Applying Machine Learning to Find Galaxy Redshifts

by **Alexander Ogden**
supervised by **Professor Carlton Baugh** &
Professor Peder Norberg

April 2022

Project report submitted for the degree
Master of Physics and Astronomy (MPhys)

at the

Institute for Computational Cosmology
Department of Physics

Abstract

Astronomy has entered the era of big data and the next generation of surveys such as LSST and SKA are expected to measure the positions and photometry of millions of galaxies. Understanding the redshift distribution of these galaxies is crucial for advances in future cosmology as it would provide unprecedented knowledge about the 3D structure of the universe and constrain key parameters of cosmological models. Spectroscopy is too time intensive to measure redshifts down to a faint magnitude limit for such large surveys, so the development of machine learning algorithms to calculate redshifts photometrically is essential. A supervised random forest algorithm is implemented to estimate photometric redshifts for galaxies in the COSMOS field using a training set of approximately 12,000 galaxies with confident spectroscopic redshifts from the zCOSMOS survey. Monte Carlo realisations are used to propagate the photometric uncertainties into the predicted redshifts and performance metrics. The photometric redshifts have a centralised scatter of $\sigma_{68} = 0.0147 \pm 0.0007$ and an outlier fraction of $\eta = 1.81\% \pm 0.04\%$, which are competitive with algorithms from literature. It is found that the inclusion of intermediate width filters in the redshifted Balmer break region greatly improved the performance of the forest, compared to when only broad filters are used.

Contents

1 Introduction

1.1 Motivation for Redshift Surveys

Redshift measurements are the cornerstones of modern cosmology. Through Hubble's law, they enable the calculation of galaxy distances which provide a crucial third dimension to astronomical observations. Therefore, by surveying the redshifts of large galaxy populations, the universe's large-scale 3D structure can be mapped and analysed (Jones et al., 2009), which has led to several of the most important scientific breakthroughs of this century such as the detection of baryon acoustic oscillations (Eisenstein et al., 2005) and the discovery of the universe's expansion (Hubble, 1929). These maps have important implications about the distribution of mass in the universe on the largest scales (Davies et al., 2015) and the evolution of galaxy clusters (Mohammad et al., 2018) which in turn are essential to the empirical testing of cosmological models since they constrain many of their key parameters such as the mass density (Peacock et al., 2001), luminosity density (Lilly et al., 1996) and the Hubble constant (Beutler et al., 2011; Wang et al., 2017).

As telescope technology and computational processing power improve, the scope of large surveys increases in both coverage and depth. Consequently, the next generation of telescopes have been assigned a range of ground-breaking, all-sky and wide-field missions (e.g. Euclid, EMU and LSST). It is predicted that understanding the redshift distributions in these novel surveys will be crucial for the coming decades' key cosmological lines of inquiry, such as explaining cosmic acceleration and constraining the dark energy equation of state (Camera et al., 2012; Collaboration et al., 2021; Vargas-Magana et al., 2019). Astronomers are therefore highly motivated to measure reliable redshifts for as many galaxies as possible. An ideal survey would precisely measure accurate redshifts for every galaxy in the sky down to an incredibly faint magnitude limit, however there are technical limitations that prevent this. As a result, two different methods for measuring redshifts have emerged that offer a trade-off between redshift quality and magnitude depth, these are spectroscopy and photometry respectively.

1.2 Spectroscopic Redshifts

Spectroscopy produces the most reliable redshift measurements (referred to as spectroscopic redshifts or z_{spec}). In this method, the spectral energy distribution (SED) of a galaxy is sampled at high resolution with a spectrograph and a distinctive emission or absorption line is identified at its redshifted wavelength λ_{obs}. By comparing this to the known rest-frame wavelength λ_{emit}, the redshift can be calculated by definition using eq. (1). Spectroscopy is extremely reliable since the process can be repeated with multiple spectral lines and because spectrographs can be manufactured to high precision.

$$z_{\text{spec}} = \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}} \tag{1}$$

In the 1980s, spectroscopy was unsuitable for surveys that aimed to measure redshifts

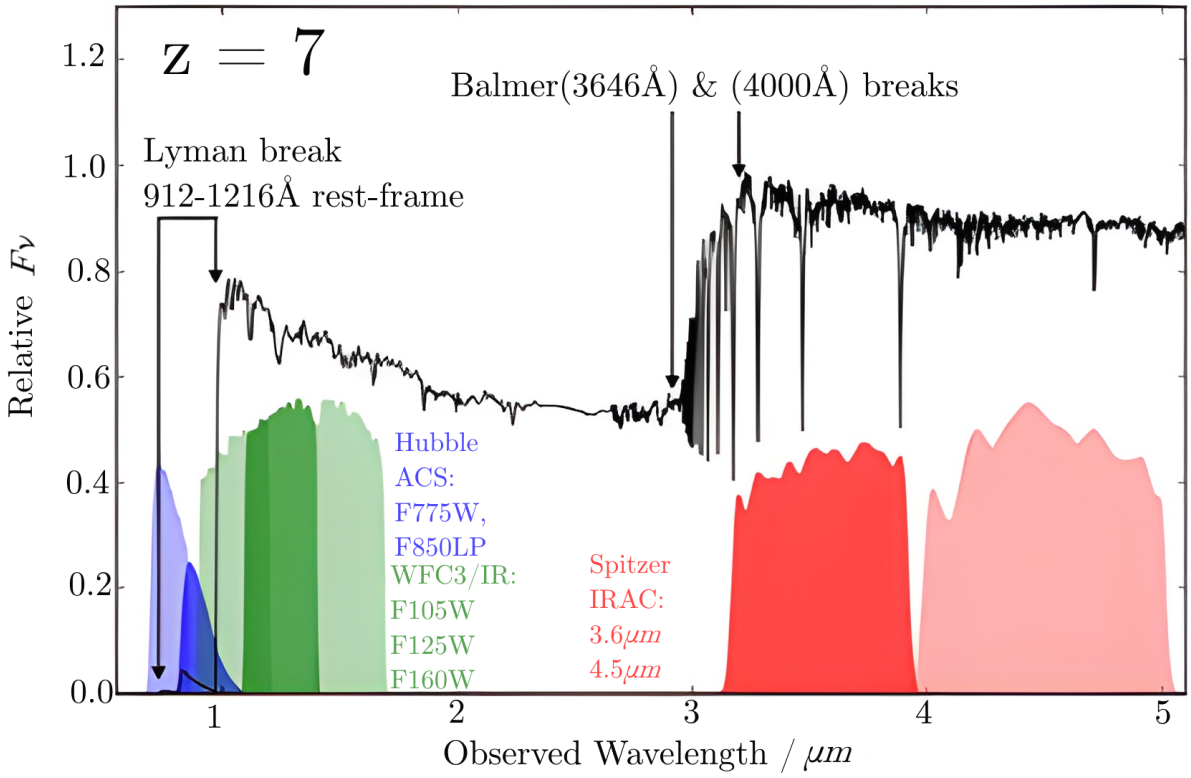

Figure 1: An illustration of a typical observed SED from a young galaxy at $z = 7$ compared to photometric filters from Hubble's ACS and WFC3 and Spitzer's IRAC instruments. The Lyman and Balmer breaks are labelled at their position in the observed SED along with their rest-frame wavelengths. Emission and absorption lines can be seen as peaks and troughs respectively in the SED. Reproduced from Dunlop (2013).

for a large field of galaxies as they relied on single-object spectroscopy. In this regime, sampling the full SED at high resolution requires long exposure times and there is always intense competition for observing time on the best spectrographs, so any large spectroscopic survey would likely not be completed in a reasonable time frame. Consequently, redshift surveys at this time were limited to only a few thousand galaxies (Davis & Peebles, 1983; Bean et al., 1983). Today, this problem is solved via the use of multiplexed fibre-optic spectrographs which allow for multiple galaxies in the field to have their SEDs measured simultaneously by targeting one optical fibre on each galaxy. DESI is one such currently operating spectroscopic instrument able to sample 5000 simultaneous SEDs which plans to measure 30 million redshifts over 14,000 deg^2 of sky across its many redshift surveys (Collaboration et al., 2016a).

However, these spectroscopic surveys are imperfect since they cannot sample every galaxy in a given field down to a faint magnitude limit. In order to do this, the fibre-optic spectrographs would have to sample each field multiple times which would cause a drastic increase in mission time. This is made worse by the fact that faint galaxies require longer exposure times to achieve the signal to noise ratio required for a high resolution SED. As a result, wide-field spectroscopic surveys are biased towards bright, nearby (low redshift) and easy to measure galaxies with well defined spectral features. For example, one third

of all galaxies measured by DESI will be in the Bright Galaxy Survey (BGS) with an estimated median redshift of $z \approx 0.2$ (Collaboration et al., 2016b).

1.3 Photometric Redshifts

To overcome this problem, photometry must be used instead of spectroscopy. In this method, the galaxy's magnitude or colour is measured in a range of different photometric filters/bands. Photometric bands are typically too broad to have the resolution required to detect individual spectral lines as spectrographs do, however they can instead detect much larger spectral features such as the Lyman break, Lyman-alpha forest and the Balmer break. In this way, the SED is sampled at low resolution and its general shape can in theory be recovered and used to calculate a redshift. In fig. 1 an example SED is shown, it is clear that the filters are too broad to locate spectral lines but filters chosen sensibly close to large spectral features could reproduce a low resolution SED.

The redshifts calculated in this way are known as photometric redshifts or z_{phot}. Photometry is much faster than spectroscopy and can be performed on any galaxy regardless of its depth, although due to the reduced resolution, the accuracy and precision of the redshift measurement is greatly reduced. The definition shown by eq. (1) should not be used in this regime since the uncertainties in λ_{obs} are now too large and, as a result, it is non-trivial to recover the photometric redshift. There are two leading methods for calculating photometric redshift: SED template fitting and machine learning (ML).

1.3.1 SED Template Fitting

First proposed by Baum (1962) to calculate redshifts for galaxies too faint for spectroscopy, the SED template fitting method involves constructing an estimate of the SED from the photometric measurements and then comparing it to ideal theoretical SED templates. In the beginning, this method was limited to only comparing galaxies of the same type and required manual comparison of SEDs. In recent decades though, it has been developed into a robust technique (Ilbert et al., 2009; Eriksen et al., 2019) with many publicly available computer programs such as EAZY (Brammer et al., 2008), LePHARE (Arnouts & Ilbert, 2011) and hyperz (Bolzonella et al., 2000).

There are some downsides to this method however. The quality of redshifts produced depends entirely on the quality of the templates, which in turn depend on the quality of the real galaxy SEDs they are based off. If the real SEDs have a high signal to noise ratio with well defined, clear spectral features then they are ideal for constraining templates. However, these ideal SEDs typically belong to bright galaxies which have low redshifts, so the templates should only be used for such galaxies. Although photometry can measure galaxies to faint depths, it can be challenging to create a template which is applicable for them case since their SED quality is so poor.

Moreover, real galaxy SEDs can be affected by a variety of additional factors (Salvato et al., 2019) e.g. the behaviour of the filter transmission curves (Beck et al., 2017); the presence of active galactic nuclei (AGNs) (Salvato et al., 2009) (Salvato et al., 2011); the

star formation history of the galaxy (Tanaka, 2015); the prevalence nebular emission lines (Yuan et al., 2019); and the amount of dust attenuation from the interstellar medium and Milky Way (Massarotti et al., 2001). The inclusion of these factors in templates presents its own challenges since they are complex and often require approximations. Each of these factors must be studied individually and incorporated into a unique template, resulting in templates that are only applicable to a specific type of object (e.g. galaxies with AGNs), making comparisons between them challenging. Selecting and justifying the factors accounted for by a given template is an important challenge for this technique. If the templates are biased with respect to the galaxy population being studied, they will fail to accurately reproduce the observed SEDs and cause the photometric redshifts to be estimated incorrectly.

1.3.2 Machine Learning

On the other hand, the machine learning (ML) technique builds an algorithm to estimate redshift which automatically optimises its own performance. This period of optimisation is known as learning *training* the algorithm, and once complete the machine learner is ready to be tested or *verified*. The way in which ML algorithms train can be used to classify them, there are two main groups: supervised and unsupervised learning. During unsupervised learning, the machine estimates redshift by searching for patterns in only the photometry it receives from real galaxies. However, this report will focus on supervised learning. Here, the machine learns the photometry-redshift relationship for galaxies with previously measured photometry and reliable spectroscopic redshifts (known as the training set). By interpolating in the training photometry-redshift space, the machine can estimate photometric redshifts for new objects not included in the training set. The quality of the photometric redshifts is heavily dependent on the quality of the training set which is in turn dependent on many factors. Firstly, the training set must be representative of the galaxies in the verification set it is designed to be used on: if the training spectroscopy contains bias then the estimated redshifts will also. Secondly, the training set must contain enough galaxies for the algorithm to effectively learn and approximate the photometry-redshift function. Thirdly, each galaxy must have high quality photometry so that as many large scale SED features are captured as possible (Norris et al., 2019). This can be achieved by having photometric filters covering a sensible region of the electromagnetic spectrum (e.g. either side of the Balmer or Lyman breaks); having a continuous set of filters which span the spectrum; and by having a high filter density (i.e. many narrow filters), since this begins to resemble a spectrograph. ML algorithms have been tested on both broad (Hoyle et al., 2015; Pasquet et al., 2019) and narrow (Eriksen et al., 2020) surveys. Note that, like spectroscopy, performing a narrow photometric survey for enough galaxies to constitute a sufficiently large training set can be prohibitively time consuming, so the balance between training set size and filter density must be chosen carefully. If the training set is of sufficient size and quality, then the ML algorithm can estimate photometric redshifts to a very high accuracy and precision.

There is great variety amongst supervised ML algorithms, two of the most popular

choices for photometric redshift computation are random forests and neural networks (NNs). Random forests are simple, easy to implement algorithms that estimate redshifts using a *forest* of decision trees (see § 2.1). Despite their simplicity, they remain a relevant and effective algorithm (Mucesh et al., 2021) and can be extended to include measurement errors (Carrasco Kind & Brunner, 2013). In some cases they can even outperform more complicated NNs (Rau et al., 2015). Neural networks process data using layers of nodes or 'neurons' to apply matrix transformations to the input data until the final neuron contains only the output data. A deep neural network (DNN) is a type of NN that has a large number of layers and neurons per layer. Although this increases the training time and required training set size, this generally leads to a more capable algorithm. Many types of DNN have been successful at estimating photometric redshifts, in particular convolutional networks (CNNs) (Syarifudin et al., 2019; Stivaktakis et al., 2020; Hoyle, 2016; Pasquet et al., 2019) and mixture density networks (MDNs) (D'Isanto & Polsterer, 2018; Eriksen et al., 2020).

ML algorithms can offer some benefits over SED template fitting, however there is no consensus on the preferred technique, they are both prevalent and are suitable for different scenarios. While ML algorithms are dependent on the quality of the spectroscopy in their training set, they are insensitive to biases in the photometry, so they outperform SEDs in fields with poor photometry and SEDs outperform ML in fields with rich photometry (Norris et al., 2019; Salvato et al., 2019). Although training a machine learner can be slow, trained algorithms can calculate photometric redshifts much faster than template fitting codes (Vanzella et al., 2004). Machine learning overcomes some problems of SED fitting by allowing the machine to automatically find the link between photometry and redshift, rather than requiring a template. However, this means ML algorithms produce redshifts without physical justification and it can be hard for humans to understand why they give the answers they do. For this reason they are often called 'black boxes'.

1.4 Summary

Due to limited spectroscopy at high redshift, ML algorithms have so far mainly been used in low redshift ($z_{\text{spec}} < 0.4$), wide-field surveys using broad band photometry (Pasquet et al., 2019) and their applications in deeper fields has been limited in comparison. Recently, mixture density networks using the PAU narrow band survey on the COSMOS field have demonstrated their effectiveness for intermediate redshift ($z_{\text{spec}} < 1.4$) galaxies (Eriksen et al., 2020). To what extent can these results be achieved by a computationally simpler random forest algorithm using broad and intermediate band photometry instead?

In this report, a supervised random forest ML algorithm is presented and used to calculate photometric redshifts for galaxies in the COSMOS field, using spectroscopic redshifts measured by the zCOSMOS survey as a training set. In § 2, the machine learning algorithm and data processing techniques are outlined. Then, in § 3 the photometric redshifts are presented, evaluated and compared to results from literature. Finally, this work is summarised and future avenues for research are discussed in § 4.

2 Method

In § 2.1 the theory behind regression trees and random forests is presented before the metrics used to evaluate the forest's performance are given. Then, the hyperparameter values are explained and optimised using the performance metrics. In § 2.2 the choice of dataset is explained and the data analysis and reduction undertaken when constructing the training set is described.

2.1 Machine Learning

2.1.1 Regression Trees

The Random Forest (RF) is a form of supervised machine learning algorithm composed of many individual decision trees grouped together into a 'forest'. A decision tree is a data structure which consists of a binary tree that recursively splits a dataset based on its features until each subset of the data can be assigned an output value. Decision trees can either be classification or regression trees depending on whether their output values are discrete or continuous respectively. In this case, the features of each galaxy are its photometric colours and the galaxies are assigned continuous photometric redshifts as output values by regression trees. The tree contains decision nodes which apply a binary condition to split the *parent* data into two *child* subsets based on the galaxy colours. Each subset is further split by additional decision nodes until it reaches a node at the end of the tree known as a *leaf node*. There are photometric redshift values associated with each leaf node, which are assigned to galaxies placed there by the decision nodes.

If the binary conditions are chosen at random, the decision tree is unlikely to sort the data effectively. Decision tree learning is a supervised ML algorithm that optimises the performance of a decision tree by finding the most effective binary conditions for its training set. The *purity* of a node is any measure of its subset's z_{spec} scatter. An effective condition is one that splits the data such that the purity of the child nodes is greater than that of the parent node. Purity can be measured quantitatively using a *splitting criterion*, the most appropriate of which, in this case, is the variance or mean squared error (MSE):

$$\text{MSE} = \frac{\sum_i (\bar{z}_{\text{spec}} - z^i_{\text{spec}})^2}{n} \tag{2}$$

Where $\bar{z}_{\text{spec}}$ is the mean of the z_{spec} subset at that node, which contains n galaxies with redshifts z^i_{spec}. If the parent node contains galaxies across a wide redshift distribution, the MSE will be large. If a binary condition effectively splits the parent data into distinct child subsets, each with a tight redshift distribution, the MSE total across both child nodes will be less than the parent MSE. Therefore, the optimum splitting condition at each parent node can be chosen by minimising the MSE sum across its child nodes. Once every condition across the tree has been optimised in this way, it is said that the tree is trained (note that this is a greedy optimisation i.e. the best condition at each individual parent node is selected with no regard for how it might negatively impact the conditions of other nodes further along the tree). Once the tree has been trained, it can be used to

estimate photometric redshift for unseen objects. The unseen objects are sorted into leaf nodes by the optimised decision nodes. At the end of training, each leaf node contains a small galaxy sample with a tight redshift distribution. The mean of this distribution ($\bar{z}_{\rm spec}$) is then the photometric redshift estimated for the unseen objects at that leaf node.

Decision tree learning is effective and fast. The tree itself will sort data efficiently and, unlike many ML algorithms, it is a 'white box' model i.e. humans can analyse the decision nodes and interpret the reasoning behind its classifications. However the method has crucial flaws. Large decision trees have a tendency to overfit, meaning they categorise objects in the training set with low bias but have a high scatter when new objects are introduced (see § 2.1.3 for discussion about bias and scatter). This is often because the tree has learnt to recognise random noise in the training set which wont be present for unseen objects.

2.1.2 Random Forest

By grouping many decision trees together in a random forest, these problems can be overcome with a few adjustments. The first step is to *bootstrap* the data. This is where the original dataset is randomly sampled (with replacement) to produce a new *bootstrapped* dataset of the same size, the only difference is that the new dataset may contain some duplicates (note that, the random sampling can be reproduced for repeat experiments by using a known seed). Each tree in the forest is given a different bootstrapped dataset to train with, so while each tree overfits to the bootstrapped data, no trees can overfit the full training data because they do not see it in its entirety. The predictions of a single tree can therefore be poor and sensitive to noise, but the mean of the whole forest's predictions is accurate and precise (this is known as *aggregation*). Put simply, the trees that underestimate the redshift are balanced out by trees that overestimate it. The performance of the algorithm is improved hugely since the variance is reduced with only marginal increases in the bias. This is similar to the 'wisdom of the crowd' principal. Combining bootstrapping and aggregating in this way is known as *bagging*.

Feature bagging is a similar process where each tree only sees a small selection of the features of the data it is given e.g. for a galaxy with N photometric colours, each tree would only see $\sqrt{N}$ or $\log_2(N)$ of them. This can further decrease the variance of the forest following the same reasoning as above. The random forest was built using the scikit-learn python package (see § 5).

2.1.3 Metrics

A variety of metrics are available for measuring the performance of ML algorithms. Here, those commonly used in photometric redshift estimation are presented (Salvato et al., 2019). It is useful to define the separation between the true (spectroscopic) and predicted (photometric) redshift as the 'fractional difference' or Z:

$$Z = \frac{z_{\rm phot} - z_{\rm spec}}{1 + z_{\rm spec}} \tag{3}$$

At redshift z the widths of photometric filters widen by a factor $1 + z_{\text{spec}}$. Without correcting for this by dividing by $1 + z_{\text{spec}}$ in the definition of Z, comparisons between measurements at different z_{spec} would be meaningless. The bias metric is the mean fractional difference:

$$\text{bias} = E[Z] = \left\langle \frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}} \right\rangle \tag{4}$$

If the bias is significantly above or below zero then a systematic error is causing the forest to disproportionately over or under estimate the photometric redshifts respectively. Mean is used in the definition of bias instead of the median so that it can be used as an indicator of whether the outliers in the photometric redshift distribution are causing these systematic errors.

The outlier fraction (η) is the fraction of sources with unusually incorrect photometric redshifts i.e. catastrophic failures. In particular, the fraction of galaxies whose absolute fractional difference is greater than 0.15:

$$\eta = \frac{\sum_i |Z_i|}{N} \qquad \forall \qquad |Z_i| > 0.15 \tag{5}$$

Where the total number of galaxies in the sample is N and where each galaxy has fractional difference Z_i. Due to the black-box nature of the forest, it is difficult to precisely determine why some galaxies are outliers. However, it could be due to poor photometry of the galaxy or a mistake by the forest e.g. confusing the Lyman and Balmer breaks. In this case an outlier threshold of 0.15 was chosen as this is typical for photometric redshifts derived from surveys that use broad bands (Bilicki et al., 2018) or both broad bands and intermediate bands like COSMOS (Salvato et al., 2009). Photometric redshifts from some narrow band surveys have extremely tight distributions, so a threshold can be chosen up to an order of magnitude smaller in this case (Eriksen et al., 2020).

To measure the scatter of photometric redshifts, the standard deviation cannot be used since it would be heavily impacted by the small fraction of extreme outliers. Instead, the centralised scatter (σ_{68}) is defined as half the difference between the 84^{th} and 16^{th} percentiles of the fractional difference:

$$\sigma_{68} = \frac{1}{2}(Z_{84.1} - Z_{15.9}) \tag{6}$$

These percentiles correspond to one standard deviation away from the mean of a Gaussian distribution and would contain 68% of the data. However, percentiles are affected less by outliers than the mean (used in the definition of standard deviation) so σ_{68} is an estimate of the standard deviation which is insensitive to outliers. It effectively calculates the standard deviation for the Gaussian-like central region of the fractional difference distribution and ignores the tails of the distribution which contain outliers. An ideal machine learner will predict z_{phot} values that are identical to the z_{spec} values. Hence the target values for all metrics is zero.

2.1.4 Hyperparameter Optimisation

A hyperparameter is a parameter that affects the structure and procedure of the algorithm training, as opposed to parameters which are produced during training e.g. the decision nodes' binary splitting criteria. Hyperparameter values must be chosen carefully for the machine learner to have maximum performance. The random forest's hyperparameters are:

• **Number of Trees:** Since each tree receives a bootstrapped sample, having too few trees results in poor performance as the whole training set is not seen collectively by the forest. The performance generally improves with an increasing number of trees. However, having too many trees will increase training time for only marginal performance improvements and may also lead to overfitting. In the left column of § 2.1.4, the number of trees is varied and its effect on the metrics is shown. The absolute bias seems to be relatively constant at ~ 0.015 with some large deviations at low numbers of trees (since a single tree performing poorly will skew the results more heavily when the total tree number is low). Both the outlier fraction and σ_{68} improve greatly when increasing tree number from 10 to 100. Once the forest grows to over 200 trees, σ_{68} begins to only improve marginally but the outlier fraction begins to marginally get worse, this is likely due to overfitting. As a result, the optimum number of trees was determined as 200.

• **Depth:** The number of decision node layers that a galaxy is sorted by before reaching a leaf node is known as depth. It is a measure of the size of each tree. Trees with low depth struggle to estimate redshift to high precision because they have an insufficient quantity of decision layers, meaning they see fewer colour measurements and sample less of the SED. As with trees, having too much depth can lead to overfitting and unneccesarily long training times. In the right column of § 2.1.4, the effects of varying depth on the metrics are shown. At low depth, the forest behaves similarly to an individual decision tree would: low bias with high outlier fraction and scatter. After a depth of 20, there is only marginal changes in all three metrics, so an optimum depth of 20 was chosen.

• **Number of Realisations:** The standard random forest algorithm was extended by using Monte Carlo realisations to include the effects of photometric measurement uncertainty into the photometric redshifts. Rather being given trained on the set of colours C, the forest was given the set of colours perturbed by their uncertainties $C + x\sigma_c$, where x is a random number drawn from a Gaussian distribution and σ_c are the colour uncertainties. Repeating this process n times (for different values of x) means that each galaxy has n separate photometric redshifts, the distribution of which depends directly on the uncertainties in the photometry. The photometric redshift of each galaxy is then the mean of these n values with uncertaintiy equal to their standard error (note that in § 3, some plots show the mean photometric redshifts and some display all the redshifts separately). Since each of the n forests produces a value for each metric, this process can be repeated to find the uncertainties for each metric. It was found that $n = 10$ was a suitable number of realisations to calculate the metrics to 4 decimal places.

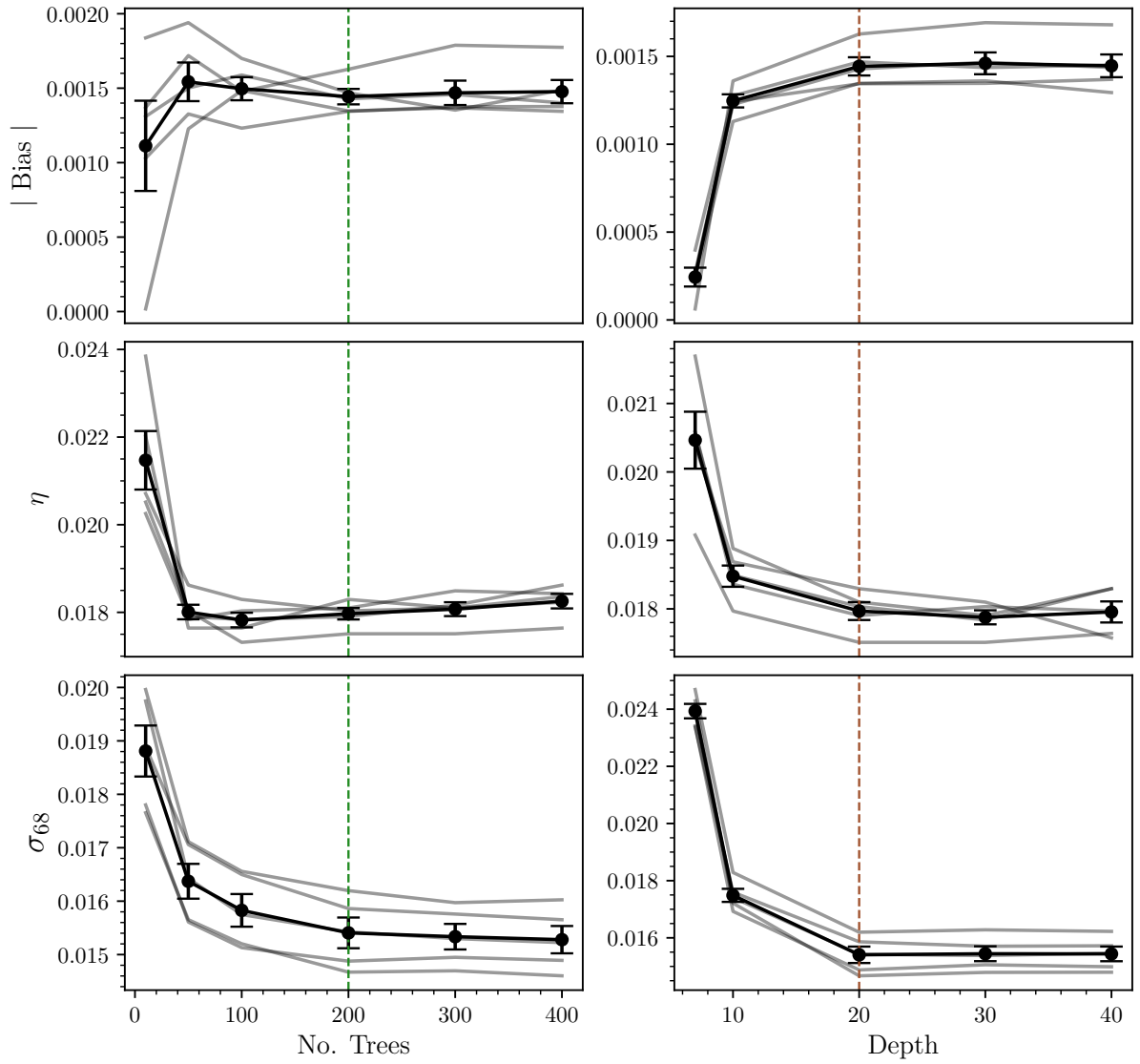

Figure 2: Relationship between metrics and hyperpatameters. In particular, the relationship between the absolute value of bias (top), outlier fraction (middle) and σ_{68} (bottom), and number of trees (left) and maximum tree depth (right). Five distinct forests with unique random seeds were trained for each column, for 6 different hyperparameter values, and plotted in grey. Note that, the points have been connected by lines only to distinguish which points belong to the same seed, forests were not trained across a continuous spectrum of hyperparameter values. The mean and standard error of these forests are plotted in black. The optimum hyperparameter values were chosen to be 200 trees with a depth of 20, shown in green and brown in their respective columns. When varying the number of trees, depth was kept at 20. When varying depth, the number of trees was kept at 200.

• **Splitting Criterion:** As mentioned previously, MSE was chosen as the splitting criterion, which measures the quality of a binary splitting condition. The alternative mean absolute error, defined as $\text{MAE} = \frac{\sum_i |\bar{z}_{\text{spec}} - z^i_{\text{spec}}|}{n}$, was not used as it vastly increased training time for no performance increase.

• **Feature Bagging:** By showing each tree only a fraction of the dataset features (in this case colours), the likelihood of overfitting is decreased. In this case it was found that the performance was best, and there was no overfitting, when each tree was shown every colour i.e. no feature bagging. This further suggests that the trees perform best when they have as much information about the SED as possible.

2.2 Data Processing

The training set consists of galaxies from the COSMOS field: a 2 deg^2 area of sky that has been observed by a diverse collection of astronomical instruments over the last 2 decades. The photometry of some 1.7 million galaxies measured by these observations was retrieved from the COSMOS 2020 catalogue (Weaver et al., 2022). The diverse range of surveys undertaken in the COSMOS field mean that the galaxies have photometry in a wide selection of both broad and intermediate width filters (see fig. 3), making it ideal for a training set. The photometry was combined with some 16,500 spectroscopic redshift measurements from the zCOSMOS bright 20k catalogue (Knobel et al., 2012) to construct the training set. The spectroscopic survey was conducted over a range of $15.00 < \text{I}_{\text{AB}}(814) < 22.50$ where $\text{I}_{\text{AB}}(814)$ is the ACS/HST filter F814W, resulting in a redshift range of $0 < z_{\text{spec}} < 4$. However, after removing poor quality spectra, the maximum redshift drops to $z_{\text{spec}} < 2.5$ with the vast majority of redshifts in the range $0 < z_{\text{spec}} < 1.2$. This means zCOSMOS is deeper than wide-field spectroscopic surveys like DESI's BGS but has a smaller sample size. In the following sections, the processing and analysis performed on the zCOSMOS and COSMOS datasets are outlined and the construction of the training set is described:

2.2.1 Confidence Cut

The quality of SEDs for objects in zCOSMOS varies and is assigned a confidence class by Lilly et al. (2009) to qualitatively describe this. By removing objects with low confidence, the training set will only be left with galaxies that have high quality SEDs with reliable redshifts, therefore improving the performance of the machine learner. Further information about the confidence cut can be found in § A. The confidence cut reduces the size of zCOSMOS from the original 20,690 galaxies to 17,538. The effects of the confidence cut on the zCOSMOS magnitude distribution can be seen in fig. 4. As expected, the confidence cut removes objects outside of the zCOSMOS magnitude limit since these likely have poor spectra. The slopes of the raw zCOSMOS data and the confident sample are approximately equal within the magnitude limits, implying that the confidence cut is independent of magnitude.

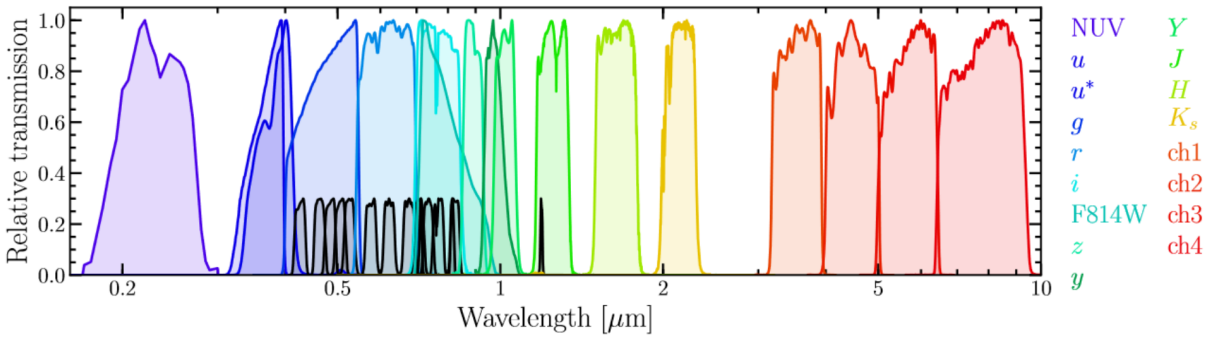

Figure 3: Relative transmission curves of photometric bands used for COSMOS 2020 photometry. The broad bands have been coloured and normalised to a peak transmission of 1.0, the intermediate and narrow bands are shown in black and normalised to 0.3 peak transmission. The UVISTA NB118 narrow band is the thin black peak inside the J broad band. The Subaru Suprime-Cam intermediate bands are the wider black peaks inside the g,r and i bands. Reproduced from (Weaver et al., 2022).

2.2.2 Redshift Cut

Distinguishing between stars and galaxies with certainty is difficult without detailed analysis. It is impractical to conduct such analysis on datasets as large as this, so as a result stars are often present. Stars have fundamentally different SEDs to galaxies, since galaxy SEDs contain contributions from many stars of different spectral classification as well as from the interstellar medium and nebulae. Removing stars from the dataset is important since the machine learner can struggle to approximate the photometry-redshift function for two distinct SED shapes simultaneously. Since stars are not moving with the Hubble flow, they should have redshifts close to zero. Applying a redshift cut to remove objects with $z < 0.002$ should therefore remove most stars from the dataset. This is verified explicitly in § B since the large peak of objects at $z = 0$ is removed after the confidence and redshift cuts. 16,649 objects remain after the redshift cut.

2.2.3 Magnitude Cut

The zCOSMOS bright catalogue is intended to have a magnitude range of $15.00 < I_{AB}(814) < 22.50$ where $I_{AB}(814)$ is the ACS/HST filter F814W. Objects outside these limits were removed, leaving 16,562 objects. The effects of the strict magnitude and redshift cuts on the zCOSMOS magnitude distribution can be seen in fig. 4. At faint magnitudes, the histogram slopes are equal, implying that neither cut has magnitude dependence in this region. However, at bright magnitudes the strict slope is steeper than those for the confident and raw datasets. This is due to the redshift cut removing stars from the dataset.

2.2.4 Matching

To generate a training dataset where every galaxy has both photometry and a known spectroscopic redshift, every zCOSMOS object needs to be matched to its photometry

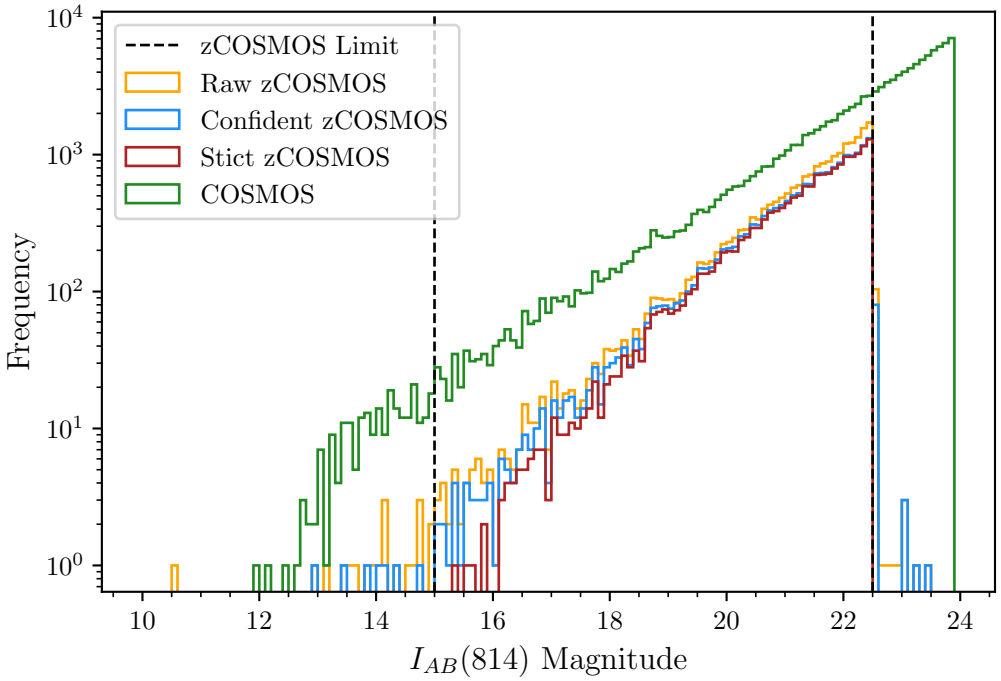

Figure 4: Logarithmic histogram showing the magnitude distribution of galaxies in the zCOSMOS catalogue during various stages of data processing. The unaltered zCOSMOS dataset is shown in orange, the blue shows zCOSMOS after the confidence cut and red shows zCOSMOS after a $z > 0.002$ redshift cut and strict magnitude cut of $15.00 < \mathrm{I_{AB}}(814) < 22.50$ which is the magnitude limit of the survey, shown as black dashed lines ($\mathrm{I_{AB}}(814)$ is the ACS/HST filter F814W). COSMOS is shown in green for comparison, note that these are only the COSMOS galaxies located within the zCOSMOS field (area of 1.64 deg^2) down to 24$^{\mathrm{th}}$ magnitude. Each bin is 0.1 mag wide.

in COSMOS. However between the two catalogues, there are no common ID numbers and the objects have marginally different positions. Therefore, each zCOSMOS object was matched to its closest object in COSMOS. A search radius of 0.5" was chosen and is justified in fig. 5. The vast majority of zCOSMOS objects are matched to a COSMOS object within 0.2", implying that they are indeed the same object. Also, the distribution tends to zero at high separation, if an unsuitably large radius is chosen then the histogram would have a second peak at some high separation as objects begin to get erroneously matched. 15,539 (93.8%) of the remaining 16,562 galaxies were matched successfully with this search radius.

2.2.5 Completing Dataset

Since the COSMOS photometry is sourced from a variety of surveys with different filters, areas and targets, it is inevitable that some galaxies in the training set are unmeasured in certain filters. If a galaxy has a high fraction of its photometry missing then the machine learner is unable to sufficiently sample its SED and is likely to estimate a photometric redshift in disagreement with its spectroscopic redshift measurement. Additionally, if such galaxies are in the training set then the machine learner may recognise patterns

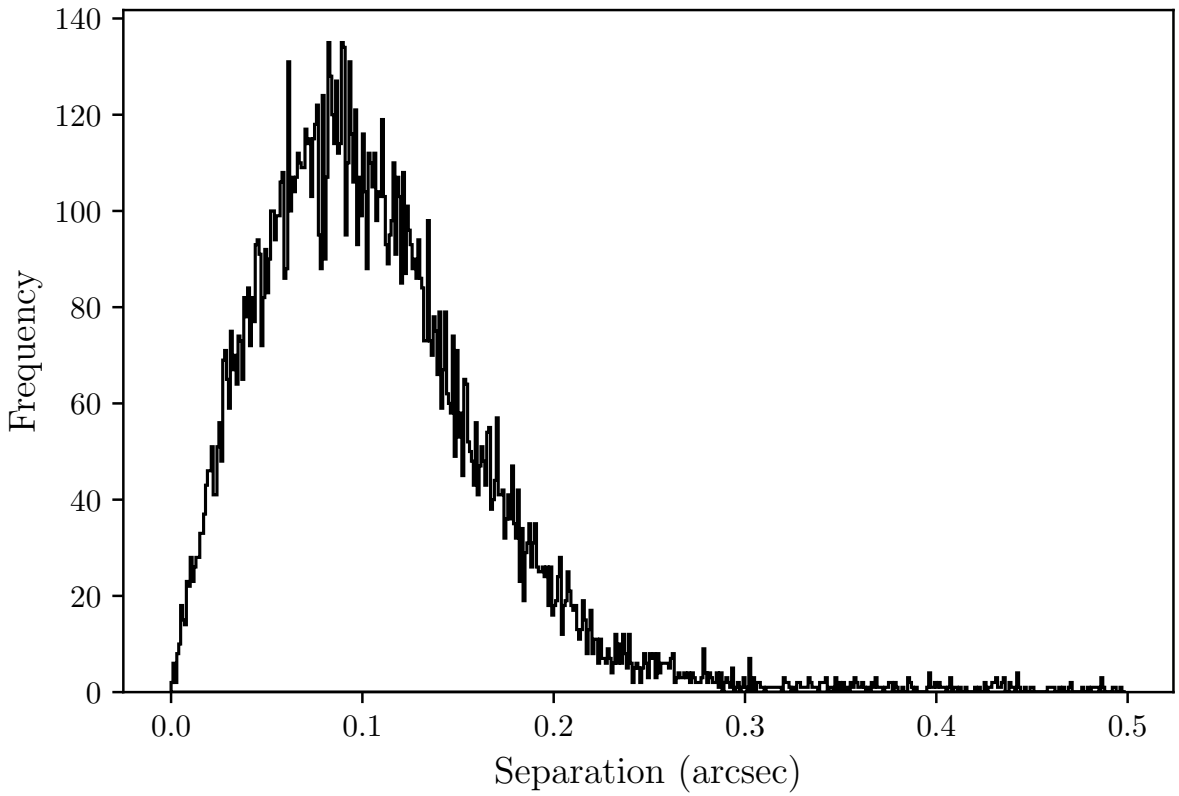

Figure 5: Histogram of the separation between each zCOSMOS objects and its matched COSMOS object. Bin widths are 0.001 arcsec.

in the galaxies with missing data e.g. bright, nearby (low redshift) galaxies are less likely to be targeted by infrared surveys, so the algorithm may associate missing infrared photometry with low redshift. The learner would then perform badly on a dataset where this systematic pattern is not present. This is an example of overfitting: when the ML algorithm performs too well on the training data such that it performs poorly on testing data (often as a result of recognising systematic errors and biases in the training set which are not present in the testing set). To avoid this, any filters with more than 1,000 missing galaxies were removed, then any galaxies with more than 10% of their photometry missing were removed. This resulted in the NUV and ch4 broad bands, and the UVISTA NB118 narrow band being removed (taking the total number of filters from 38 to 35) as well as reducing the number of galaxies to 15,277. The remaining galaxies still had a small fraction of missing photometry so to complete it, each missing magnitude was replaced with the mean of the other galaxies' photometric magnitudes in that particular band. This is a known technique for completing random forest training datasets (Xia et al., 2017).

2.2.6 Calculate Colours

A colour was generated for every unique pair of photometric filters remaining in the dataset. The forest was then trained on this set of colours. Uncertainties in the magnitude

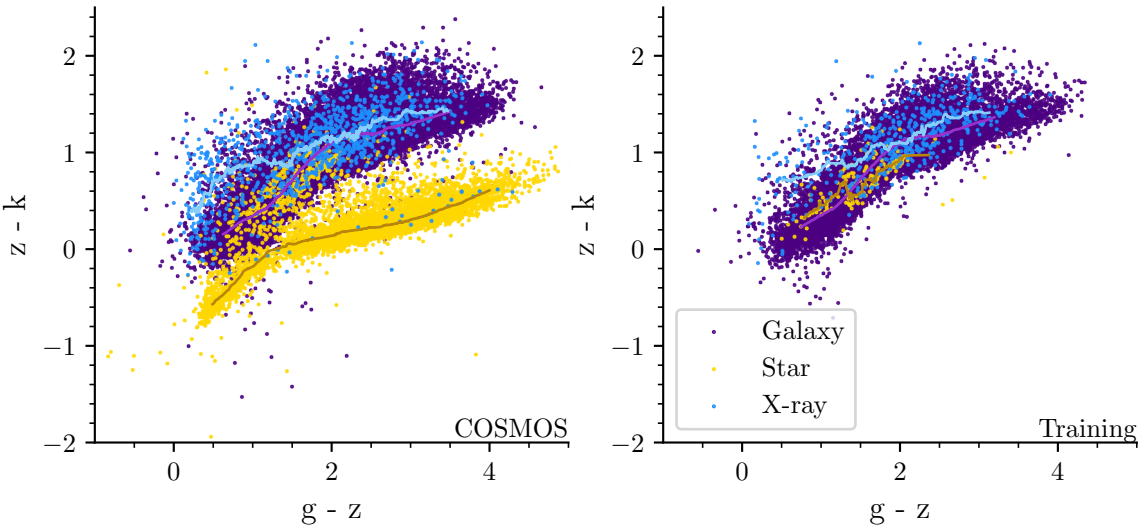

Figure 6: Colour-colour plots for magnitude cut COSMOS objects (left) and objects in the training set (right). The COSMOS objects have been cut such that their Hubble I band magnitude has the same extent as the training set i.e. $15.00 < \mathrm{I_{AB}}(814) < 22.50$. Star-galaxy separation was estimated by Weaver et al. (2022) and used to colour the galaxies purple, the stars yellow and X-ray sources blue. Rolling median lines with 10% window sizes have been plotted in dark yellow and lighter blue and purple for the stars, X-ray sources and galaxies respectively.

were propagated using the method described in § C. In principle, this data contains the same information as the magnitudes alone, however it was discovered in preliminary experiments that the forest saw increased performance when given the colours of each galaxy rather than the raw photometric magnitudes[1]. The 35 magnitudes remaining in COSMOS produced 595 unique colours in the training set. This same observation has been made and quantified using feature importance analysis in similar studies (Brescia et al., 2019). Further work is needed to precisely determine why this is the case for this data. However, it could be because the colours provide the forest with more direct information about the SED than the magnitudes do. If two filters are placed either side of a large spectral feature like the Balmer or Lyman breaks, the difference in photometric magnitude (or colour) between them will be large. Since the regression trees only analyse one magnitude/colour per node, by explicitly providing colour, the forest can identify the large features more easily. If only magnitude is provided then the trees need two nodes to identify these features and there is no guarantee that a given tree will make such a comparison (since the nodes are selected on a greedy basis). Note that, although in reality the forest does not know about the spectral features, it is a useful analogy for explaining its behaviour. In reality it can only interpolate within the multi-dimensional photometry space and it appears that photometry-redshift function is constrained more

[1]Metrics σ_{68} and η improved by factors of 1.46x and 1.61x respectively. Note that since N magnitudes have $\Sigma_{i=1}^{N} i = \frac{N(N-1)}{2}$ colour combinations, the algorithm training time increases since there are more features to consider. In the case of 200 trees and 20 depth, the training time changed by a factor of ∼5x (from ∼7 minutes to ∼35 minutes).

precisely in the colour-redshift space than in the magnitude-colour space. Note also that: in preliminary experiments it was found that training on both colours and magnitudes provided no performance increase to when only colours were trained on.

2.2.7 Test-Train Split

It is important that the training set is representative of the verification set, since if there are statistical differences between them, then they can negatively influence the quality of the photometric redshifts. The filamentary nature of the large-scale, 3D structure of the universe means that the redshift and colour distributions of the training field are unlikely to be uniform by default. For example, a galaxy cluster located in the training field will provide an abundance of galaxies in a particular redshift range, as a result the algorithm will be highly effective at estimating photometric redshifts in that range (and vice-versa for a sparsely populated redshift range, this motivates the need for a large training set). If the algorithm is then tested in a field where there is a different redshift distribution (i.e. there is a dearth of galaxies where there was previously an abundance or vice-versa), then it will perform poorly since it has never seen galaxies in this region of colour-redshift space before. Put simply, the algorithm can only do what it is trained to do. To ensure that the training and verification sets are representative, they are both generated from the same parent dataset. 80% of the 15,277 galaxies in the dataset were used for training and the other 20% were used for verification. This was done at random to ensure there were no statistical differences between them. Therefore the final training set contained 12,216 objects and the verification set contained 3061 objects.

The success of this entire process is demonstrated by fig. 6, which compares the colour-colour distributions of the final 15,277 training objects to the original COSMOS objects (to the same magnitude depth). Stars have colour distributions that are distinct from galaxies, this is shown in the COSMOS objects which are split into two regions in colour-colour space. This splitting corresponds well with star-galaxy separation performed by Weaver et al. (2022). However, after the data processing, the stars have been largely removed and there is only one region remaining. Note that a small fraction of remaining objects have been estimated as stars, these are either inaccurate star-galaxy classifications or highly unusual stars. This star-galaxy separation is used only for visualisation and the forest itself is only given the colour photometry of the objects.

3 Results & Discussion

In section § 3.1 the photometric redshifts estimated by the algorithm are presented and evaluated. Then, in § 3.2 the importance of different filters is discussed before both of these results are compared to literature in § 3.3.

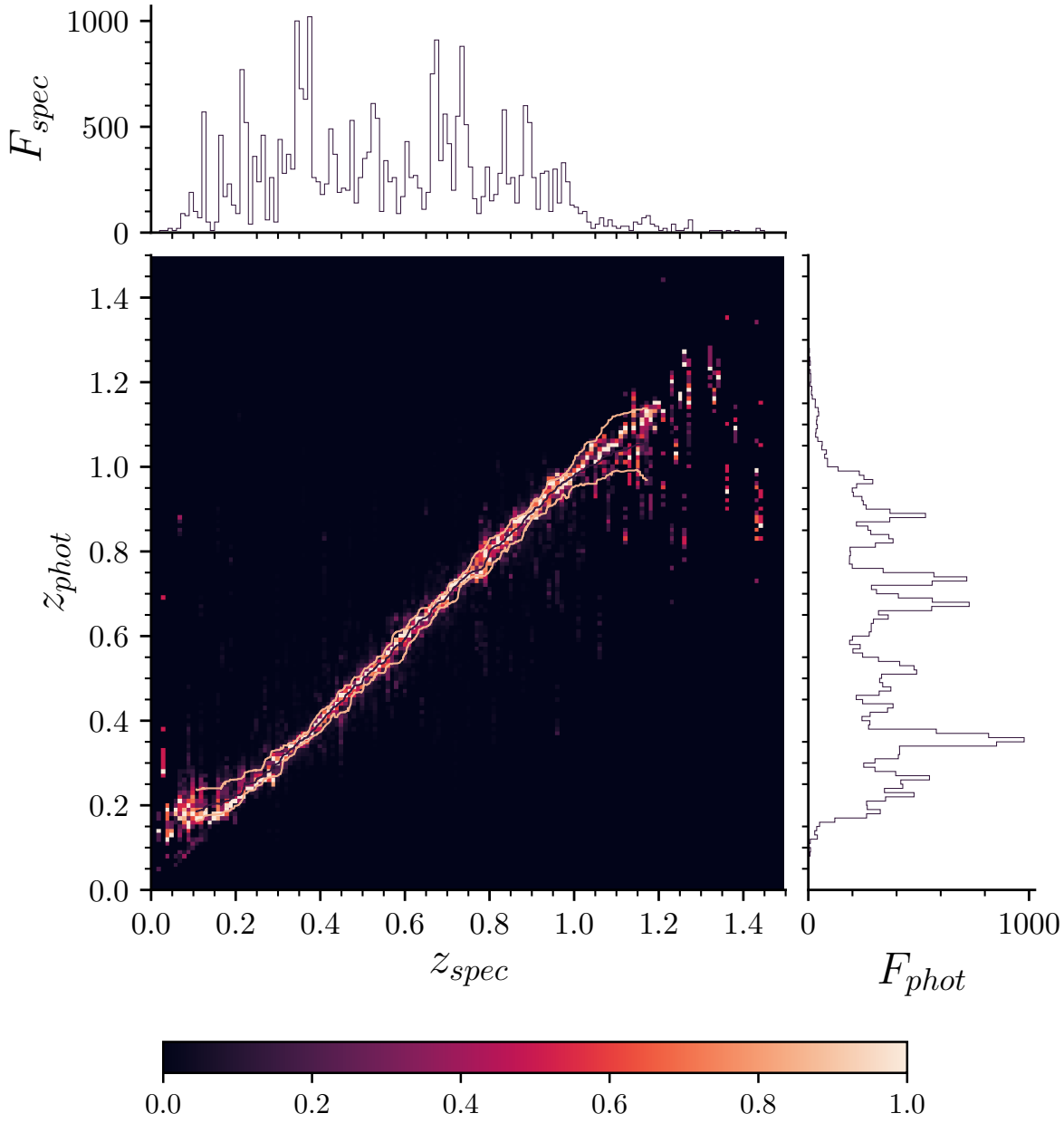

Figure 7: Heatmap showing the relationship between spectroscopic redshift and total photometric redshifts of galaxies in the verification set as estimated by all ten random forest realisations, each with 200 trees and 20 depth. The pixels have a width of 0.01 redshift, bright pixels indicate that more galaxies have been assigned that z_{phot} for a given z_{spec}. Each column's colour map is independent of each other so that the heatmap brightness is not dominated by overpopulated regions. A black rolling median line with window size 100 is superimposed over the heatmap and is flanked by orange rolling 16^{th} and 84^{th} percentile lines which are used in the definition of σ_{68}. Spectroscopic and photometric redshift distributions are shown as histograms on their respective axes.

3.1 Photometric Redshifts

The relationship between the true spectroscopic redshifts and predicted photometric redshifts for galaxies in the verification set is shown in fig. 7. An ideal forest would match z_{phot} to z_{spec} with perfect accuracy such that the heatmap is a bright, one pixel wide diagonal line resembling an identity matrix. This line is visible in fig. 7 but contains a degree of scatter since the real forest is non-ideal. By plotting the results from all 10 realisations and by colouring each column independently, the uncertainty in z_{phot} can be represented visually by the scatter around bright pixels for each z_{spec} column individually. This is particularly pronounced in regions with a dearth of galaxies, for example at very high and low redshift, where individual galaxies dominate the colour scaling of the entire z_{spec} column.

Due to the small field of view of zCOSMOS, the redshift distribution in the training set is non-uniform due to galaxy clustering in the field. As expected, the forest performs well and has a narrow spread in these regions with an abundance of galaxies in the training set e.g. $0.3 < z_{\text{spec}} < 1.0$. Similarly, the heatmap has a large spread in underdense regions e.g. $z < 0.3$ and $z > 1.0$. This is further evidenced by the rolling median and σ_{68} percentile lines which are tightly packed around the diagonal in the centre of the z_{spec} distribution but deviate at high and low redshift regions. As well as training set size, the performance is also affected by the photometry of the galaxies in these regions. For example, high redshift galaxies are fainter and have noisier photometry which results in photometric redshifts with higher variance.

The spectroscopic and mean photometric redshift distributions from all 10 forests can be qualitatively compared in the bottom half of fig. 8. Since the forest is non-ideal, there are differences between the distributions. Although the general shapes of the distributions appear similar, the spectroscopic distribution has pronounced peaks and troughs which are not captured by its photometric equivalent. This is because the photometry given to the forest does not have the resolution required for the forest to determine photometric redshifts to a precision high enough to capture the fine detail of the spectroscopic distribution.

The corresponding cumulative distribution histograms can be compared in the top half of fig. 8. If, for a given redshift bin, there are equally many galaxies in the z_{spec} and z_{phot} distributions, the slope of the cumulative histogram in that bin will be equal. Therefore, if the z_{phot} histogram slope is greater than that of the z_{spec} histogram, the forest has placed too many galaxies in that z_{phot} bin. Note that the slopes being equal does not mean the forest has estimated z_{phot} correctly, it only indicates the quantity of values in the histogram bin. It can be seen that the forest performs poorly for galaxies below $z_{\text{spec}} < 0.3$. The forest never estimates that a galaxy has $z_{\text{phot}} < 0.1$, as evidenced by the standard histograms and the cumulative z_{phot} histogram having a much shallower slope than z_{spec} does. Additionally, the forest places disproportionately many galaxies in the region $0.1 < z_{\text{spec}} < 0.3$, causing the z_{phot} slope to steepen and catch up to the z_{spec} histogram. The histograms then have roughly equal slopes for the remainder of the

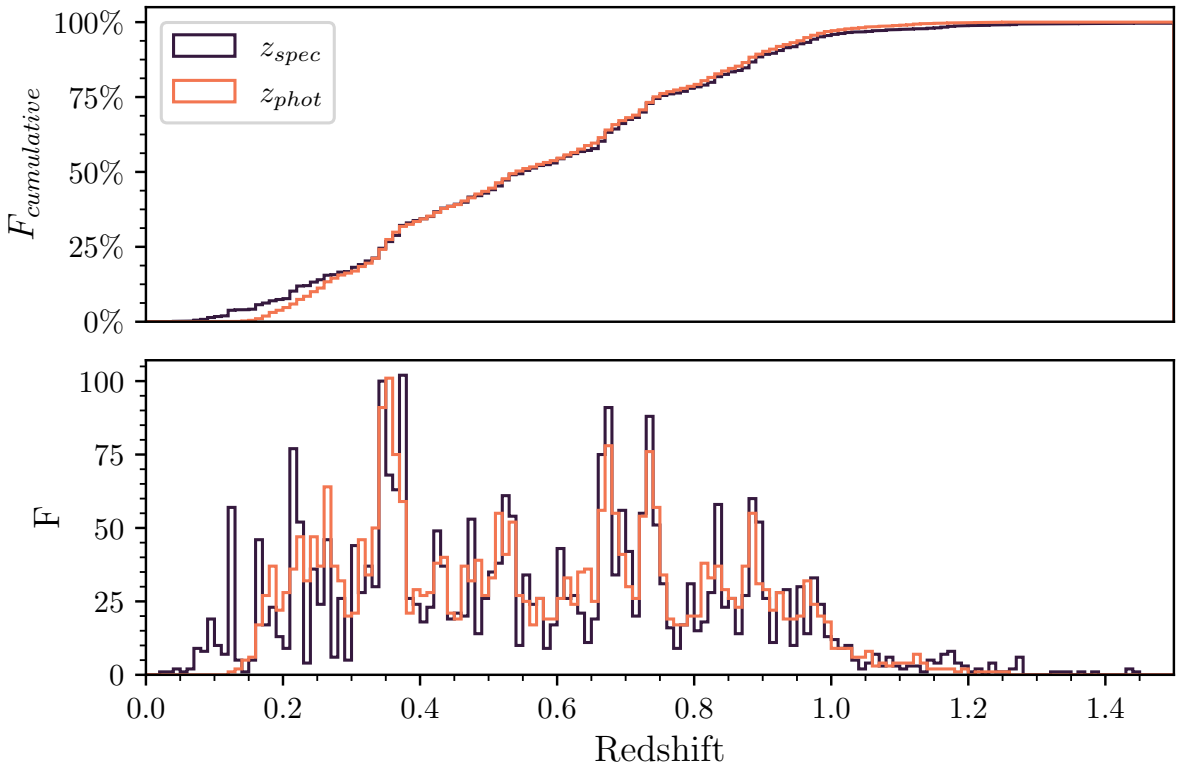

Figure 8: The photometric (orange) and spectroscopic (black) redshift distributions are superimposed as histograms (bottom) and cumulative histograms (top). Photometric redshifts are the mean predictions from all 10 realisations, each with 200 trees and 20 depth. Each histogram bin has a width of 0.01 redshift.

redshift distribution. By analysing fig. 7, it can be seen that this is explained by the forest overestimating redshifts for galaxies in the region $z_{\text{spec}} < 0.1$ by assigning them redshifts in the $0.1 < z_{\text{phot}} < 0.3$ region. Since the forest is a 'black box', it is hard to determine why this is the case without further study.

The fractional difference distribution is presented in fig. 9 along with the metrics for the algorithm with 200 trees, 20 depth and 10 realisations. The bias is close to zero and the modal histogram class is at zero, which indicates that any tendancy to over or under estimate the photometric redshifts is small. The bias being slightly positive is a result of the overestimation of galaxies in the region $z_{\text{spec}} < 0.3$. This is visible in both the scatter plot and in the histogram, since the positive tail has a higher frequency than the negative tail in the region $0.05 < \frac{|z_{\text{spec}} - z_{\text{phot}}|}{1+z_{\text{spec}}} < 0.10$. The rolling bias line confirms this since it becomes highly positive at low redshift. The rolling metric lines allow for a quantitative measure of the algorithm's performance as a function of z_{spec}. They agree with the prior observations that the forest performs poorly at very high and low redshift, since the lines become highly positive or negative in these regions. Equally, the forest performs well at intermediate redshifts when the metrics are close to zero.

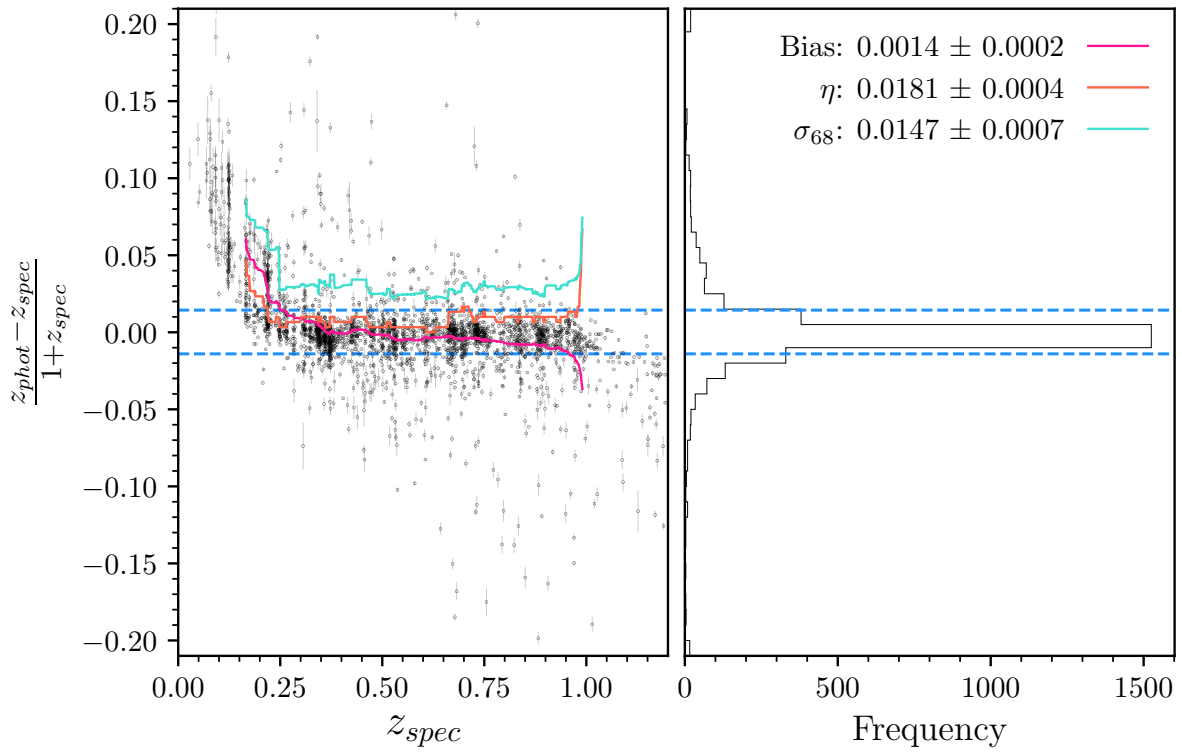

Figure 9: Fractional difference plotted as a function of spectroscopic redshift (left) and the corresponding histogram (right) with bin widths of 0.01 redshift. Photometric redshifts are the mean predictions with standard error from all 10 realisations, each with 200 trees and 20 depth. The small fraction of galaxies beyond the limits of the plot (with absolute fractional differences greater than 0.20) have been added to the outermost bins. The dashed blue lines indicate the 16^{th} and 84^{th} percentiles respectively, these are used in the calculation of σ_{68}. The metrics are displayed to 4 d.p. in the top right corner. To show how the metrics vary as a function of spectroscopic redshift: bias, outlier fraction and σ_{68} have been plotted on a rolling basis in pink, orange and cyan respectively with window sizes of 300.

3.2 Filter Importance

For the COSMOS filter set, broad band photometry gives the forest a low resolution view of a wide wavelength range of the galaxy SEDs, whereas intermediate bands provide a higher resolution view across a smaller range. It is important to understand the respective influences that these filter subsets have on the photometric redshifts since it can guide future research. For example, what combination of filter widths is best for the forest's performance and can future surveys be designed to produce the best training sets for machine learners? In fig. 10 the change in performance is shown when the forest is trained using different colour subsets. As expected, when the GRIZ forest is given only four broad band magnitudes the performance of the forest is poor and the values of all three metrics are high, since the forest sees very little information about the SED. The uncertainties in the metrics for the GRIZ forest are small, implying that the photometric redshifts estimated in each realisation were very similar and that the uncertainties in the GRIZ photometry were low.

Redshift	0	0.1	0.3	0.7	1.2
Balmer (Å)	3646	4011	4740	6198	8021
Lyman (Å)	912	1003	1186	1550	2189

Table 1: Example values of observed Balmer and Lyman break wavelengths at different redshifts, to nearest Å.

In general, adding more photometry to the training set will increase performance and reduce the metric values, however the effects can vary depending on the width and placement of the photometry. For example in fig. 10, σ_{68} increases when the forest is trained on the broad bands compared to the intermediate bands, despite there being more filters in the training set. However the outlier fraction shows the opposite and decreases when training on broad bands. These observations indicate that the intermediate bands allow the forest to estimate higher accuracy photometric redshifts for the central 68% majority of galaxies than the broad bands do, but that the broad bands allow for fewer outliers. This can be explained by comparing the wavelengths in table 1 with the range of wavelengths covered by the intermediate filters in COSMOS as shown by fig. 3 (the minimum intermediate filter peak transmission is at 4263 Å, and the maximum is at 8245 Å). The majority of zCOSMOS galaxies are in the range $0.3 < z_{\text{spec}} < 1$. Therefore, for the majority of galaxies in the verification set, the Balmer break has been redshifted into the domain of the intermediate bands. The intermediate bands therefore provide the forest with higher resolution sampling of this important region of the SED which allows it to estimate photometric redshifts with high accuracy, producing a low σ_{68} value. When the forest is trained using only intermediate bands, it has no way of locating the Balmer break if it has been redshifted outside their wavelength domain, which results in many outlier photometric redshifts and a high value for η. The reverse is true for the forest trained only on broad band photometry. It has information across a much larger wavelength domain, so the Balmer break is less likely to be unfindable (resulting in a lower η), but it has low

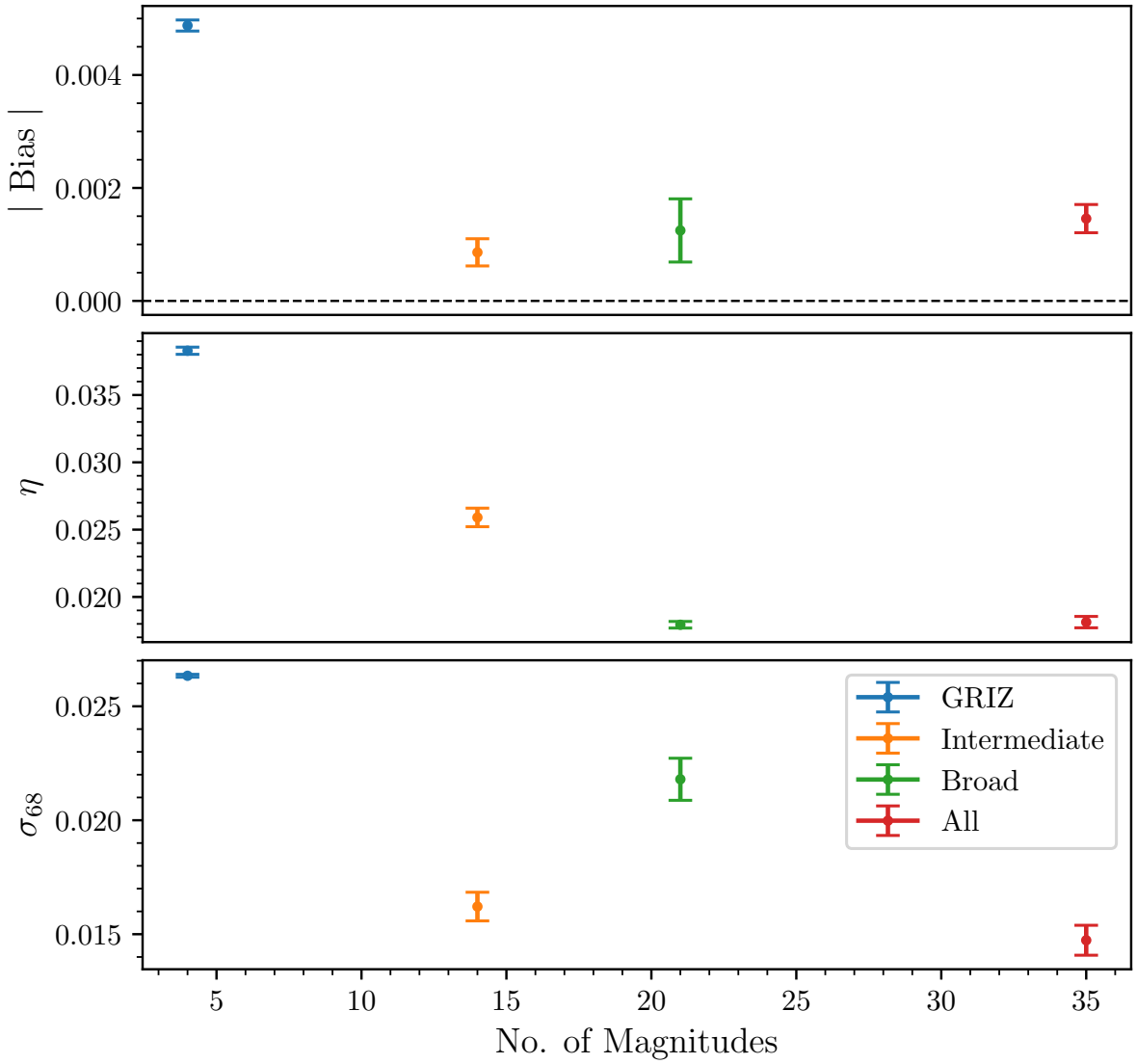

Figure 10: The relationship between the metrics (absolute value of bias (top), outlier fraction (middle) and σ_{68} (bottom)) and the number of magnitudes used in the training set for four forests, each trained on a different magnitude subsets. These four subsets are: GRIZ (only g,r,i,z broad filters), Intermediate (only intermediate width filters), Broad (only broad width filters) and All (both intermediate and broad filters). They are shown in blue, orange, green and red respectively. Note that the forests were trained on all of the unique colour combinations derived from magnitudes in their subset i.e. the GRIZ forest was given colours: g-r, g-i, g-z, r-i, r-z and i-z. All forests contained 200 trees with 20 depth and 10 realisations. A dotted line is shown at the target value $|\text{bias}| = 0$.

resolution sampling of the SED in the region where most of the Balmer breaks are (so it achieves a higher σ_{68}). When both broad and intermediate bands are used then the forest has improved performance, since it has both high resolution information about the Balmer breaks of most galaxies and low resolution information about the very high and low redshift galaxies. Note that the Lyman break is not captured by any of the COSMOS filters at these redshifts after the NUV band is removed.

This demonstrates that including intermediate band photometry in the training set can be hugely beneficial if their wavelength domain matches the redshifted wavelength of important large scale spectral features, in this case the Balmer break. Similar findings have been made in the COSMOS field before for AGNs (Salvato et al., 2011). It follows that narrow band surveys across a wide wavelength range would enable the machine learner to produce high quality redshifts. However, a narrow band survey will, by definition, contain more bands than a broad band survey with the same wavelength range. This means that it takes much longer for a narrow band survey to measure as many galaxies as a broad band survey, which results in the sample sizes being much smaller. Also, narrow bands are noisier than broad bands so require longer exposure times, which further compounds the problem. Therefore, future work is needed to determine the optimum compromise between sample size versus photometric band width and wavelength range.

3.3 Comparison

The results from this report are competitive with current literature values as shown in table 2. Note that many of the literature algorithms presented have trained on galaxies in different fields with varying photometric depth and uncertainty. Due to this, it is not possible to conclude that one algorithm is better than another without additional study, they are only shown to situate the performance of this random forest roughly within the context of the literature.

This forest has achieved σ_{68} and outlier fraction values within the same order of magnitude as some literature random forests (RFs) and convolutional and mixture density neural networks (CNNs and MDNs respectively) trained on broad band (BB) photometry. The neural networks produced by Pasquet et al. (2019) and Eriksen et al. (2020) have achieved σ_{68} values which are an order of magnitude smaller than that of this forest. An avenue of further study could potentially be to investigate the extent to which this forest's performance can be improved by using any techniques employed by these networks (see § 4.3). The MDN presented by Eriksen et al. (2020) is particularly relevant since it is trained on the COSMOS field with narrow band (NB) photometry from the PAU survey (Alarcon et al., 2021) and achieved a very low σ_{68} value. To what extent can this precision be attributed to the narrow photometry and could a random forest achieve the same precison when trained on the PAU survey?

Source	Photometry	Algorithm	σ_{68}	η (%)
Syarifudin et al. (2019)	BB	CNN	0.089	3.35
Hoyle (2016)	BB	CNN	0.030	1.71
Carrasco Kind & Brunner (2013)	BB	RF	0.021*	1.54
This Report	BB & IB	RF	0.0147	1.81
Pasquet et al. (2019)	BB	CNN	0.006	9†
Eriksen et al. (2020)	NB	MDN	0.004	10†

Table 2: Comparison between the performances of machine learners from literature and the random forest. Note that metrics marked with (*) use standard deviation rather than σ_{68} and those marked with (†) have a stricter definition of outlier fraction.

4 Conclusions & Further Work

In § 4.1 the results are collated and summarised before avenues of future research are discussed. In particular, potential improvements to the forest's performance are discussed in § 4.2 and new applications to apply the forest to are discussed in § 4.3.

4.1 Conclusion

A random forest algorithm has been implemented and used to estimate photometric redshifts for galaxies in the COSMOS field. The training set used by the forest contained 12,216 galaxies and was verified using 3061 unseen galaxies with representative redshift and magnitude distributions. Monte Carlo realisations have been used to extend the standard random forest to propagate the effects of uncertainties in the photometry into the redshifts and performance metrics. The forest's optimised hyperparameters were found to be: 200 trees; with 10 realisations; where each tree has 20 depth, is shown the complete colour data and uses MSE as its splitting criterion.

The forest's performance is characterised by the metric values: bias $= 0.0014 \pm 0.0002$, $\eta = 1.81\% \pm 0.04\%$ and $\sigma_{68} = 0.0147 \pm 0.0007$. These are found to be competitive with other ML algorithms used to estimate redshifts using broad band photometry. The forest performs well for galaxies in the central redshift region $0.3 < z_{\text{spec}} < 1.0$ for two reasons. Firstly, the training set contains an abundance of galaxies in this region. Secondly, the redshifted Balmer break wavelengths for galaxies in this region have been sampled at higher resolution by intermediate width filters. Similarly, the galaxy performs poorly for galaxies at high ($z_{\text{spec}} > 1.0$) and low ($z_{\text{spec}} < 0.3$) redshift due to a dearth of galaxies in the training set and a lack of intermediate band photometry in the required wavelength regions. It is found that forests shown only broad band photometry can have their performances increase significantly if they are given intermediate band photometry in a region close to a redshifted large spectral feature. Similarly, forests shown only narrow photometry will perform better when given broad photometry that covers a wider wavelength spectrum.

The reported value of $\sigma_{68} = 0.0147 \pm 0.0007$ is much less precise than the estimated

precision of DESI of $\sigma_{68} = 0.0005$ (Collaboration et al., 2016b). As a result, the photometric redshifts presented here are unsuitable for use in high precision calculations such as studying BAO. However this value is similar to results from other ML codes (see table 2) and SED template codes e.g. EAZY has achieved $\sigma_{68} = 0.034$ in some fields (Brammer et al., 2008). Photometric redshifts known to this precision are suitable for low precision purposes such as studying galaxy evolution (Fontana et al., 2000) and galaxy clustering (Finoguenov et al., 2007).

4.2 Forest Improvements

The forest performed best when the photometry was given as colours rather than magnitudes. It was conjectured that this was because the colours enabled the trees to identify patterns in feature space using one decision node rather than two. To explore this idea, feature importance analysis can be used to assign a weight to each feature in the training set to determine which colour or magnitude is most critical to the forest's performance. An example of this is shown in Brescia et al. (2019) and Hoyle et al. (2015). The efficiency of the forest can be increased by discarding colours and magnitudes that are not important to redshift calculation. Dedicated surveys could then be designed with this consideration in mind. This technique would also allow for more sophisticated comparisons to be made between the intermediate and broad filters, since the total weight of each filter set could be found qualitatively.

The missing photometric magnitudes were replaced with the mean of the other galaxies' magnitudes in the same filter. This is a simple method of dataset completion and it can have the effect of reducing the variance of the photometry distributions. To investigate whether this has had a negative impact on the forest, alternative completion methods can be introduced. For example, Bayesian methods can used to complete missing data in a more sophisticated and "statistically principled" manner Ma & Chen (2018) or the random forest itself can be modified to function without complete data Xia et al. (2017).

This random forest consists of a relatively small number of deep trees. However, if the forest is instead built from many shallow trees, it can be gradient-boosted to improve its performance. Rather than building the entire forest at once, the gradient-boosted forest builds it gradually, so that it can train trees on data that have been classified poorly by previous trees. This forest struggled at high and low redshift where there was a dearth of galaxies, however if it were gradient-boosted then more trees would be assigned to these regions and the performance may improve. Gradient-boosted forests have been used for photometric redshift estimation and generally outperform traditional random forests (Gerdes et al., 2010; Hoyle et al., 2015).

4.3 Further Work

It has been shown that a combination of broad bands and well-placed intermediate bands can generate a robust training set for the forest. However, there are many open questions about whether this training set can be improved upon. Can the performance of the

forest be improved if its training data was replaced by the PAU survey which consists of the standard u,g,r,i,z,y filters and 40 narrow filters spanning the wavelength range 4500 Å to 8500 Å (Alarcon et al., 2021)? The narrow band range of PAU is similar to the intermediate filter range of COSMOS, would there be a measurable difference when training on only the narrow/intermediate bands? Would the outlier fraction of the PAU-trained forest be improved by the addition of COSMOS broad band photometry outside the PAU wavelength range i.e. redder than the J-band? Can a random forest match the impressive σ_{68} achieved by Eriksen et al. (2020) using narrow bands, or does their addition of transfer learning and use of a MDN prevent this?

5 Acknowledgements

- Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under ESO programme ID 179.A-2005 and on data products produced by TERAPIX and the Cambridge Astronomy Survey Unit on behalf of the UltraVISTA consortium.
- Based on zCOSMOS observations carried out using the Very Large Telescope at the ESO Paranal Observatory under Programme ID: LP175.A-0839.
- This report made use of the TOPCAT software (Taylor, 2005).
- The following python packages were used: scikit-learn Pedregosa et al. (2011), numpy Harris et al. (2020), pandas Reback et al. (2022), matplotlib Hunter (2007), seaborn Waskom (2021) and scipy Virtanen et al. (2020).
- The author would like to thank supervisors Prof. Baugh and Prof. Norberg for their astute advice and guidance.

References

Alarcon A., et al., 2021, Monthly Notices of the Royal Astronomical Society, 501, 6103

Arnouts S., Ilbert O., 2011, Astrophysics Source Code Library, p. ascl:1108.009

Baum W. A., 1962, 15, 390

Bean A. J., et al., 1983, Monthly Notices of the Royal Astronomical Society, 205, 605

Beck R., et al., 2017, Astronomy and Computing, 19, 34

Beutler F., et al., 2011, Monthly Notices of the Royal Astronomical Society, 416, 3017

Bilicki M., et al., 2018, A&A, 616, A69

Bolzonella M., Miralles J.-M., Pello' R., 2000, arXiv:astro-ph/0003380

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Brescia M., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 663

Camera S., et al., 2012, Monthly Notices of the Royal Astronomical Society, 427, 2079

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

Collaboration D., et al., 2016b, arXiv:1611.00036 [astro-ph]

Collaboration D., et al., 2016a, arXiv:1611.00037 [astro-ph]

Collaboration T. L. D. E. S., et al., 2021, arXiv:1809.01669 [astro-ph]

D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111

Davies L. J. M., et al., 2015, Monthly Notices of the Royal Astronomical Society, 452, 616

Davis M., Peebles P. J. E., 1983, The Astrophysical Journal, 267, 465

Dunlop J. S., 2013, arXiv:1205.1543 [astro-ph 10.1007/978-3-642-32362-1_5, 396, 223

Eisenstein D. J., et al., 2005, ApJ, 633, 560

Eriksen M., et al., 2019, Monthly Notices of the Royal Astronomical Society, 484, 4200

Eriksen M., et al., 2020, Monthly Notices of the Royal Astronomical Society, 497, 4565

Finoguenov A., et al., 2007, ApJS, 172, 182

Fontana A., et al., 2000, AJ, 120, 2206

Gerdes D. W., et al., 2010, ApJ, 715, 823

Harris C. R., et al., 2020, Nature, 585, 357

Hoyle B., 2016, Astronomy and Computing, 16, 34

Hoyle B., et al., 2015, Monthly Notices of the Royal Astronomical Society, 449, 1275

Hubble E., 1929, Proceedings of the National Academy of Sciences, 15, 168

Hughes I., Hase T. P. A., 2010, Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis. New York : Oxford University Press, Oxford

Hunter J. D., 2007, Computing in Science and Engineering, 9, 90

Ilbert O., et al., 2009, ApJ, 690, 1236

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

Knobel C., et al., 2012, ApJ, 753, 121

Lilly S. J., et al., 1996, ApJ, 460

Lilly S. J., et al., 2009, ApJS, 184, 218

Ma Z., Chen G., 2018, J. Korean Stat. Soc., 47, 297

Massarotti M., et al., 2001, A&A, 380, 425
Mohammad F. G., et al., 2018, A&A, 610, A59
Mucesh S., et al., 2021, Monthly Notices of the Royal Astronomical Society, 502, 2770
Norris R. P., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, Art. No. 108004
Pasquet J., et al., 2019, A&A, 621, A26
Peacock J. A., et al., 2001, Nature, 410, 169
Pedregosa F., et al., 2011, J. Mach. Learn. Res., 12, 2825
Rau M. M., et al., 2015, Mon. Not. R. Astron. Soc., 452, 3710
Reback J., et al., 2022, Pandas-Dev/Pandas: Pandas 1.4.2, Zenodo, doi:10.5281/zenodo.6408044
Salvato M., et al., 2009, ApJ, 690, 1250
Salvato M., et al., 2011, ApJ, 742, 61
Salvato M., Ilbert O., Hoyle B., 2019, Nat Astron, 3, 212
Stivaktakis R., et al., 2020, IEEE Trans. Big Data, 6, 460
Syarifudin M. R. I., Hakim M. I., Arifyanto M. I., 2019, J. Phys.: Conf. Ser., 1231, 012013
Tanaka M., 2015, ApJ, 801, 20
Taylor M. B., 2005, 347, 29
Vanzella E., et al., 2004, A&A, 423, 761
Vargas-Magana M., et al., 2019, arXiv:1901.01581 [astro-ph]
Virtanen P., et al., 2020, Nat Methods, 17, 261
Wang Y., Xu L., Zhao G.-B., 2017, ApJ, 849, 84
Waskom M. L., 2021, Journal of Open Source Software, 6, 3021
Weaver J. R., et al., 2022, ApJS, 258, 11
Xia J., et al., 2017, Pattern Recognition, 69, 52
Yuan F.-T., et al., 2019, A&A, 631, A123

Appendices

A Confidence Classes

The zCOSMOS redshifts have a range of reliabilities and each object has been assigned a confidence class by Lilly et al. (2009) to represent this. The confidence class is calculated from the spectroscopic verification rate and the comparison between the z_{spec} from zCOSMOS and the z_{phot} from COSMOS. An ideal galaxy has a well defined SED which enables a secure z_{spec} measurement (i.e. verification rate of 100%) in good agreement with its z_{phot}. The data release description recommends only keeping objects with good confidence classes, hereafter referred to as the confidence cut. These are objects in any of the following categories:

- Secure (99%) or very secure (99.8%) spectroscopic redshifts, regardless of COSMOS photometric redshift data.
- Probable redshift (94%), as long as it not in disagreement with any z_{phot} measurements.
- Insecure redshift (70%), if it has a z_{phot} measurement that it is in agreement with.
- Redshift measured using only one spectral line, if not in disagreement with z_{phot}. If this spectral line is from an AGN then it must instead have a z_{phot} measurement that is in agreement with it.

This confident sample comprises 88% of zCOSMOS and has a mean spectroscopic verification rate of 99% Lilly et al. (2009). The confidence cut should remove objects with insecure redshifts, these are likely objects that are too faint for spectroscopy or so bright that they could be stars. If the confidence cut is valid then we expect to have fewer objects outside the intended magnitude limits and this is seen in fig. 4. The histogram slopes are parallel when inside the magnitude limits, meaning that the confidence cut is independent of magnitude for regular objects, but outside the limits the number of objects is reduced significantly.

B Redshift Distribution

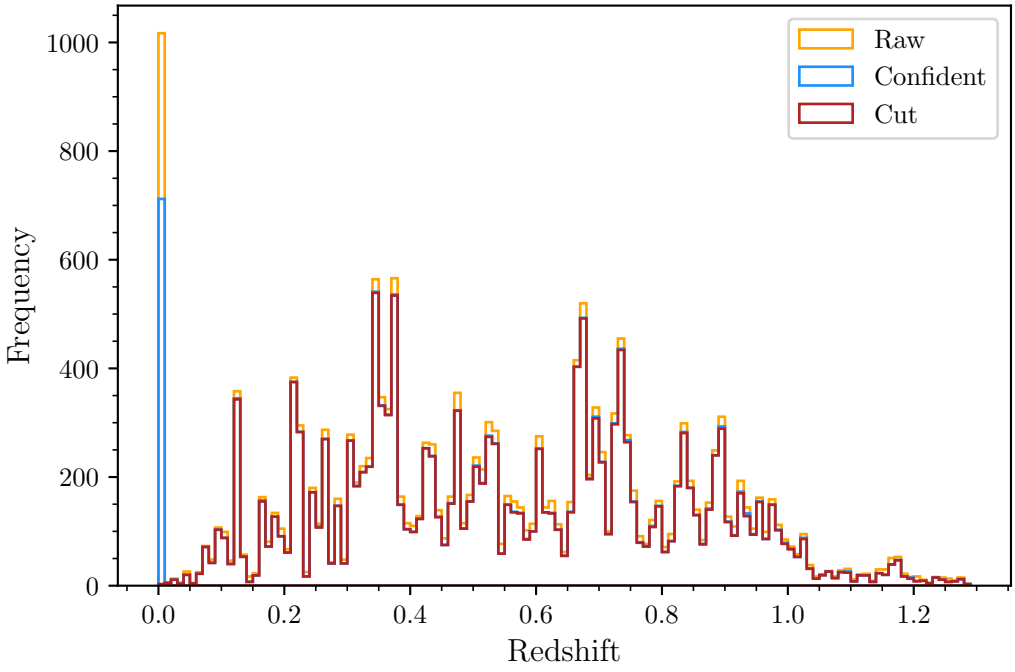

Figure 11: Redshift distribution of unedited zCOSMOS (orange) as well as how it changes after being affected by first the confidence cut (blue) and then the $z > 0.002$ redshift cut (red). The histogram bins have width $\Delta z = 0.01$. The redshift axis has been shown over the range $0 < z < 1.3$ since this is where the vast majority of objects are located.

C Error Propagation

Errors are propagated from photometric magnitudes into colours using the functional method described by Hughes & Hase (2010). A general colour C is defined as the difference between two magnitudes A and B:

$$C(A, B) = A - B \tag{7}$$

The uncertainties in A and B are α_A and α_B respectively. The uncertainties in C due to the uncertainty in A and B respectively (α_C^A and α_C^B) are then:

$$\alpha_C^A = |C(A + \alpha_A, B) - C(A, B)| = |\alpha_A| \tag{8}$$

$$\alpha_C^B = |C(A, B + \alpha_B) - C(A, B)| = |\alpha_B| \tag{9}$$

The total error in C (α_C) is then simply:

$$(\alpha_C)^2 = (\alpha_A)^2 + (\alpha_B)^2 \tag{10}$$