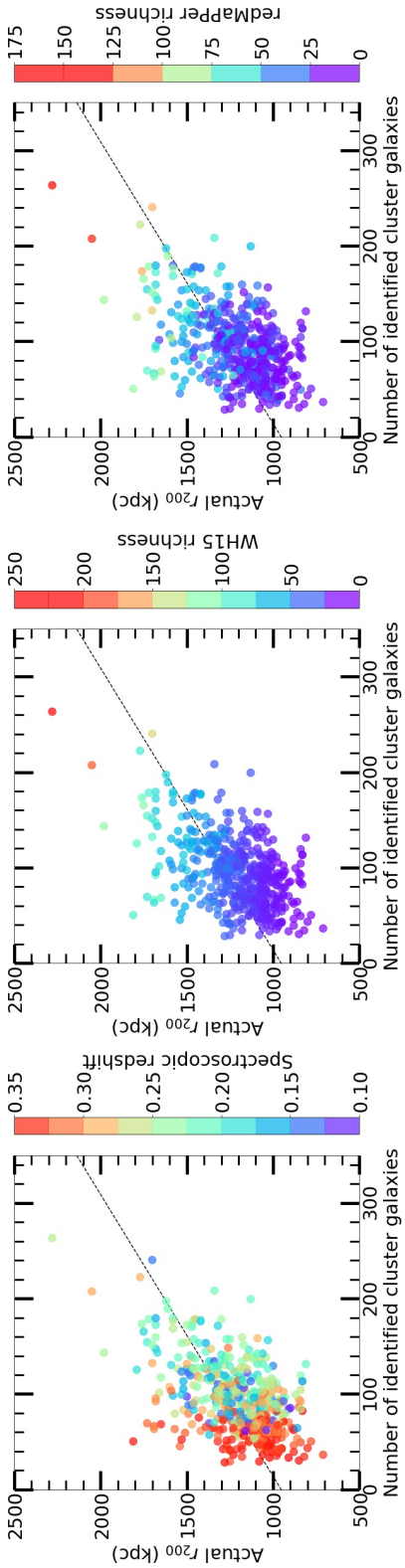


**Figure 4.19:** This figure is equivalent to Figure 4.10 except we overlaid our learned scaling relation (black dotted line) on clusters in our CMWR test set.

Finally, we examined the predictive performance of the optimal $r$ filter absolute magnitude bin size and best fit parameter values for $M^*$ and $\alpha$ in the Schechter function on individual clusters in our CMWR-$r_{200}$ test set. In Figure 4.21, we noticed that the overall trends between our estimated cluster richnesses and WH15 richnesses with spectroscopic redshifts, 'actual' $r_{200}$ and redMaPPer richnesses were again consistent with Figure 4.13, where our estimated cluster richnesses had no distinct correlation with spectroscopic redshifts and our estimated cluster richnesses linearly increased with 'actual' $r_{200}$ and redMaPPer richnesses. Subsequently, we obtained a root mean squared error of 18.04 and a median absolute percentage error of 33.50 per cent between our estimated cluster richnesses and WH15 richnesses within $r_{200}$.

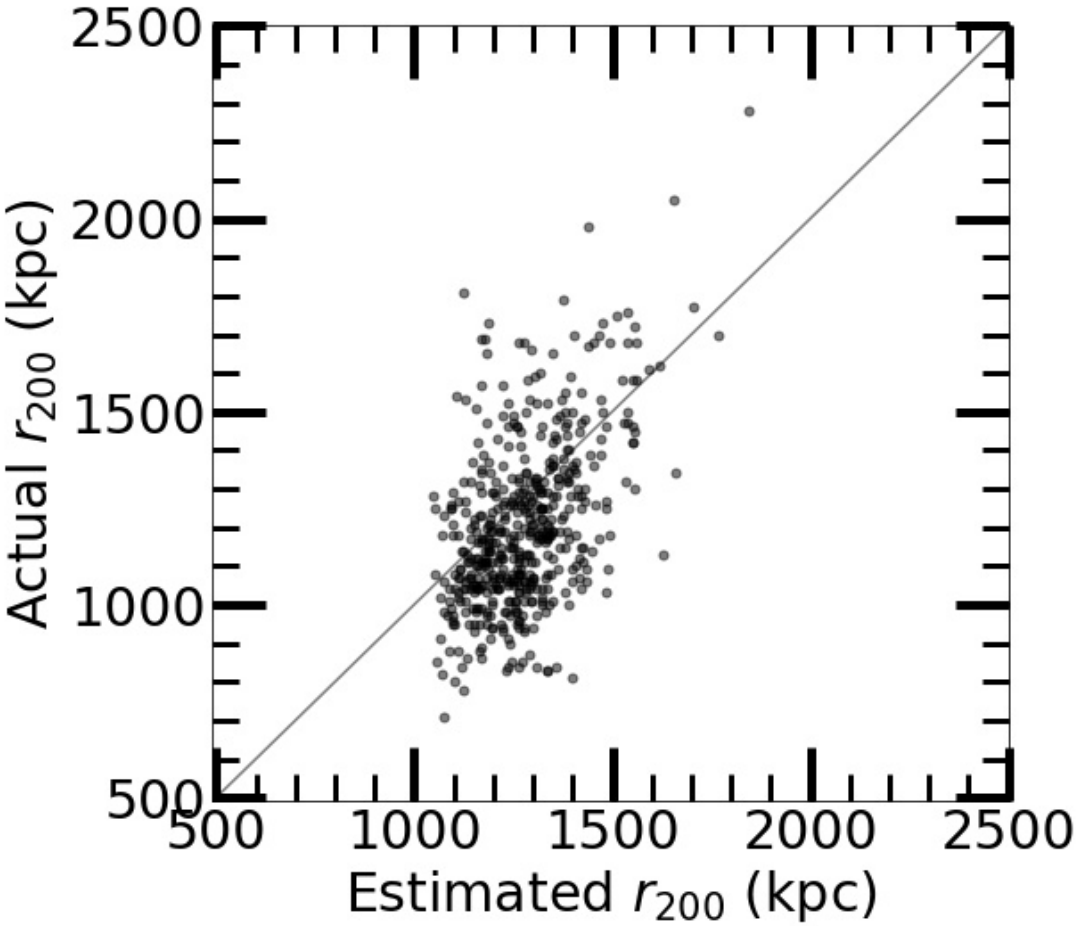


**Figure 4.20:** This figure is equivalent to Figure 4.11 except it compared the predicted and 'actual' $r_{200}$ of clusters in our CMWR test set.

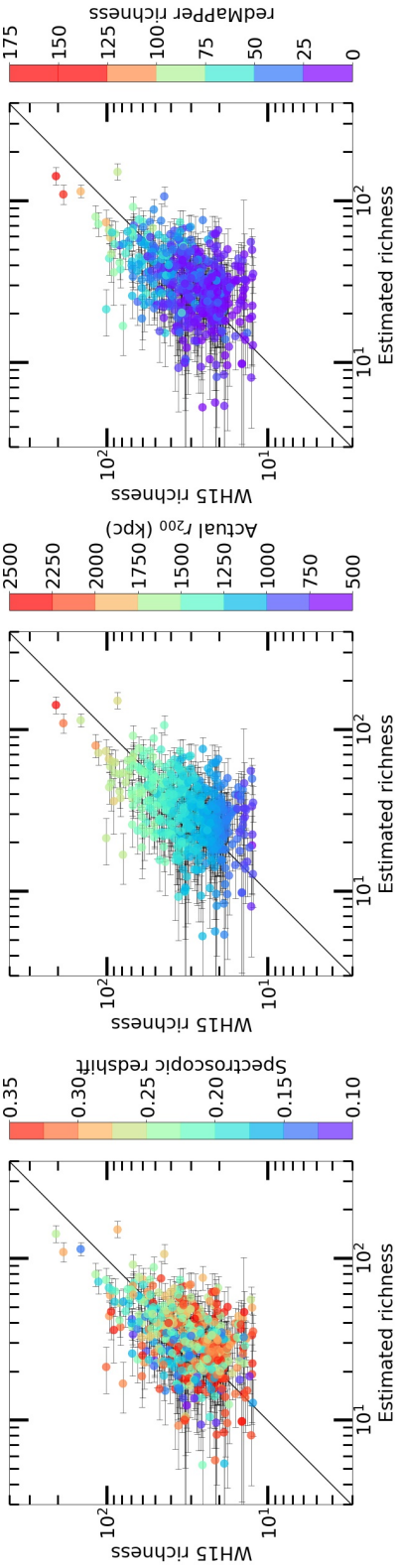


**Figure 4.21:** This figure is equivalent to Figure 4.13 except it was applied to unseen clusters from the CMWR-$r_{200}$ test set.

### 4.3.3 Examining the importance of input features to our background subtraction model

In Figure 4.22, we examined the importance of each input feature to our background subtraction model. This involved randomly shuffling the data of each input feature and then applying our background subtraction model on the dataset to observe how the shuffled feature impacted the predictive performance. This strategy is known as permutation feature importance testing (Breiman, 2001), where the permutation scores were based on the number of 'actual' cluster galaxies identified by our background subtraction model. In particular, a lower permutation score for an input feature implied greater reliance of our background subtraction model on that specific input feature to provide good predictive performance, because randomly shuffling the data for an important input feature would result in fewer 'actual' cluster galaxies being identified. We applied this permutation feature importance test to galaxies in our test set, which originally contained 3750 cluster galaxies. Subsequently, we observed that $g$, $r$, $i$, $z$, $u-g$, $u-r$, $g-i$ and $g-z$ appeared to have greater significance to our background subtraction model whereas $u$, $g-r$, $r-i$, $i-z$, $r-z$, $u-i$ and $u-z$ appeared to have lesser significance to our background subtraction model. Although, it is important to note that our background subtraction model had effectively utilised all the input features since the number of identified 'actual' cluster galaxies for each input feature was still only a fraction of the original number of 'actual' cluster galaxies.

## 4.4 Discussion

In Figure 4.5, we observed that the photometric redshift distribution of galaxies in our cluster galaxy sample was skewed towards higher redshifts, such that higher redshift cluster galaxies were overrepresented. To achieve a fair representation of cluster galaxies at different redshifts in our background subtraction model, we randomly sampled a fixed number of cluster galaxies within fixed redshift bin sizes of 0.01 when creating our training, validation and test sets. This ensured

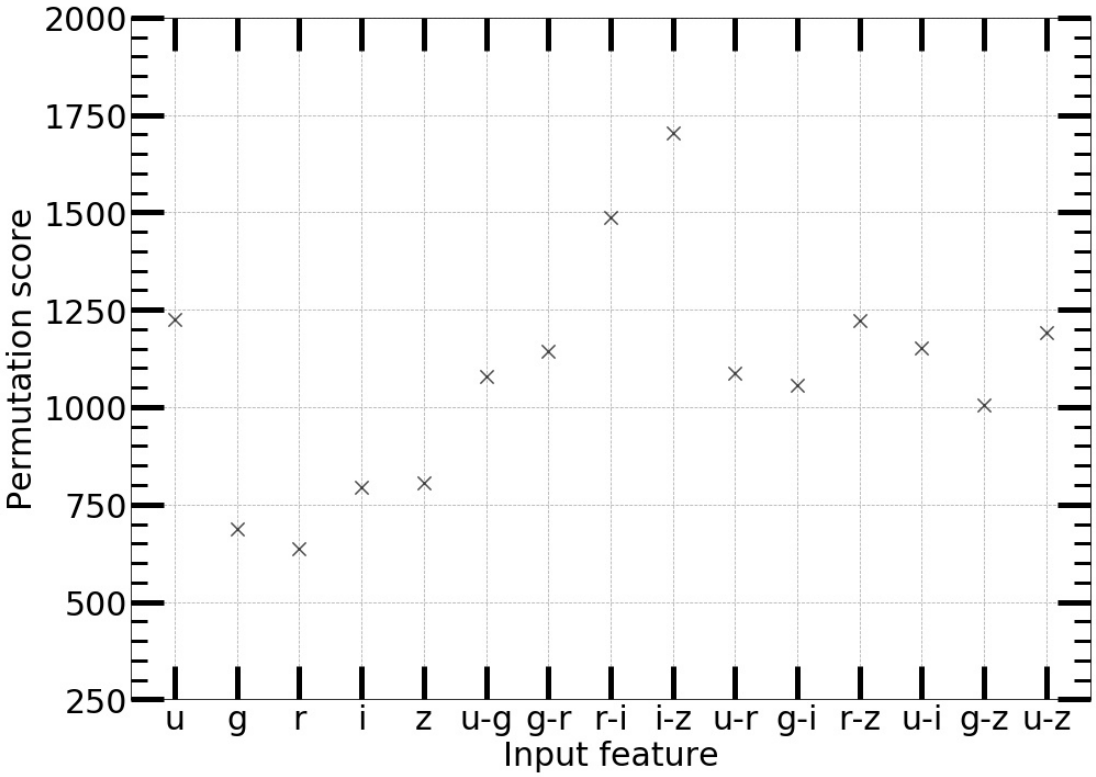


**Figure 4.22:** This figure shows the importance (N.B. a lower permutation score signifies greater importance) of each input feature to our background subtraction model, where the permutation score was based on the number of 'actual' cluster galaxies identified by our background subtraction model after randomly shuffling the data for each input feature.

that our background subtraction model was exposed to equal numbers of cluster galaxies at various redshifts within colour-magnitude space. We also exposed our background subtraction model to equal numbers of cluster and field galaxies during its training. This was to ensure a fair representation of the different galaxy classes in our background subtraction model.

We initially constrained our training sample to only spectroscopically confirmed cluster galaxies from the AMF11 catalogue but we quickly noticed that the training sample itself had a significant drop in the number of faint cluster galaxies across all redshifts when compared to the non-spectroscopically confirmed cluster galaxies. As such, we decided not to adopt this constraint when training our background subtraction model. Furthermore, we did not utilise spectroscopically confirmed field galaxies since it was difficult to acquire a sample that was representative of all potential foreground and background galaxies encountered within a random field. Although, in future work this may be possible since the number of spectroscopically confirmed cluster and field galaxies would naturally increase over time.

When constructing our scaling relation, we employed 'actual' $r_{200}$ values that were estimated from a scaling relation (see Equation 1 in Wen et al. (2012)) that was based on the total $r$ filter luminosity of identified cluster galaxies within a 2.5 Mpc radius from the cluster center at each clusters redshift. This meant that the errors from their estimated $r_{200}$ values would have carried over into our estimated $r_{200}$ values too. In future work, we could instead consider employing $r_{200}$ values from X-ray catalogues as X-ray emission measurements of clusters are not as significantly influenced by projection effects (Ebeling et al., 2010). This would improve the overall precision of our 'actual' $r_{200}$ values and thus improve the precision of our cluster richness estimates within $r_{200}$. Furthermore, we can establish a scaling relation for any radii, not only $r_{200}$, as long as we have sufficient data to enable the construction of a scaling relation for the radii.

In Figures 4.13 and 4.21 we did not observe any redshift biases in our estimated cluster richnesses after we applied completeness corrections to account for fewer observed galaxies at the faint end of the luminosity distribution of individual clusters. This indicated that incompleteness of our cluster galaxy sample from the AMF11 catalogue had a bigger impact on estimating cluster richnesses than

incompleteness from misclassifications by our background subtraction model. Although, in Figures 4.10, 4.15 and 4.19 we observed a larger drop in the number of identified cluster galaxies at higher redshifts (i.e. $z > 0.3$) when compared to the number of identified cluster galaxies at lower redshifts. We believed that this could be due to the cluster galaxies at higher redshifts having larger photometric errors than cluster galaxies at lower redshifts, which can be seen in Figure 4.18 by the increased scatter between data points as redshift increased. Naturally, this would make it more difficult for our background subtraction model to identify them. As such, we would expect there to be fewer cluster galaxies identified at higher redshifts, since we did not truly account for cluster galaxies having larger photometric errors at higher redshifts in our background subtraction model. In the future, it would be beneficial to obtain and utilise a larger cluster galaxy sample when training our background subtraction model, which would hopefully reduce this effect by exposing the model to more examples. It may also be beneficial to employ an algorithm that can learn to interpolate regions within colour-magnitude space in order to account for the larger photometric errors at higher redshifts, such as a variational autoencoder (Kingma & Welling, 2013).

In this work, we used the Schechter function to fit to the luminosity distribution of identified cluster galaxies, where it was important to review the cluster membership status of each individual galaxy in order to minimise severe contamination from bright interloping field galaxies when fitting the Schechter function. Although, we were aware of alternative luminosity functions that could be used to fit to the luminosity distribution of cluster galaxies. Two other commonly used luminosity functions[42] include the Zwicky function (Zwicky, 1957) and Abell function (Abell, 1975). Briefly, the Zwicky function is fitted by considering the difference in magnitude of each cluster galaxy from the brightest cluster galaxy whereas the Abell function is fitted by combining two separately fitted analytical functions. This means that the Zwicky function requires identifying the brightest cluster galaxy beforehand whereas the Abell function is not continuous at all luminosities. We decided to use the Schechter function over these other luminosity functions because the Schechter function did not have strict prerequisite

[42]We recommend the reader to refer to Sarazin (1986) for an overview of different luminosity functions.

conditions and offers continuity (i.e. it was composed of a power law and an exponential function) at all luminosities (Sarazin, 1986).

We fitted the Schechter function to a composite luminosity distribution that consisted of a subsample of identified cluster galaxies with high completeness to obtain best fit parameter values of $M^* = -22.81$ with a standard deviation of $\pm 0.5$; $n^* = 159.82$ with a standard deviation of $\pm 154.62$ and $\alpha = -1.99$ with a standard deviation of $\pm 0.37$. We did not allow $\alpha$ to be greater than $-1$ or lesser than $-2$ when performing Chi-squared fitting, as we assumed that it would be unphysical for the number of cluster galaxies to be decreasing or increasing rapidly at fainter magnitudes respectively. We also did not set any specific bounds for $M^*$ and $n^*$, since these parameters were more dependent on the given data. We attempted to compare our best fit parameter values for $M^*$ and $\alpha$ to the best fit parameter values of $M^*$ and $\alpha$ found in the literature from cluster studies to determine whether our best fit parameter values for $M^*$ and $\alpha$ were appropriate as 'universal' values. However, we found that the literature contained a wide range of values for $M^*$ and $\alpha$ that depended on a variety of different factors (e.g. photometric system used, magnitude range examined, redshift range examined, cluster mass range examined, composition of galaxy types in cluster sample, background subtraction method used). Although, we noticed that some typical values obtained for $M^*$ and $\alpha$ span approximately from $-23$ to $-20$ and $-2.1$ to $-0.8$ respectively (e.g. Oegerle et al. 1987; Oegerle & Hoessel 1989; Valotto et al. 1997; Wilson et al. 1997; Rauzy et al. 1998; Paolillo et al. 2001; Yagi et al. 2002; de Propris et al. 2003; Popesso et al. 2005; González et al. 2006; Alshino et al. 2010; de Filippis et al. 2011; Moretti et al. 2015; Lan et al. 2016). This suggested that our assumptions for $M^*$ and $\alpha$ were not unreasonable.

We remind the reader that our approach for estimating cluster richness was based only on the number of cluster galaxies identified by our background subtraction model within a defined magnitude range and given search area. This would be particularly beneficial for cosmological studies (Sarazin, 1986), such as comparing simulated and observed halo mass functions (e.g. Castro et al. 2016; Yennapureddy & Melia 2019), since it would reduce the complexity of modeling an appropriate selection function to correct for biases from post-processing (e.g. incorrect star/galaxy classification, deblending/interpolation issues, misestimated

photometric redshifts) or survey conditions (e.g. flux limitations, oversaturation by bright stars, different aperture sizes) (Melin et al., 2005). In addition, our background subtraction method provides robustness when estimating cluster richness along any line-of-sight environment since it assesses the cluster membership status of galaxies based only on their photometric measurements. This is not easily achievable when using simple statistical-based or colour-based background subtraction methods. Furthermore, our background subtraction method does not require us to make any assumptions about the properties of the cluster and field galaxies since these properties are self-learned by the AE algorithm. This means that our background subtraction method is not intrinsically biased towards selecting different galaxy types.

In future work, it would be interesting to examine the applicability of our background subtraction method on different usage cases. These include studying the properties and evolution of identified cluster galaxies or deciding spectroscopic follow-ups of potential galaxy members in clusters or measuring the observed radial density, luminosity and redshift profiles of clusters. We also aim to extend this current work by also establishing an empirical scaling relation between our richness estimates and cluster dark matter halo masses, that have been inferred via weak gravitational lensing, in order to construct an observed halo mass function for constraining cosmological parameters. In addition, we intend to integrate our background subtraction method with our own cluster finder model (i.e. Deep-CEE) and photometric redshift estimator model (i.e. Z-Sequence) to mask or remove interloping line-of-sight galaxies in image data or photometric catalogue data respectively to further minimise their model predictions errors.

We note that there are various other types of conventional machine learning algorithms[43] available, aside from AE's, which could be used for the task of performing background subtraction. These could include the K-nearest neighbours algorithm, K-means algorithm, isolation forest algorithm (Liu et al., 2008), support vector machine algorithm and XGBoost algorithm (Chen & Guestrin, 2016). The reason we chose to utilise an AE over other conventional machine learning algorithms was due to the fact that an AE is a deep neural network, which is

[43]We recommend the reader to refer to `https://pyod.readthedocs.io/en/latest/pyod.models.html` for an extensive list of outlier detection algorithms (Zhao et al., 2019).

capable of self-learning the importance of input features. On the other hand, most conventional machine learning algorithms require important features to be manually extracted in order to attain good predictive performance, which can be time-consuming and difficult to do when there are many complex features (Liang et al. 2017; Notley & Magdon-Ismail 2018; O' Mahony et al. 2019; Liu et al. 2022a).

When training our background subtraction model, we used a Monte Carlo cross-validation strategy to determine an optimal hyper-parameter combination that offered the best predictive performance possible across different weight initialisations by performing random subsampling of our training and validation sets. Although, this may have resulted in some galaxies not being utilised at all (i.e. if the galaxy was not randomly chosen to be in any of our training, validation or test sets), which is not maximising data efficiency. In future work, we could employ a k-fold cross-validation strategy for hyper-parameter tuning and model evaluation. This would improve data efficiency and model generalisation as our background subtraction model would be evenly examined across all available data during its training and testing phases.

We performed permutation feature importance testing to determine which input features were deemed as important by our background subtraction model when identifying 'actual' cluster galaxies. From which, we found that the following input features displayed high significance: $g$, $r$, $i$, $z$, $u-g$, $u-r$, $g-i$ and $g-z$. This tells us that our background subtraction model had learned to utilise most of the available photometric information in high dimensional colour-magnitude space. This was more efficient than only utilising a two dimensional colour-magnitude diagram, which is typically used when attempting to detect cluster galaxies within colour-magnitude space (e.g. Yee et al. 1999; Gladders & Yee 2005; Stott et al. 2009; Valentinuzzi et al. 2011). We believe that these specific input features were important to our background subtraction model due to two main reasons. Firstly, in Figure 4.14 it can be seen that the majority of the cluster and field galaxy population can be distinguished via filter magnitudes within colour-magnitude space. This explained why our background subtraction model prioritised several filter magnitudes when performing background subtraction. Secondly, in Figure 4.14 it can also be seen that a minority of cluster galaxies

overlapped with field galaxies within colour magnitude space. This explained why our background subtraction model also prioritised several colours in combination with the filter magnitudes to distinguish between these overlapping cluster and field galaxies. Based on these reasons, it is not unreasonable to assume that our background subtraction model can recognise the broad spectral features (e.g. 4000$\mathring{A}$ break) and overall shape of the observed spectral energy distribution[44] of cluster galaxies[45] at different redshifts.

In future work, it would be interesting to examine the impact from including additional features such as galaxy sizes, morphology and surface brightness as inputs for our background subtraction model. However, we note that we cannot easily reapply our method to galaxy surveys that do not readily provide information for all our required input features. Furthermore, our background subtraction model is not provided with redshift information as an input feature when distinguishing between cluster and field galaxies. Instead, we wanted our background subtraction model to self-learn about the photometric properties of cluster galaxies belonging to different redshift intervals, which is similar to how photometric redshifts of individual galaxies are estimated by empirical algorithms.

## 4.5 Conclusion

We present a proof-of-concept study of AutoEnRichness, a hybrid empirical and analytical approach that uses a multi-stage machine learning algorithm and a conventional luminosity distribution fitting approach to perform background subtraction and estimate cluster richnesses respectively. We utilised photometric data from the SDSS-IV DR16 to train our background subtraction model, which learned to reconstruct the photometry of cluster galaxies in order to distinguish between cluster and field galaxies. We then examined the predictive performance of our background subtraction model at distinguishing between cluster and field galaxies in a test set, which resulted in a balanced accuracy of 83.20 per cent.

[44] We recommend the reader to refer to Kennicutt (1992) for further details on the observed spectral energy distribution of different galaxies.

[45] Clusters typically have a majority population of elliptical and lenticular galaxies with a minority population of spiral galaxies (Dressler, 1980b).

Subsequently, we constructed a scaling relation that estimated $r_{200}$ when given the number of cluster galaxies identified by our background subtraction model within a search radius of 2.5 Mpc at each cluster's spectroscopic redshift. We utilised this learned scaling relation to resample galaxies within an $r_{200}$ radius for each cluster. Next, we fitted the Schechter function to a composite luminosity distribution that consisted of a subsample of cluster galaxies identified by our background subtraction model within $r_{200}$ that had high completeness. We used a Chi-squared fitting approach to determine an optimal $r$ filter absolute magnitude bin size of 0.52 and best fit parameter values of $M^* = -22.81$ with a standard deviation of $\pm 0.5$; $n^* = 159.82$ with a standard deviation of $\pm 154.62$ and $\alpha = -1.99$ with a standard deviation of $\pm 0.37$. We then used the optimal $r$ filter absolute magnitude bin size and best fit parameter values for $M^*$ and $\alpha$ to fit the Schechter function to the luminosity distribution of individual clusters. We estimated cluster richnesses within $r_{200}$ by computing the integral of the locally fit Schechter function. Lastly, we applied the optimal $r$ filter absolute magnitude bin size and best fit parameter values for $M^*$ and $\alpha$ to another test set of clusters to obtain a median absolute percentage error of 33.50 per cent between our estimated cluster richnesses and WH15 richnesses within $r_{200}$. We note that the only cluster prerequisites for AutoEnRichness were the astronomical coordinates of the approximate cluster location as well as an initial cluster redshift estimate for computing appropriate cluster radii. We intend for AutoEnRichness to be combined with the Deep-CEE and Z-Sequence models to obtain the key measurements (i.e. position from cluster detection and distance from redshift estimation respectively) needed for conducting astrophysics and cosmology research. In future work, it would be beneficial to develop a data pipeline that integrates AutoEnRichness with these other methods into an end-to-end process in preparation for usage on upcoming large-scale galaxy surveys.

# Chapter 5

# Conclusion

## 5.1 Summary of our findings

In this thesis, we explored how galaxy cluster cataloguing strategies can be infused with modern data science techniques. The underlying objective of conducting our proof-of-concept studies was to provide researchers with new data-driven tools that can improve the overall efficiency of data-processing in large-scale galaxy surveys. We also demonstrated that the tools yielded prediction errors that were comparable to existing cluster cataloguing methods. To achieve these tasks, we decided to employ specific machine learning algorithms that we believed were highly applicable for performing cluster detection, redshift estimation and richness estimation. We will now summarise the key findings of our research from chapters 2, 3 and 4.

### 5.1.1 Galaxy cluster detection

In chapter 2 we introduced a cluster detection model known as Deep-CEE. We described how an object detection algorithm can be utilised to identify clusters in wide-field colour images as well as predict the astronomical coordinates of cluster cores in each detection. The motivation behind this work was to investigate the possibility of replicating the behaviour of traditional human identification of

clusters (e.g. Abell clusters) with an automated machine. We decided to employ the Faster R-CNN multi-stage algorithm for this task. We note that we did not focus on optimising the hyper-parameters of this algorithm since we were instead concentrated on examining its applicability.

We initially exposed the algorithm to examples of the core regions of known clusters (N.B. between a photometric redshift range of $0.1 < z < 0.2$ and above a minimum number of observed galaxies) and random background regions in wide-field colour images that were created from combining photometric imaging of *g, r, i* filters. The algorithm learned to process images by first simplifying the images into its low-level features (e.g. straight lines, curves, edges, blobs) and then proposed regions within the simplified images that 'look' like the visual features of cluster cores rather than random background regions. A confidence score was subsequently generated for each proposed region, which allowed us to manually set a confidence score threshold to maximise the precision and recall of finding known clusters.

We obtained an F1 score of approximately 80 per cent for cluster identification when we applied the algorithm on images of unseen redMaPPer clusters. This indicated that the algorithm could identify many of the known clusters with relatively few misclassifications, which is somewhat comparable to the estimated $70-85$ per cent completeness of the Abell catalogue (Lucey, 1983). We obtained a standard error of regression value of approximately 12 kpc when comparing our predicted astronomical coordinates of true positive detections with the known astronomical coordinates of the redMaPPer clusters, that had their image positions randomly translated. This suggested that the algorithm could adequately locate the core region of clusters, where the optical core radius of clusters is approximately 250 kpc.

The primary benefit of adopting this approach for cluster detection is that there is no need for making prior assumptions on the physical appearance of observed objects since the importance of different visual features is automatically learned by the algorithm, which makes it more straightforward when modeling a selection function. In addition, this approach also minimises the necessity for performing extensive pre-processing since it works directly on images rather than requiring to wait for the creation of photometric galaxy catalogues, which are

typically used by researchers for cluster detection. Furthermore, this approach can be quickly deployed to discover many new clusters in upcoming large-scale galaxy surveys since the only training prerequisite for the algorithm is that it needs to be tuned on a set of survey images containing known clusters.

### 5.1.2 Galaxy cluster redshift estimation

In chapter 3 we introduced a cluster redshift estimation model known as Z-Sequence. We described how an ensemble regression algorithm can be utilised to estimate the photometric redshifts of clusters from photometric data. The motivation behind this work was to develop an approach that had no reliance on individual galaxy redshifts or specific photometric features whilst also being able to accurately predict the photometric redshift of clusters. We decided to employ the sequential random k-nearest neighbours algorithm for this task. We preferred to use an ensemble of k-nearest neighbour algorithms rather than an individual k-nearest neighbour algorithm due to the individual k-nearest neighbour algorithm having greater sensitivity to minor changes in the training set. We note that the k-nearest neighbour algorithm did not learn internal parameters to make predictions but it instead computed the average of the 'k' nearest neighbour target labels in a training set to make predictions.

We initially created a photometric training set (i.e. filter magnitudes and colours) of line-of-sight galaxies within a fixed search radius of the core region of known clusters (N.B. between a photometric redshift range of $0.05 \leq z \leq 0.6$ and above a minimum number of observed galaxies). We assigned target labels for each galaxy based on the photometric redshift of its host cluster. This assumed that line-of-sight galaxies were at the same redshift as the host cluster. We performed bootstrap sampling with replacement of the training set to generate a randomly sampled training set for each of the internal k-nearest neighbour algorithms. The output of the ensemble was computed from the combined median of all the 'k' nearest neighbour target labels of each internal k-nearest neighbour algorithm. In addition, we implemented a feature selection strategy, known as sequential forward selection, to allow the internal k-nearest neighbour algorithms to automatically learn appropriate combinations of photometric features that

minimised the photometric redshift prediction error. Furthermore, we tuned the hyper-parameters of the sequential random k-nearest neighbours algorithm to also minimise the photometric redshift prediction error.

We obtained a photometric redshift prediction error (i.e. median value of $|\Delta z|/(1+z)$) of approximately 0.011 when we applied the algorithm onto the photometric data of unseen WHL12 and redMaPPer clusters using a 10 arcseconds search radius. We note that the photometric redshift prediction error increased by approximately 30 to 50 per cent when the search radius was enlarged from 10 arcseconds to 21 and 32 arcseconds respectively, which was likely due to the inclusion of additional interloping field galaxies. It should be noted that we compared our estimated photometric redshifts with the photometric redshift estimates obtained by the WHL12 and redMaPPer methods to measure our photometric redshift prediction error. Meanwhile, the WHL12 and redMaPPer clusters compared their estimated photometric redshifts with spectroscopic redshifts to measure their photometric redshift prediction errors. The reason we decided to use the photometric redshift of clusters instead of the spectroscopic redshift of clusters was due to there being at least three times as many more clusters with photometric redshifts available. From which, these results indicated that the algorithm could precisely estimate the photometric redshift of clusters based only on the similarity of observed galaxies within colour-magnitude space, where our photometric redshift prediction error was still on par with the photometric redshift prediction errors obtained by the WHL12 and redMaPPer clusters of approximately 0.008 and 0.007 respectively.

This approach comes with multiple practical benefits when attempting to estimate the photometric redshift of clusters. Firstly, it does not utilise individual galaxy redshifts since it assumes that line-of-sight galaxies have the same redshift as the host cluster, which means the algorithm is practical in survey regions that have no known individual galaxy redshifts. This also minimises the need to wait for the creation of galaxy redshift catalogues. Secondly, it is compatible with any input combination of photometric features since the algorithm automatically determines the most appropriate photometric features for accurately estimating the photometric redshift of clusters, which means that no prior assumptions are needed for the photometry of observed galaxies. As such, we can efficiently work

with incomplete filter sets. Thirdly, it can be easily combined with cluster finders (e.g. Deep-CEE) that do not intrinsically estimate the redshift of clusters since the algorithm only needs to know the astronomical coordinate of the core region of clusters to effectively start making predictions. Furthermore, this approach can be quickly deployed to estimate the photometric redshift of many clusters in upcoming large-scale galaxy surveys since the only training prerequisite for the algorithm is that it needs to be tuned on the survey photometry of line-of-sight galaxies around the core region of known clusters.

### 5.1.3 Galaxy cluster richness estimation

In chapter 4 we introduced a cluster richness estimation model that is nominally known as AutoEnRichness. We described how a reconstruction algorithm can be utilised to distinguish between cluster and field galaxies from photometric data and we also described how an analytical function can be utilised to estimate the richness of clusters based on the luminosity of identified cluster galaxies. The motivation behind this work was to develop an approach that determined whether line-of-sight galaxies were associated to clusters in order to minimise severe contamination when estimating the richness of clusters. We decided to employ the autoencoder algorithm and Schechter function for this task. We note that the reason we used the Schechter function was to ensure that we accounted for incompleteness in the luminosity distribution of cluster galaxies when estimating the richness of clusters, where the Schechter function in this case described the number of cluster galaxies per unit magnitude.

We initially obtained a photometric training set of identified cluster galaxies (N.B. between a photometric redshift range of $0.1 \leq z \leq 0.35$) from an existing cluster galaxy catalogue. We also obtained a photometric training set of random field galaxies from manually identified field regions. The algorithm was specifically trained to recreate the photometry of cluster galaxies by first compressing the photometric data of cluster galaxies into a reduced feature space of low-level photometric features and then the algorithm decompressed the photometric data into its original format. This encouraged the algorithm to learn internal parameters that minimised the difference between the input and output values of cluster

galaxies, such that smaller differences indicated a cluster galaxy whilst larger differences indicated a field galaxy. The difference between the input and output values was subsequently passed to a logistic regression algorithm that was trained to perform classification. This yielded a probability of whether the input data was more likely to be a cluster or field galaxy, which allowed us to manually set a probability threshold to maximise the precision and recall of finding known cluster galaxies. We also tuned the hyper-parameters of the autoencoder algorithm to maximise the area under the precision-recall curve. Next, we counted the number of cluster galaxies found by the algorithm within a 2.5 Mpc search radius from the core region of known clusters. We then established a scaling relation that predicted a characteristic radius of $r_{200}$ for clusters, where the scaling relation was constructed between the number of identified cluster galaxies and $r_{200}$ values of known clusters. Lastly, we resampled galaxies within $r_{200}$ of the known clusters to produce a composite luminosity distribution from $r$ filter magnitudes of identified cluster galaxies within $r_{200}$. We fitted the Schechter function to the composite luminosity distribution to determine universal parameter values of the Schechter function. We reutilised these universal parameter values to refit the Schechter function to the luminosity distribution of individual clusters. It should be noted that we applied a completeness correction of the faint magnitude bins when refitting the Schechter function to individual clusters. We then integrated the fitted Schechter function to compute the richness of clusters within $r_{200}$, where we only considered cluster galaxies that were brighter than an absolute $r$ filter magnitude limit of $-20.5$.

We obtained an F1 score of approximately 73 per cent for only cluster galaxy identification and a balanced accuracy of 83 per cent for cluster and field galaxy classification when we applied the algorithm onto the photometric data of unseen cluster and field galaxies from the AMF11 clusters and our proposed field regions. This implied that the algorithm could identify many of the known cluster galaxies with relatively few misclassifications as well as the algorithm consistently being able to accurately distinguish between cluster and field galaxies. We obtained a root mean squared error of approximately 18 and a median absolute percentage error of approximately 33 per cent when comparing our estimated cluster

richnesses within $r_{200}$ with known cluster richnesses within $r_{200}$ from WH15 clusters. Furthermore, we obtained a strong linear correlation between our estimated cluster richnesses within $r_{200}$ and redMaPPer cluster richnesses, which were not scaled within a characteristic radius. These results indicated that the algorithm could precisely estimate the richness of clusters, given that the richness range of clusters in our test set approximately spanned from a few to several hundred cluster galaxies within $r_{200}$. However, it was difficult to directly compare the precision of our richness estimates with the precision obtained by other conventional cluster richness estimation methods since they did not state the average difference between their estimated and known richnesses.

This approach comes with multiple practical benefits when attempting to estimate the richness of clusters. Firstly, it does not make prior assumptions on the photometric properties of cluster and field galaxies since the importance of different photometric features is automatically learned by the algorithm, where the photometric properties of cluster and field galaxies becomes more difficult to visibly separate at fainter magnitudes such that manual feature selection becomes extremely inefficient. Secondly, it compares the photometric properties of cluster and field galaxies when performing background subtraction, which is advantageous over statistical background subtraction methods that do not examine the true membership status of galaxies. Thirdly, it can be easily combined with cluster finders (e.g. Deep-CEE) or cluster redshift estimators (e.g. Z-Sequence) that do not intrinsically perform background subtraction since the algorithm only needs to know the astronomical coordinate of the core region of clusters to effectively start making predictions. Furthermore, this approach can be quickly deployed to estimate the richness of many clusters in upcoming large-scale galaxy surveys since the only training prerequisite for the algorithm is that it needs to be tuned on the survey photometry of galaxies that are likely to be from clusters and field regions. This approach also requires an approximated redshift for clusters in order to establish a scaling relation for predicting the $r_{200}$ of clusters.

## 5.2 Future work

The research presented in this thesis is composed of several of our initial proof-of-concept studies for conducting cluster cataloguing with modern data science techniques. As such, we are aware that there are still further considerations and enhancements that can be made to our existing models. It would also be beneficial to consider the possibilities of applying machine learning to catalogue additional properties of clusters.

The following points briefly describes some of the potential considerations, enhancements and possibilities for future work based on our current findings:

- Explore using different combinations of photometric filters when creating wide-field colour images for the Deep-CEE model.
- Compare the effectiveness from using different object detection algorithms for cluster detection in the Deep-CEE model.
- Examine the significance of tuning the hyper-parameters for the Faster R-CNN algorithm in the Deep-CEE model.
- Assess the importance of visual features that are learned by the Faster R-CNN algorithm in the Deep-CEE model.
- Modify the Deep-CEE model to yield uncertainty estimates for the predicted astronomical coordinates of clusters.
- Compare the stability of the Deep-CEE model from using k-fold and Monte Carlo cross-validation strategies to create different training and testing sets.
- Analyse the implications from using training and testing samples of clusters for the Deep-CEE model that are found by X-ray, SZ effect and weak gravitational lensing detection methods.
- Determine the detection rate of clusters identified by the Deep-CEE model that are also found by X-ray, SZ effect and weak gravitational lensing detection methods.

- Use Monte Carlo simulations to approximately measure the completeness of cluster detections by the Deep-CEE model.
- Examine the significance from using a larger 'k' value of the number of nearest neighbours for the SRKNN algorithm in the Z-Sequence model.
- Compare the effectiveness from using different nearest neighbour weighting schemes for the SRKNN algorithm in the Z-Sequence model.
- Evaluate the overall photometric redshift prediction error from using photometric redshifts versus spectroscopic redshifts of clusters in the training sample of the Z-Sequence model.
- Assess the stability of the SRKNN algorithm in the Z-Sequence model from using k-fold and Monte Carlo cross-validation strategies to create different training and testing sets.
- Integrate the background subtraction algorithm from the AutoEnRichness model into the Deep-CEE and Z-Sequence models to mask/remove interloping field galaxies when detecting clusters and estimating cluster redshifts.
- Compare prediction errors for the Deep-CEE and Z-Sequence models before and after removing interloping field galaxies along the line-of-sight of clusters.
- Examine the significance from tuning the hyper-parameters of the logistic regression algorithm in the AutoEnRichness model.
- Compare the effectiveness from using different outlier detection algorithms in the AutoEnRichness model.
- Evaluate the overall richness prediction error from establishing a scaling relation to predict $r_{200}$ that employs X-ray determined $r_{200}$ values in the AutoEnRichness model.
- Assess the stability of the background subtraction algorithm in the AutoEnRichness model from using a k-fold cross-validation strategy to create different training and testing sets.

- Examine the significance from using different training and testing sets of clusters when determining universal parameter values of the Schechter function in the AutoEnRichness model.
- Incorporate uncertainties from observed photometry errors into all of our models.
- Assess how prediction errors are impacted from using Bayesian optimisation to tune the hyper-parameters in our models.
- Train all of our models on simulated data in order to understand the output of the algorithms in more detail.
- Create a data pipeline that combines the outputs of all our models to measure the time taken to process large quantities of observational data in preparation for upcoming large-scale galaxy surveys.
- Probe the limitations of applying our models on a larger positional, distance and mass coverage.
- Investigate the prospect of utilising machine learning to predict the morphology and dynamical state of clusters.

# Appendix A

# Appendices

| Step | Total loss | RPN objectness loss | RPN box regression loss | DN classification loss | DN box regression loss |
|---|---|---|---|---|---|
| 147 | 0.5954 | 0.3825 | 0.0535 | 0.1162 | 0.0432 |
| 204 | 0.3448 | 0.2023 | 0.0538 | 0.0605 | 0.0282 |
| 313 | 0.3185 | 0.1242 | 0.0515 | 0.0794 | 0.0634 |
| 450 | 0.2841 | 0.0834 | 0.0487 | 0.0994 | 0.0526 |
| 581 | 0.2559 | 0.0694 | 0.0436 | 0.0838 | 0.0591 |
| 934 | 0.1641 | 0.0441 | 0.0380 | 0.0552 | 0.0268 |
| 1065 | 0.1347 | 0.0401 | 0.0302 | 0.0367 | 0.0277 |
| 1361 | 0.1399 | 0.0344 | 0.0283 | 0.0543 | 0.0229 |
| 1455 | 0.2004 | 0.0327 | 0.0268 | 0.1129 | 0.0280 |
| 1582 | 0.1242 | 0.0302 | 0.0279 | 0.0443 | 0.0217 |
| 1709 | 0.1049 | 0.0307 | 0.0256 | 0.0274 | 0.0212 |
| 1836 | 0.1465 | 0.0280 | 0.0261 | 0.0646 | 0.0277 |
| 1960 | 0.1344 | 0.0281 | 0.0301 | 0.0531 | 0.0232 |
| 2083 | 0.1090 | 0.0290 | 0.0246 | 0.0377 | 0.0177 |
| 2209 | 0.1322 | 0.0274 | 0.0237 | 0.0521 | 0.0290 |
| 2335 | 0.1540 | 0.0284 | 0.0256 | 0.0712 | 0.0288 |
| 2461 | 0.1096 | 0.0267 | 0.0247 | 0.0365 | 0.0218 |
| 2588 | 0.0942 | 0.0273 | 0.0245 | 0.0244 | 0.0181 |
| 2716 | 0.1184 | 0.0269 | 0.0354 | 0.0339 | 0.0223 |
| 2842 | 0.0986 | 0.0253 | 0.0235 | 0.0324 | 0.0173 |
| 2968 | 0.0969 | 0.0247 | 0.0222 | 0.0310 | 0.0190 |
| 3187 | 0.1048 | 0.0251 | 0.0275 | 0.0326 | 0.0196 |
| 3312 | 0.1293 | 0.0234 | 0.0233 | 0.0611 | 0.0214 |
| 3438 | 0.1125 | 0.0240 | 0.0224 | 0.0458 | 0.0204 |
| 3566 | 0.0900 | 0.0250 | 0.0232 | 0.0243 | 0.0175 |
| 3697 | 0.0901 | 0.0247 | 0.0213 | 0.0256 | 0.0185 |
| 3823 | 0.1237 | 0.0233 | 0.0238 | 0.0548 | 0.0219 |
| 3948 | 0.1189 | 0.0233 | 0.0229 | 0.0520 | 0.0208 |
| 4073 | 0.1059 | 0.0231 | 0.0208 | 0.0420 | 0.0201 |
| 4198 | 0.1015 | 0.0239 | 0.0217 | 0.0337 | 0.0223 |
| 4321 | 0.1177 | 0.0224 | 0.0204 | 0.0509 | 0.0240 |
| 4543 | 0.1118 | 0.0231 | 0.0221 | 0.0461 | 0.0205 |
| 4675 | 0.0943 | 0.0232 | 0.0215 | 0.0294 | 0.0202 |
| 4805 | 0.1042 | 0.0232 | 0.0264 | 0.0297 | 0.0249 |
| 4944 | 0.1024 | 0.0222 | 0.0194 | 0.0398 | 0.0209 |
| 5072 | 0.1038 | 0.0223 | 0.0201 | 0.0423 | 0.0191 |
| 5198 | 0.0881 | 0.0241 | 0.0205 | 0.0253 | 0.0182 |
| 5330 | 0.1047 | 0.0220 | 0.0202 | 0.0425 | 0.0199 |
| 5462 | 0.1275 | 0.0219 | 0.0241 | 0.0541 | 0.0274 |
| 5587 | 0.0969 | 0.0233 | 0.0220 | 0.0294 | 0.0222 |
| 5917 | 0.0972 | 0.0224 | 0.0265 | 0.0265 | 0.0219 |
| 6048 | 0.0943 | 0.0221 | 0.0184 | 0.0344 | 0.0194 |
| 6179 | 0.1038 | 0.0228 | 0.0228 | 0.0313 | 0.0268 |
| 6495 | 0.1066 | 0.0235 | 0.0231 | 0.0352 | 0.0248 |
| 6698 | 0.1035 | 0.0216 | 0.0184 | 0.0446 | 0.0189 |
| 6824 | 0.1022 | 0.0223 | 0.0190 | 0.0427 | 0.0182 |
| 6951 | 0.1045 | 0.0217 | 0.0187 | 0.0460 | 0.0180 |
| 7077 | 0.1029 | 0.0217 | 0.0188 | 0.0378 | 0.0247 |
| 7204 | 0.0910 | 0.0208 | 0.0180 | 0.0341 | 0.0181 |
| 7332 | 0.0974 | 0.0215 | 0.0176 | 0.0403 | 0.0181 |
| 7458 | 0.0926 | 0.0249 | 0.0211 | 0.0268 | 0.0198 |

**Table A1:** This table displays the Total, RPN and DN loss values at different steps during the training of our model when evaluated on the test set.

# References


Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, http://tensorflow.org/

Abazajian K., et al., 2004, The Astronomical Journal, 128, 502

Abell G. O., 1958, The Astrophysical Journal Supplement Series, 3, 211

Abell G. O., 1959, Leaflet of the Astronomical Society of the Pacific, 8, 121

Abell G. O., 1965, Annual Review of Astronomy and Astrophysics, 3, 1

Abell G. O., 1975, Stars and Stellar Systems IX: Galaxies and the Universe, p. 601

Abell G. O., Corwin Jr. H. G., Olowin R. P., 1989, The Astrophysical Journal Supplement Series, 70, 1

Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, Monthly Notices of the Royal Astronomical Society, 479, 415

Adelman-McCarthy J. K., et al., 2008, The Astrophysical Journal Supplement Series, 175, 297

Aggarwal C. C., Hinneburg A., Keim D. A., 2001, in Van den Bussche J., Vianu V., eds, Database Theory — ICDT 2001. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 420–434

Aha D. W., Bankert R. L., 1995, R0, 1

Aha D. W., Kibler D., Albert M., 1991, Machine Learning, 6, 37

Ahn C. P., et al., 2012, The Astrophysical Journal Supplement Series, 203, 21

Ahumada R., et al., 2020, The Astrophysical Journal Supplement Series, 249, 3

Aihara H., et al., 2011, The Astrophysical Journal Supplement Series, 193, 29

Alam S., et al., 2015, The Astrophysical Journal Supplement Series, 219, 12

Alpaslan M., et al., 2012, Monthly Notices of the Royal Astronomical Society, 426, 2832

Alshino A., Khosroshahi H., Ponman T., Willis J., Pierre M., Pacaud F., Smith G. P., 2010, Monthly Notices of the Royal Astronomical Society, 401, 941

Altman D. G., Bland J. M., 2005, BMJ, 331, 903

Altman N., Krzywinski M., 2015, Nature Methods, 12, 999

Amodeo S., Ettori S., Capasso R., Sereno M., 2016, Astronomy & Astrophysics, 590, A126

Annis J., et al., 1999, in American Astronomical Society Meeting Abstracts. p. 12.02

Arulkumaran K., Deisenroth M., Brundage M., Bharath A., 2017, IEEE Signal Processing Magazine, 34

Ashby W. R., 1952, Design for a brain. Wiley

Babbage C., Davy H., 1822, A letter to Sir Humphry Davy, Bart., President of the Royal Society, from Charles Babbage, Esq., M.A., F.R.S., on the Application of Machinery to the Calculating and Printing Mathematical Table

Babbedge T. S. R., et al., 2004, Monthly Notices of the Royal Astronomical Society, 353, 654

Bahcall N. A., 1999, in Dekel A., Ostriker J. P., eds, Formation of Structure in the Universe. p. 135

Bailey S. I., 1908, Annals of Harvard College Observatory, 60, 199

Baldi P., Sadowski P., Whiteson D., 2014, Nature Communications, 5, 4308

Baldry I. K., et al., 2010, Monthly Notices of the Royal Astronomical Society, 404, 86

Balland C., Blanchard A., 1995, arXiv e-prints, pp astro–ph/9510130

Barati Farimani A. H. M., Aluru N. R., 2018, npj 2D Materials and Applications, 2

Bautz L. P., Morgan W. W., 1970, The Astrophysical Journal, 162, L149

Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, Monthly Notices of the Royal Astronomical Society, 460, 1371

Bellman R., 1961, Adaptive Control Processes: A Guided Tour. Princeton Legacy Library, Princeton University Press

Bengio Y., Grandvalet Y., 2004, The Journal of Machine Learning Research, 5, 1089

Bentley J. L., 1975, Communications of the ACM, 18, 509

Bergstra J., Bengio Y., 2012, The Journal of Machine Learning Research, 13, 281

Bilicki M., et al., 2018, Astronomy & Astrophysics, 616, A69

Blake C., et al., 2008, Astronomy & Geophysics, 49, 5.19

Bocquet S., et al., 2019, The Astrophysical Journal, 878, 55

Böhringer H., et al., 2004, Astronomy & Astrophysics, 425, 367

Bolzonella M., Miralles J. M., Pelló R., 2000, Astronomy & Astrophysics, 363, 476

Bottou L., 2010, in Lechevallier Y., Saporta G., eds, Proceedings of COMPSTAT'2010. Physica-Verlag HD, Heidelberg, pp 177–186

Boué G., Adami C., Durret F., Mamon G. A., Cayatte V., 2008, Astronomy & Astrophysics, 479, 335

Boyd K., Eng K. H., Page C. D., 2013, in Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III. ECMLPKDD'13. Springer-Verlag, Berlin, Heidelberg, pp 451–466, doi:10.1007/978-3-642-40994-3˙29, https://doi.org/10.1007/978-3-642-40994-3_29

Breiman L., 2001, Machine Learning, 45, 5

Breiman L., Friedman J., Olshen R., Stone C., 1984, Classification And Regression Trees, doi:10.1201/9781315139470.

Briscoe E., Feldman J., 2011, Cognition, 118, 2

Brodersen K. H., Ong C. S., Stephan K. E., Buhmann J. M., 2010, in 2010 20th International Conference on Pattern Recognition. pp 3121–3124, doi:10.1109/ICPR.2010.764

Bruzual G., Charlot S., 2003, Monthly Notices of the Royal Astronomical Society, 344, 1000

Bühlmann P., 2012, Handbook of Computational Statistics, pp 985–1022

Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001, Monthly Notices of the Royal Astronomical Society, 321, 559

Calvi R., Poggianti B. M., Vulcani B., 2011, Monthly Notices of the Royal Astronomical Society, 416, 727

Carrasco Kind M., Brunner R. J., 2013, Monthly Notices of the Royal Astronomical Society, 432, 1483

Castro T., Marra V., Quartin M., 2016, Monthly Notices of the Royal Astronomical Society, 463, 1666

Chalmers C., Fergus P., Wich S., Longmore S. N., 2021, arXiv e-prints, p. arXiv:2103.07276

Charnock T., Moss A., 2017, The Astrophysical Journal Letters, 837, L28

Chase Lipton Z., Elkan C., Narayanaswamy B., 2014, arXiv e-prints, p. arXiv:1402.1892

Chen T., Guestrin C., 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. Association for Computing Machinery, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785

Chollet F., et al., 2015, Keras, https://keras.io

Cohn J. D., Battaglia N., 2019, Monthly Notices of the Royal Astronomical Society, 491, 1575

Colless M., et al., 2001, Monthly Notices of the Royal Astronomical Society, 328, 1039

Comon P., 1994, Signal Processing, 36, 287

Cortes C., Vapnik V., 1995, Machine Learning, 20, 273

Costanzi M., et al., 2018, Monthly Notices of the Royal Astronomical Society, 482, 490

Cover T., Hart P., 1967, IEEE Transactions on Information Theory, 13, 21

Csabai I., Dobos L., Trencséni M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astronomische Nachrichten, 328, 852

Culbertson J. T., 1950, Consciousness and behavior: a neural analysis of behavior and of consciousness. Wm. C. Brown Co.

Cunningham P., Cord M., Delany S. J., 2008, Supervised Learning. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 21–49, doi:10.1007/978-3-540-75171-7˙2, https://doi.org/10.1007/978-3-540-75171-7_2

Curtis H. D., 1918, Publications of Lick Observatory, 13, 9

Dark Energy Survey Collaboration et al., 2016, Monthly Notices of the Royal Astronomical Society, 460, 1270

Davis J., Goadrich M., 2006. , doi:10.1145/1143844.1143874

Davis M., et al., 2003, in Guhathakurta P., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. pp 161–172 (arXiv:astro-ph/0209419), doi:10.1117/12.457897

Dey A., et al., 2019, The Astronomical Journal, 157, 168

Diaferio A., Geller M. J., 1997, The Astrophysical Journal, 481, 633

Diaferio A., Geller M. J., Rines K. J., 2005, The Astrophysical Journal, 628, L97

Dickey J. M., Keller D. T., Pennington R., Salpeter E. E., 1987, The Astronomical Journal, 93, 788

Dietrich J. P., Erben T., Lamer G., Schneider P., Schwope A., Hartlap J., Maturi M., 2007, Astronomy & Astrophysics, 470, 821

Dietterich T. G., 2000, in Multiple Classifier Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–15

Djorgovski S. G., Weir N., Fayyad U., 1994, in Crabtree D. R., Hanisch R. J., Barnes J., eds, Astronomical Society of the Pacific Conference Series Vol. 61, Astronomical Data Analysis Software and Systems III. p. 195

Dong F., Pierpaoli E., Gunn J. E., Wechsler R. H., 2008, The Astrophysical Journal, 676, 868

Doré O., et al., 2016, arXiv e-prints, p. arXiv:1606.07039

Dressler A., 1980a, The Astrophysical Journal, 236, 351

Dressler A., 1980b, The Astrophysical Journal, 236, 351

Dressler A., Shectman S. A., 1987, The Astronomical Journal, 94, 899

Dressler A., et al., 1997, The Astrophysical Journal, 490, 577

Dreyer J. L. E., 1888, Memoirs of the Royal Astronomical Society, 49, 1

Duda R. O., Hart P. E., 1973, Pattern classification and scene analysis

Duda R., Hart P., Stork D., 2001, Pattern Classification, 2 edn. John Wiley & Sons, New York

Dudani S. A., 1976, IEEE Transactions on Systems, Man, and Cybernetics, SMC-6, 325

Dunlop J., 1828, Philosophical Transactions of the Royal Society of London Series I, 118, 113

Eales S. A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 481, 1183

Ebeling H., Edge A. C., Henry J. P., 2001, The Astrophysical Journal, 553, 668

Ebeling H., Edge A. C., Mantz A., Barrett E., Henry J. P., Ma C. J., van Speybroeck L., 2010, Monthly Notices of the Royal Astronomical Society, 407, 83

Efron B., 1979, The Annals of Statistics, 7, 1

Efron B., Tibshirani R., 1986, Statistical Science, 1, 54

Efron B., Tibshirani R., 1994, An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, doi:10.1201/9780429246593

Eifler T., et al., 2021, Monthly Notices of the Royal Astronomical Society, 507, 1746

Eisenstein D. J., et al., 2011, The Astronomical Journal, 142, 72

Endsley R., Behroozi P., Stark D. P., Williams C. C., Robertson B. E., Rieke M., Gottlöber S., Yepes G., 2020, Monthly Notices of the Royal Astronomical Society, 493, 1178

Estrada-Piedra T., Torres-Papaqui J. P., Terlevich R., Fuentes O., Terlevich E., 2004, Age determination of the nuclear stellar population of Active Galactic Nuclei using Locally Weighted Regression. Astronomical Society of the Pacific Conference Series Vol. 314, (arXiv:astro-ph/0311627)

Ettori S., Donnarumma A., Pointecouteau E., Reiprich T. H., Giodini S., Lovisari L., Schmidt R. W., 2013, Space Science Reviews, 177, 119

Falco E. E., et al., 1999, The Publications of the Astronomical Society of the Pacific, 111, 438

Farrens S., Abdalla F. B., Cypriano E. S., Sabiu C., Blake C., 2011, Monthly Notices of the Royal Astronomical Society, 417, 1402

Fayyad U. M., Doyle R. J., Weir W. N., Djorgovski S., 1993, in The earth and space Science information System (ESSIS). pp 405–417, doi:10.1063/1.44408

Fix E., 1951, Technical report, Discriminatory analysis: nonparametric discrimination, consistency properties. Randolph Field, Texas, USA

Fotopoulou S., Paltani S., 2018, Astronomy & Astrophysics, 619, A14

Frenk C. S., White S. D. M., Efstathiou G., Davis M., 1990, The Astrophysical Journal, 351, 10

Friedman J., Hastie T., Tibshirani R., 2001, The elements of statistical learning. Vol. 1, Springer New York

Fuentes O., Gulati R. K., 2001, Prediction of Stellar Atmospheric Parameters from Spectra, Spectral Indices and Spectral Lines Using Machine Learning. Revista Mexicana de Astronomia y Astrofisica Conference Series Vol. 10

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, The Astronomical Journal, 111, 1748

Fukushima K., 1980, Biological Cybernetics, pp 193–202

Gail M. H., Green S. B., 1976, Journal of the American Statistical Association, 71, 757

Gallagher John S. I., Ostriker J. P., 1972, The Astronomical Journal, 77, 288

Gavazzi R., Soucail G., 2007, Astronomy & Astrophysics, 462, 459

Geman S., Bienenstock E., Doursat R., 1992, Neural Computation, 4, 1

Ghahramani Z., 2004, Unsupervised Learning. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 72–112, doi:10.1007/978-3-540-28650-9˙5, https://doi.org/10.1007/978-3-540-28650-9_5

Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 2000, The Astrophysical Journal, 544, 616

Girardi M., Biviano A., Giuricin G., Mardirossian F., Mezzetti M., 1995, The Astrophysical Journal, 438, 527

Girshick R., 2015, arXiv e-prints, p. arXiv:1504.08083

Gladders M. D., Yee H. K. C., 2000, The Astronomical Journal, 120, 2148

Gladders M. D., Yee H. K. C., 2005, The Astrophysical Journal Supplement Series, 157, 1

Gladders M. D., López-Cruz O., Yee H. K. C., Kodama T., 1998, The Astrophysical Journal, 501, 571

Glorot X., Bengio Y., 2010, in Teh Y. W., Titterington M., eds, Proceedings of Machine Learning Research Vol. 9, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, Chia Laguna Resort, Sardinia, Italy, pp 249–256, https://proceedings.mlr.press/v9/glorot10a.html

González R. E., Lares M., Lambas D. G., Valotto C., 2006, Astronomy & Astrophysics, 445, 51

Gorecki A., Abate A., Ansari R., Barrau A., Baumont S., Moniez M., Ricol J.-S., 2014, Astronomy & Astrophysics, 561, A128

Goto T., et al., 2002, The Astronomical Journal, 123, 1807

Goto T., et al., 2003, Publications of the Astronomical Society of Japan, 55, 739

Goutte C., Gaussier E., 2005, in Proceedings of the 27th European Conference on Advances in Information Retrieval Research. ECIR'05. Springer-Verlag, Berlin, Heidelberg, pp 345–359, doi:10.1007/978-3-540-31865-1˙25, https://doi.org/10.1007/978-3-540-31865-1_25

Gunn J. E., Gott J. Richard I., 1972, The Astrophysical Journal, 176, 1

Gupta N., Reichardt C. L., 2020, The Astrophysical Journal, 900, 110

Gutierrez-Osuna R., 2000, Statistical Pattern Recognition, https://pdfs.semanticscholar.org/f607/88c6be96a8f62df7a8d8d65824a63f465415.pdf

Hall D., Dayoub F., Kulk J., McCool C., 2017, arXiv e-prints, p. arXiv:1702.01247

Hansen S. M., McKay T. A., Wechsler R. H., Annis J., Sheldon E. S., Kimball A., 2005, The Astrophysical Journal, 633, 122

Harvey A. S., Fotopoulos G., 2016, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 41B8, 423

Hayek F. A., 1952, The sensory order. University of Chicago Press

He K., Zhang X., Ren S., Sun J., 2015. p. arXiv:1502.01852 (arXiv:1502.01852)

Hebb D. O., 1949, The organization of behavior: a neuropsychological theory. Science editions

Herschel W., 1785, Philosophical Transactions of the Royal Society of London Series I, 75, 213

Herschel J. F. W., 1864, Philosophical Transactions of the Royal Society of London Series I, 154, 1

Hetterscheidt M., Erben T., Schneider P., Maoli R., van Waerbeke L., Mellier Y., 2005, Astronomy & Astrophysics, 442, 43

Heydon-Dumbleton N. H., Collins C. A., MacGillivray H. T., 1989, Monthly Notices of the Royal Astronomical Society, 238, 379

Hilton M., et al., 2021, The Astrophysical Journal Supplement Series, 253, 3

Ho M., Rau M. M., Ntampaka M., Farahi A., Trac H., Póczos B., 2019, The Astrophysical Journal, 887, 25

Hoekstra H., Bartelmann M., Dahle H., Israel H., Limousin M., Meneghetti M., 2013, Space Science Reviews, 177, 75

Hoerl A. E., Kennard R. W., 1970, Technometrics, 12, 55

Hosang J., Benenson R., Schiele B., 2017, arXiv e-prints, p. 1705.02950

Hsu C.-w., Chang C.-c., Lin C.-J., 2003, A Practical Guide to Support Vector Classification

Hu L., Huang M., Ke S., Tsai C., 2016, SpringerPlus, 5, 1304

Huang J., et al., 2016, arXiv e-prints, p. arXiv:1611.10012

Hubble E. P., 1926, The Astrophysical Journal, 64, 321

Hubble E. P., 1929a, Proceedings of the National Academy of Science, 15, 168

Hubble E. P., 1929b, The Astrophysical Journal, 69, 103

Huchra J. P., Geller M. J., 1982, The Astrophysical Journal, 257, 423

Huchra J. P., et al., 2012, The Astrophysical Journal Supplement Series, 199, 26

Ilbert O., et al., 2009, The Astrophysical Journal, 690, 1236

Innes R. T. A., 1924, Circular of the Union Observatory Johannesburg, 61, 243

Ishikawa S., et al., 2020, The Astrophysical Journal, 904, 128

Ivezić Ž., et al., 2019, The Astrophysical Journal, 873, 111

Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, Monthly Notices of the Royal Astronomical Society, 321, 372

Johnston D. E., et al., 2007, arXiv e-prints, p. arXiv:0709.1159

Jones D. H., et al., 2009, Monthly Notices of the Royal Astronomical Society, 399, 683

Kang H., 2013, Korean Journal of Anesthesiology, 64, 402

Kauffmann G., et al., 2003, Monthly Notice of the Royal Astronomical Society, 341, 33

Kennicutt Robert C. J., 1992, The Astrophysical Journal Supplement Series, 79, 255

Kepner J., Fan X., Bahcall N., Gunn J., Lupton R., Xu G., 1999, The Astrophysical Journal, 517, 78

Kingma D. P., Welling M., 2013, arXiv e-prints, p. arXiv:1312.6114

Kiran B. R., Sobh I., Talpaert V., Mannion P., Sallab A. A. A., Yogamani S., Pérez P., 2022, IEEE Transactions on Intelligent Transportation Systems, 23, 4909

Kleene S. C., 1956, Annals of Mathematics Studies, 34, 3

Kodama T., Arimoto N., 1997, Astronomy & Astrophysics, 320, 41

Kodama T., Arimoto N., Barger A. J., Arag'on-Salamanca A., 1998, Astronomy & Astrophysics, 334, 99

Kodama T., Smail I., Nakata F., Okamura S., Bower R. G., 2001, The Astrophysical Journal, 562, L9

Koester B. P., et al., 2007, The Astrophysical Journal, 660, 221

Kohavi R., 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 1137–1143

Kravtsov A. V., Borgani S., 2012, Annual Review of Astronomy and Astrophysics, 50, 353

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Pereira F., Burges C., Bottou L., Weinberger K., eds, Vol. 25, Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kron R. G., 1980, The Astrophysical Journal Supplement Series, 43, 305

Kurtz M. J., Huchra J. P., Beers T. C., Geller M. J., Gioia I. M., Maccacaro T., Schild R. E., Stauffer J. R., 1985, The Astronomical Journal, 90, 1665

L. McHugh M., 2008, Biochemia Medica, 18, 7

Laigle C., et al., 2016, The Astrophysical Journal Supplement Series, 224, 24

Lan T.-W., Ménard B., Mo H., 2016, Monthly Notices of the Royal Astronomical Society, 459, 3998

Larson R. B., Tinsley B. M., Caldwell C. N., 1980, The Astrophysical Journal, 237, 692

Lassell W., Marth A., 1867, Memoirs of the Royal Astronomical Society, 36, 45

Lauberts A., Valentijn E. A., 1989, The surface photometry catalogue of the ESO-Uppsala galaxies

Lawrence A., et al., 2007, Monthly Notices of the Royal Astronomical Society, 379, 1599

Le Fèvre O., et al., 2005, Astronomy & Astrophysics, 439, 845

LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, Proceedings of the IEEE, 86, 2278

LeCun Y., Cortes C., Burges C., 2010, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2

LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436

Liang H., Sun X., Yunlei S., Gao Y., 2017, EURASIP Journal on Wireless Communications and Networking, 2017

Lidman C., et al., 2008, Astronomy & Astrophysics, 489, 981

Lilly S. J., et al., 2007, The Astrophysical Journal Supplement Series, 172, 70

Lin Y.-T., Mohr J. J., 2004, The Astrophysical Journal, 617, 879

Lin T.-Y., et al., 2014, arXiv e-prints, p. 1405.0312

Lin S.-C., Su Y., Liang G., Zhang Y., Jacobs N., Zhang Y., 2022, Monthly Notices of the Royal Astronomical Society, 512, 3885

Liske J., Lemon D. J., Driver S. P., Cross N. J. G., Couch W. J., 2003, Monthly Notices of the Royal Astronomical Society, 344, 307

Liu F. T., Ting K. M., Zhou Z.-H., 2008, in 2008 Eighth IEEE International Conference on Data Mining. pp 413–422, doi:10.1109/ICDM.2008.17

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C., 2015, arXiv e-prints, p. arXiv:1512.02325

Liu Q., Zhang J., Liu J., Yang Z., 2022a, International Journal of Machine Learning and Cybernetics, 13, 1685

Liu A., et al., 2022b, Astronomy & Astrophysics, 661, A2

Lopes P. A. A., 2007, Monthly Notices of the Royal Astronomical Society, 380, 1608

Lucey J. R., 1983, Monthly Notices of the Royal Astronomical Society, 204, 33

Lupton R. H., Gunn J. E., Szalay A. S., 1999, The Astronomical Journal, 118, 1406

Macqueen J., 1967, in In 5-th Berkeley Symposium on Mathematical Statistics and Probability. pp 281–297

Maddox S. J., Sutherland W. J., Efstathiou G., Loveday J., 1990, Monthly Notices of the Royal Astronomical Society, 243, 692

Martinez M., Stiefelhagen R., 2018, arXiv e-prints, p. arXiv:1810.05075

McCarthy J., Minsky M. L., Rochester N., Shannon C. E., 2006, AI Magazine, 27, 12

McClintock T., et al., 2019, Monthly Notices of the Royal Astronomical Society, 482, 1352

McCulloch W. S., 1950, Dialectica, 4, 192

McGlynn T., Scollick K., White N., 1998, in McLean B. J., Golombek D. A., Hayes J. J. E., Payne H. E., eds, Vol. 179, New Horizons from Multi-Wavelength Sky Surveys. p. 465

Mcculloch W., Pitts W., 1943, Bulletin of Mathematical Biophysics, 5, 115

Mehrtens N., et al., 2012, Monthly Notices of the Royal Astronomical Society, 423, 1024

Mei S., et al., 2009, The Astrophysical Journal, 690, 42

Melin J. B., Bartlett J. G., Delabrouille J., 2005, Astronomy & Astrophysics, 429, 417

Menke T., HäSe F., Gustavsson S., Kerman A., Oliver W., Aspuru-Guzik A., 2018, in APS March Meeting Abstracts. p. S39.004

Merritt D., Graham A. W., Moore B., Diemand J., Terzić B., 2006, The Astronomical Journal, 132, 2685

Messier C., 1781, Catalogue des Nébuleuses et des Amas d'Étoiles (Catalog of Nebulae and Star Clusters), Connoissance des Temps ou des Mouvements Célestes, for 1784, p. 227-267

Meyer H., Kühnlein M., Appelhans T., Nauss T., 2016, Atmospheric Research, 169, 424

Minsky M. L., 1956, Some Universal Elements for Finite Automata. Princeton University Press, pp 117–128, doi:10.1515/9781400882618-005

Mishra A., Reddy P., Nigam R., 2019, arXiv e-prints, p. arXiv:1908.04682

Miyazaki S., et al., 2002, The Astrophysical Journal, 580, L97

Mohanapriya M., Jayabalan L., 2018, Journal of Physics: Conference Series, 1142, 012011

Moore G. E., 1965, Electronics, 38

Moore B., Quinn T., Governato F., Stadel J., Lake G., 1999, Monthly Notices of the Royal Astronomical Society, 310, 1147

Moretti A., et al., 2015, Astronomy & Astrophysics, 581, A11

Morgan W. W., 1958, Publications of the Astronomical Society of the Pacific, 70, 364

Morgan S. P., Teachman J. D., 1988, Journal of Marriage and Family, 50, 929

Müllner D., 2011, arXiv e-prints, p. arXiv:1109.2378

Muntoni A. P., Pagnani A., Weigt M., Zamponi F., 2021, arXiv e-prints, p. arXiv:2109.04105

Naim A., 1995, in American Astronomical Society Meeting Abstracts. p. 88.05

Nair V., Hinton G. E., 2010, in Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML'10. Omnipress, Madison, WI, USA, pp 807–814

Navarro J. F., Frenk C. S., White S. D. M., 1997, The Astrophysical Journal, 490, 493

Navarro J. F., et al., 2004, Monthly Notices of the Royal Astronomical Society, 349, 1039

Newman A. B., Ellis R. S., Andreon S., Treu T., Raichoor A., Trinchieri G., 2014, The Astrophysical Journal, 788, 51

Ng A. Y., 2004, in Proceedings of the Twenty-First International Conference on Machine Learning. ICML '04. Association for Computing Machinery, New York, NY, USA, p. 78, doi:10.1145/1015330.1015435, https://doi.org/10.1145/1015330.1015435

Notley S., Magdon-Ismail M., 2018, arXiv e-prints, p. arXiv:1805.02294

Ntampaka M., et al., 2019, The Astrophysical Journal, 876, 82

Nwankpa C., Ijomah W., Gachagan A., Marshall S., 2018, arXiv e-prints, p. arXiv:1811.03378

O' Mahony N., Campbell S., Carvalho A., Harapanahalli S., Velasco-Hernandez G., Krpalkova L., Riordan D., Walsh J., 2019, arXiv e-prints, p. arXiv:1910.13796

Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A., 1992, The Astronomical Journal, 103, 318

Oegerle W. R., Hoessel J. G., 1989, The Astronomical Journal, 98, 1523

Oegerle W. R., Hoessel J. G., Ernst R. M., 1986, The Astronomical Journal, 91, 697

Oegerle W. R., Hoessel J. G., Jewison M. S., 1987, The Astronomical Journal, 93, 519

Oke J. B., Gunn J. E., 1983, The Astrophysical Journal, 266, 713

Omohundro S. M., 1989, Technical report, Five balltree construction algorithms. Berkeley, California, USA

Owens E. A., Griffiths R. E., Ratnatunga K. U., 1996, Monthly Notices of the Royal Astronomical Society, 281, 153

Owers M. S., et al., 2017, Monthly Notices of the Royal Astronomical Society, 468, 1824

Paolillo M., Andreon S., Longo G., Puddu E., Gal R. R., Scaramella R., Djorgovski S. G., de Carvalho R., 2001, Astronomy & Astrophysics, 367, 59

Park C. H., Kim S. B., 2015, Expert Systems with Applications, 42, 2336

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, Astronomy & Astrophysics, 621, A26

Pearson K., 1900, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50, 157

Pearson K., 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Pérez R., Balatsky A., 2019, in APS March Meeting Abstracts. p. G70.288

Perez L., Wang J., 2017, arXiv e-prints, p. arXiv:1712.04621

Piattella O. F., 2018, arXiv e-prints, p. arXiv:1803.00070

Planck Collaboration et al., 2014, Astronomy & Astrophysics, 571, A29

Planck Collaboration et al., 2016a, Astronomy & Astrophysics, 594, A24

Planck Collaboration et al., 2016b, Astronomy & Astrophysics, 594, A27

Popesso P., Böhringer H., Brinkmann J., Voges W., York D. G., 2004, Astronomy & Astrophysics, 423, 449

Popesso P., Böhringer H., Romaniello M., Voges W., 2005, Astronomy & Astrophysics, 433, 415

Postman M., Lubin L. M., Gunn J. E., Oke J. B., Hoessel J. G., Schneider D. P., Christensen J. A., 1996, The Astronomical Journal, 111, 615

Pratt G. W., Arnaud M., Biviano A., Eckert D., Ettori S., Nagai D., Okabe N., Reiprich T. H., 2019, Space Science Reviews, 215, 25

Press W. H., Schechter P., 1974, The Astrophysical Journal, 187, 425

Raschka S., 2014, About feature scaling and normalization - and the effect of standardization for machine learning algorithms, https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Raschka S., Mirjalili V., 2017, Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, 2nd Edition, 2nd edn. Packt Publishing

Rashevsky N., 1935, Nature, 135, 528

Raudys S., Jain A., 1991, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13, 252

Rauzy S., Adami C., Mazure A., 1998, Astronomy & Astrophysics, 337, 31

Reblinsky K., Bartelmann M., 1999, Astronomy & Astrophysics, 345, 1

Reichardt C. L., et al., 2013, The Astrophysical Journal, 763, 127

Ren S., He K., Girshick R., Sun J., 2015, arXiv e-prints, p. arXiv:1506.01497

Ren Y., Zhu C., Xiao S., 2018, Mathematical Problems in Engineering, 2018

Ricci, M. et al., 2020, Astronomy & Astrophysics, 642, A126

Rish I., 2001, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 3

Rodriguez F., Merchán M., 2020, Astronomy & Astrophysics, 636, A61

Rood H. J., Sastry G. N., 1971, Publications of the Astronomical Society of the Pacific, 83, 313

Rosenblatt F., 1958, Psychological review, 65, 386

Rozo E., et al., 2009, The Astrophysical Journal, 708, 645

Ruder S., 2016, arXiv e-prints, p. arXiv:1609.04747

Rumelhart D., Hinton G., Williams R., 1985, Learning Internal Representations by Error Propagation. Institute for Cognitive Science, University of California, San Diego

Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533

Rykoff E. S., et al., 2014, The Astrophysical Journal, 785, 104

Salvato M., Ilbert O., Hoyle B., 2019, Nature Astronomy, 3, 212

Samuel A. L., 1959, IBM Journal of Research and Development, 3, 210

Sánchez C., et al., 2014, Monthly Notices of the Royal Astronomical Society, 445, 1482

Sarazin C. L., 1986, Reviews of Modern Physics, 58, 1

Saunders W., et al., 2000, Monthly Notices of the Royal Astronomical Society, 317, 55

Scaramella R., et al., 2021, arXiv e-prints, p. arXiv:2108.01201

Schechter P., 1976, The Astrophysical Journal, 203, 297

Schlafly E. F., Finkbeiner D. P., 2011, The Astrophysical Journal, 737, 103

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, The Astrophysical Journal, 500, 525

Schölkopf B., Smola A. J., Müller K.-R., 1997, in Proceedings of the 7th International Conference on Artificial Neural Networks. ICANN '97. Springer-Verlag, Berlin, Heidelberg, pp 583–588

Schultz H., 1875, Monthly Notices of the Royal Astronomical Society, 35, 135

Searle G. M., 1880, Project PHAEDRA: Preserving Harvard's Early Data and Research in Astronomy (https://library.cfa.harvard.edu/project-phaedra). Harvard College Observatory observations, p. 324

Sebok W. L., 1979, The Astronomical Journal, 84, 1526

Sensui T., Funato Y., Makino J., 1999, Publications of the Astronomical Society of Japan, 51, 943

Serra-Ricart M., 1994, Faint Object Classification Using Artificial Neural Networks. Vol. 161, Springer Netherlands

Shane C. D., Wirtanen C. A., 1954, The Astronomical Journal, 59, 285

Shectman S. A., Landy S. D., Oemler A., Tucker D. L., Lin H., Kirshner R. P., Schechter P. L., 1996, The Astrophysical Journal, 470, 172

Shen H., George D., Huerta E. A., Zhao Z., 2017, arXiv e-prints, p. arXiv:1711.09919

Sheth R. K., Tormen G., 1999, Monthly Notices of the Royal Astronomical Society, 308, 119

Simonyan K., Zisserman A., 2014, arXiv e-prints, p. arXiv:1409.1556

Smirnov N. V., 1939, Bull. Math. Univ. Moscou, 2, 3

Smith A. G., Hopkins A. M., Hunstead R. W., Pimbblet K. A., 2012, Monthly Notices of the Royal Astronomical Society, 422, 25

Snoek J., Larochelle H., Adams R. P., 2012, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. NIPS'12. Curran Associates Inc., Red Hook, NY, USA, pp 2951–2959

Stewart G. W., 1993, SIAM Review, 35, 551

Storrie-Lombardi M. C., Lahav O., Sodr L. J., Storrie-Lombardi L. J., 1992, Monthly Notices of the Royal Astronomical Society, 259, 8P

Stott J. P., Smail I., Edge A. C., Ebeling H., Smith G. P., Kneib J.-P., Pimbblet K. A., 2007, The Astrophysical Journal, 661, 95

Stott J. P., Edge A. C., Smith G. P., Swinbank A. M., Ebeling H., 2008, Monthly Notices of the Royal Astronomical Society, 384, 1502

Stott J. P., Pimbblet K. A., Edge A. C., Smith G. P., Wardlow J. L., 2009, Monthly Notices of the Royal Astronomical Society, 394, 2098

Strauss M. A., et al., 2002, The Astronomical Journal, 124, 1810

Strazzullo V., et al., 2016, The Astrophysical Journal Letters, 833, L20

Strazzullo V., et al., 2019, Astronomy & Astrophysics, 622, A117

Strobl C., Malley J., Tutz G., 2009, Psychological Methods, 14, 323

Sunayama T., et al., 2020, Monthly Notices of the Royal Astronomical Society, 496, 4468

Sunyaev R. A., Zeldovich Y. B., 1972, Comments on Astrophysics and Space Physics, 4, 173

Sutton R. S., 1988, Machine Learning, 3, 9

Sutton R. S., McAllester D., Singh S., Mansour Y., 1999, Policy Gradient Methods for Reinforcement Learning with Function Approximation. Vol. 12, MIT Press, https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf

Szabo T., Pierpaoli E., Dong F., Pipino A., Gunn J., 2011, The Astrophysical Journal, 736, 21

Szegedy C., Toshev A., Erhan D., 2013, in Burges C., Bottou L., Welling M., Ghahramani Z., Weinberger K., eds, Vol. 26, Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2013/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2015, arXiv e-prints, p. arXiv:1512.00567

Takahashi K., Sensui T., Funato Y., Makino J., 2002, Publications of the Astronomical Society of Japan, 54, 5

Takizawa M., Mineshige S., 1998, The Astrophysical Journal, 499, 82

Tanaka M., et al., 2018, Publications of the Astronomical Society of Japan, 70, S9

Tempel E., Kipper R., Tamm A., Gramann M., Einasto M., Sepp T., Tuvikene T., 2016, Astronomy & Astrophysics, 588, A14

Tharwat A., 2021, Applied Computing and Informatics, 17, 168

Thompson N. C., Greenewald K., Lee K., Manso G. F., 2020, arXiv e-prints, p. arXiv:2007.05558

Tibshirani R., 1996, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267

Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, The Astrophysical Journal, 688, 709

Tkatchenko A., 2020, Nature Communications, 11, 4125

Torrey L., Shavlik J., 2009. , doi:10.4018/978-1-60566-766-9.ch011

Turner E. L., Gott J. R. I., 1976, The Astrophysical Journal Supplement Series, 32, 409

Uttley A. M., 1956, Automata studies, p. 253

Valdes F., 1982, in Instrumentation in Astronomy IV. pp 465–472, doi:10.1117/12.933489

Valentinuzzi T., et al., 2011, Astronomy & Astrophysics, 536, A34

Valotto C. A., Nicotra M. A., Muriel H., Lambas D. G., 1997, The Astrophysical Journal, 479, 90

VanderPlas J., 2016, Python Data Science Handbook: Essential Tools for Working with Data, 1st edn. O'Reilly Media, Inc., Sebastopol, California, USA

Varshney K. R., Willsky A. S., 2011, IEEE Transactions on Signal Processing, 59, 2496

Vikhlinin A., et al., 2009, The Astrophysical Journal, 692, 1060

Virtanen P., et al., 2020, Nature Methods, 17, 261

Voges W., et al., 1999, Astronomy & Astrophysics, 349, 389

Wadadekar Y., 2005, The Publications of the Astronomical Society of the Pacific, 117, 79

Walcher J., Groves B., Budavári T., Dale D., 2011, Astrophysics and Space Science, 331, 1

Watkins C., Dayan P., 1992, Machine Learning, 8, 279

Weinstein M. A., et al., 2004, The Astrophysical Journal Supplement Series, 155, 243

Weir N., Fayyad U. M., Djorgovski S., 1995, The Astronomical Journal, 109, 2401

Wen Z. L., Han J. L., 2013, Monthly Notices of the Royal Astronomical Society, 436, 275

Wen Z. L., Han J. L., 2015, The Astrophysical Journal, 807, 178

Wen Z. L., Han J. L., Liu F. S., 2009, The Astrophysical Journal Supplement Series, 183, 197

Wen Z. L., Han J. L., Liu F. S., 2012, The Astrophysical Journal Supplement Series, 199, 34

Wilson G., Smail I., Ellis R. S., Couch W. J., 1997, Monthly Notices of the Royal Astronomical Society, 284, 915

Wittman D., Dell'Antonio I. P., Hughes J. P., Margoniner V. E., Tyson J. A., Cohen J. G., Norman D., 2006, The Astrophysical Journal, 643, 128

Wold H., 1983, Iiasa collaborative paper, Systems Analysis by Partial Least Squares, http://pure.iiasa.ac.at/id/eprint/2336/. IIASA, Laxenburg, Austria, http://pure.iiasa.ac.at/id/eprint/2336/

Wright E. L., et al., 2010, The Astronomical Journal, 140, 1868

Wylezalek D., et al., 2014, The Astrophysical Journal, 786, 17

Xu Q., Liang Y.-Z., 2001, Chemometrics and Intelligent Laboratory Systems, 56, 1

Yagi M., Kashikawa N., Sekiguchi M., Doi M., Yasuda N., Shimasaku K., Okamura S., 2002, The Astronomical Journal, 123, 87

Yan Z., Mead A. J., Van Waerbeke L., Hinshaw G., McCarthy I. G., 2020, Monthly Notices of the Royal Astronomical Society, 499, 3445

Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005, Monthly Notices of the Royal Astronomical Society, 356, 1293

Yee H. K. C., Gladders M. D., López-Cruz O., 1999, in Weymann R., Storrie-Lombardi L., Sawicki M., Brunner R., eds, Astronomical Society of the Pacific Conference Series Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies. p. 166 (arXiv:astro-ph/9908001)

Yee H. K. C., et al., 2000, The Astrophysical Journal Supplement Series, 129, 475

Yennapureddy M. K., Melia F., 2019, European Physical Journal C, 79, 571

York D. G., et al., 2000, The Astronomical Journal, 120, 1579

Zhao Y., Nasrullah Z., Li Z., 2019, Journal of Machine Learning Research, 20, 1

Zwicky F., 1937, The Astrophysical Journal, 86, 217

Zwicky F., 1957, Morphological Astronomy. Springer Verlag

Zwicky F., Herzog E., Wild P., Karpowicz M., Kowal C. T., 1961, Catalogue of galaxies and of clusters of galaxies, Vol. I

d'Abrusco R., Longo G., Paolillo M., de Filippis E., Brescia M., Staiano A., Tagliaferri R., 2007, arXiv e-prints, pp astro–ph/0701137

de Andres D., et al., 2022, in European Physical Journal Web of Conferences. p. 00013 (arXiv:2111.01933), doi:10.1051/epjconf/202225700013

de Filippis E., Paolillo M., Longo G., La Barbera F., de Carvalho R. R., Gal R., 2011, Monthly Notices of the Royal Astronomical Society, 414, 2771

de Propris R., et al., 2003, Monthly Notices of the Royal Astronomical Society, 342, 725

de Vaucouleurs G., 1959, Handbuch der Physik, 53, 275

van Haarlem M. P., 1996, in Coles P., Martinez V., Pons-Borderia M.-J., eds, Astronomical Society of the Pacific Conference Series Vol. 94, Mapping, Measuring, and Modelling the Universe. p. 191 (arXiv:astro-ph/9601081)

van Haarlem M. P., Frenk C. S., White S. D. M., 1997, Monthly Notices of the Royal Astronomical Society, 287, 817

van Uitert E., Gilbank D. G., Hoekstra H., Semboloni E., Gladders M. D., Yee H. K. C., 2016, Astronomy & Astrophysics, 586, A43

van den Bergh S., 1960, The Astrophysical Journal, 131, 215

van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579