Title

Photometric Redshift and Ellipticity Measurements for Cosmology with Probabilisitic Neural Networks

Author

Jones, Evan

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Photometric Redshift and Ellipticity Measurements for Cosmology with Probabilisitic Neural Networks

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Astronomy & Astrophysics

by

Evan Jones

2024

ABSTRACT OF THE DISSERTATION

Photometric Redshift and Ellipticity Measurements for Cosmology with Probabilisitic Neural Networks

by

Evan Jones
Doctor of Philosophy in Astronomy & Astrophysics
University of California, Los Angeles, 2024
Professor Tuan H. Do, Chair

Cosmological weak lensing probes can inform us of the contents and evolution of the universe, including the properties of dark matter and dark energy, which collectively make up $\sim 95\%$ of the universe. We live in an exciting period in scientific history; large scale astronomical surveys such as the Legacy Survey of Space and Time (LSST) will soon provide imaging for over a billion celestial objects, which timely coincides with recent advancements in probabilistic image-based machine learning. It is incumbent on scientists to leverage recent advancements to extract as much information as possible from large scale astronomical surveys to probe our universe. This thesis contains my contribution toward this objective.

Precision cosmological measurements require accurate data analysis with precise uncertainties. The two critical data analysis tasks for weak lensing cosmological probes are 1) photometric redshift (photo-z) estimation and 2) galaxy shear estimation. These quantities allow us to map the distribution of galaxies in the sky and quantify the distribution of dark matter. Here we present results for photo-z estimation and galaxy shape estimation using probabilistic neural networks, using a novel dataset derived from the Hyper Suprime-Cam

(HSC) Survey.

In Chapter 1, we provide an introduction to weak lensing cosmological probes, photo-z estimation, and shear estimation. In Chapter 2, we introduce the machine-learning-ready dataset derived from HSC consisting of galaxy photometry, galaxy images, and spectroscopic redshifts. We make this dataset publicly available and utilize it for all photo-z estimation analyses in this work. In Chapter 3, we present a probabilistic photo-z estimation model using a Bayesian neural network (BNN) and compare its performance to alternative methods. In Chapter 4, we present an image-based probabilistic photo-z estimation model using a Bayesian convolutional neural network (BCNN) and compare its performance to alternative methods. In Chapter 5, we present an image-based probabilistic model for galaxy ellipticity estimation (as a proxy for shear estimation) evaluated on HSC galaxy images using a custom BCNN. In the Appendix we provide a roadmap by which one can utilize the photo-z and potential shear estimation models in this thesis to perform a weak lensing measurement.

The dissertation of Evan Jones is approved.

Jack Singal

Tommaso L. Treu

Matthew Arnold Malkan

Tuan H. Do, Committee Chair

University of California, Los Angeles

2024

to my Mother, Ruth Ann Jones, and the rest of my family

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This thesis is a culmination of the efforts of many. Beyond the fact that I have stood on the shoulders of giants to make my contribution possible, I would like to thank those in my life who have provided me guidance and support. I would first like to thank my Mother for providing me every opportunity to chase my passions. I would also like to thank Carl Sagan for writing Cosmos, which initially inspired me to study Astrophysics. I would like to thank Jack Singal for accepting me as an undergraduate research student and introducing me to machine learning and data analysis. I must also thank my advisor, Tuan Do, for making me a better scientist through innumerable discussions and for offering me an incredible amount of support while I faced a number of unexpected challenges along the way.

I acknowledge that results presented in this thesis are based on published works with additional co-authors. In particular, Chapter two is derived from Do et al. 2024 (in prep), Chapter Three is derived from Jones et al. (2021b) and Chapter Four is derived from Jones et al. (2023).

VITA

2018 B.A. (Physics)
University of Richmond

2020 M.S. (Astronomy & Astrophysics)
University of California, Los Angeles.

PUBLICATIONS

Jones, E., Do, T., Boscoe, B., Singal, J., Wan, Y. and Nguyen, Z., 2024. Redshift Prediction with Images for Cosmology using a Bayesian Convolutional Neural Network with Conformal Predictions. The Astrophysical Journal, Submitted.

Jones, E., Do, T., Boscoe, B., Singal, J., Wan, Y. and Nguyen, Z., 2024. Improving Photometric Redshift Estimation for Cosmology with LSST Using Bayesian Neural Networks. The Astrophysical Journal, 964(2), p.130.

Jones, E., Do, T., Boscoe, B., Wan, Y., Nguyen, Z. and Singal, J., 2022. Photometric Redshifts for Cosmology: Improving Accuracy and Uncertainty Estimates Using Bayesian Neural Networks.

Boscoe, B., Do, T., Jones, E., Li, Y., Alfaro, K. and Ma, C., 2022. Elements of effective machine learning datasets in astronomy. arXiv preprint arXiv:2211.14401.

Singal, J., Silverman, G., Jones, E., Do, T., Boscoe, B. and Wan, Y., 2022. Machine Learning Classification to Identify Catastrophic Outlier Photometric Redshift Estimates. The Astrophysical Journal, 928(1), p.6.

Jones, E. and Singal, J., 2020. Tests of Catastrophic Outlier Prediction in Empirical Photometric Redshift Estimation with Redshift Probability Distributions. Publications of the Astronomical Society of the Pacific, 132(1008), p.024501.

Jones, E. and Singal, J., 2017. Analysis of a custom support vector machine for photometric redshift estimation and the inclusion of galaxy shape information. Astronomy Astrophysics, 600, p.A113.

Singal, J., Kogut, A., Jones, E. and Dunlap, H., 2015. Axial ratio of edge-on spiral galaxies as a test for bright radio halos. The Astrophysical Journal Letters, 799(1), p.L10.

Singal, J., Haider, J., Ajello, M., Ballantyne, D.R., Bunn, E., Condon, J., Dowell, J., Fixsen, D., Fornengo, N., Harms, B. and Holder, G., 2018. The radio synchrotron background: conference summary and report. Publications of the Astronomical Society of the Pacific, 130(985), p.036001.

CHAPTER 1

Introduction

To investigate dark matter and dark energy, the upcoming Legacy Survey of Space and Time (LSST Ivezić et al., 2008) and Euclid (Collaboration et al., 2022) will provide observations for tens of billions of objects throughout their operation. The astronomical community will require sophisticated methodology to extract scientific information from these observations. Weak lensing measurements of galaxies provide insight into the development of cosmic structure and serve as a critical cosmological probe that relies ultimately on accurately and precisely measuring both the 1) redshifts and 2) shears of hundreds of millions of galaxies with well-constrained uncertainties.

The redshift z of a galaxy is the measurable quantity that tells us how far away it is from the Earth and how far back in time the light that we observe was emitted from it. The redshift of a galaxy is defined as

$$1 + z \equiv \frac{f_{em}}{f_{obs}}, \tag{1.1}$$

where f_{em} is the frequency of emitted light from a source and f_{obs} is the frequency at which we observe that light.

Spectroscopic measurements are the most reliable method of obtaining redshift, but are too time consuming and therefore not a suitable solution for obtaining the number of redshifts required for cosmological analyses with large scale survey data. Photometric redshift (Photo-z) estimation, where the redshift is estimated from a galaxy's brightness in a limited number of wide photometric bands, can provide redshifts for billions of galaxies in a tractable amount of time. However, photo-z estimates are subject to significant systematic errors because the

spectral information of a galaxy is sampled with only a limited number of imaging bands. Obtaining sufficiently accurate photo-z estimates and understanding the error properties of these estimates is still a a major challenge (Huterer et al., 2006; Newman & Gruen, 2022).

Cosmic shear measurements are the other necessary component to infer cosmological parameters from a weak lensing measurement. Cosmic shear results from the bending of light from distant galaxies due to gravitational interactions with large scale structure. Weak lensing refers to gravitational lensing in the limit that deflected photons cause only small changes in the observed position, size, brightness, and shape of galaxies.

Obtaining accurate shear estimates and uncertainties is also a major challenge. The difficulty of developing effective shear estimation models is compounded by the absence of 'true' shear values with which to train a model using real galaxy data.

The extent to which galaxy orientations deviate from a random distribution is thought to result from lensing. Galaxy shape distortions from lensing are typically on the order of 1% the size of the galaxy, which is a far smaller contribution than typical instrinsic galaxy ellipticities (~ 0.3). This relative signal weakness is compounded by coherent distortions produced from light propagating through the atmosphere, telescope optics, and incomplete knowledge of point spread functions (PSFs). Accurately measuring lensing shears for galaxies with low signal-to-noise ratios is an on-going challenge in weak lensing analyses. Systematic errors resulting from shape measurement must be reduced by factors of 5-10 (Mandelbaum et al. (2014)).

Shear can be quantified as the bulk alignment of galaxies as a function of angular separation. Complex shear is defined as

$$\gamma = |\gamma| e^{-2i\phi}, \tag{1.2}$$

where ϕ is the angular separation over which the shear is measured. The tangential and

cross components of the shear are defined respectively as

$$\gamma_+ = -Re[\gamma], \gamma_x = -Im[\gamma]. \tag{1.3}$$

We can express the tangential and cross components of cosmic shear as two-point correlation functions:

$$\xi_+ = \langle \gamma\gamma^* \rangle = \langle \gamma_t\gamma_t \rangle + \langle \gamma_x\gamma_x \rangle \tag{1.4}$$

$$\xi_- = Re[\langle \gamma\gamma \rangle e^{-4i\phi}] = \langle \gamma_t\gamma_t \rangle - \langle \gamma_x\gamma_x \rangle \tag{1.5}$$

Weak lensing is quantified as a mapping between unlensed coordinates (x_u, y_u) and lensed coordinates (x_l, y_l) (Mandelbaum, 2018):

$$\begin{pmatrix} x_u \\ y_u \end{pmatrix} = \begin{pmatrix} 1-\gamma_1-\kappa & -\gamma_2 \\ -\gamma_2 & 1+\gamma_1-\kappa \end{pmatrix} \begin{pmatrix} x_l \\ y_l \end{pmatrix}, \tag{1.6}$$

where the complex lensing shear describing the stretching of galaxy images is given by $\gamma = \gamma_1 + i\gamma_2$, and the convergence κ reflects the change in size and brightness. This can be restated in terms of the reduced shear, $g = \gamma/(1-\kappa)$:

$$\begin{pmatrix} x_u \\ y_u \end{pmatrix} = (1-\kappa) \begin{pmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1 \end{pmatrix} \begin{pmatrix} x_l \\ y_l \end{pmatrix}, \tag{1.7}$$

For weak-lensing, $\gamma \simeq g$ is assumed because κ is small. The $(1-\kappa)$ factor influences the magnification, which is measured via observed bias in object number density or size distributions (Jee et al., 2013). For a galaxy population with random orientations, the average measured elipticity of the set provides an unbiased shear estimate (Mandelbaum et al., 2015)).

Shear systematic errors are commonly described in a linear model:

$$\hat{\gamma} = (1 + m)\gamma + c, \tag{1.8}$$

where γ is the true lensing shear and $\hat{\gamma}$ is the estimated shear. There are two bias terms: m is the multiplicative bias and c is the additive bias. The additive bias is generally linearly proportional to the PSF ellipticity (Mandelbaum et al., 2015). In order to statistically reduce uncertainties on cosmological parameters, we need to reduce the systematic errors in weak lensing by reducing both shear bias terms.

Systematic errors in photo-z and shear estimation can manifest as outlier predictions that are far from their true value, biases in the distribution of predictions, and large scatter in predictions (e.g. Newman & Gruen, 2022; Mandelbaum, 2018). These systematics strongly affect science goals such as weak lensing inferences of cosmological parameters since photo-z and shear uncertainties will be propagated into models constraining cosmological quantities. Any photo-z or shear model developed for the potential application to these science missions must produce uncertainties on predictions.

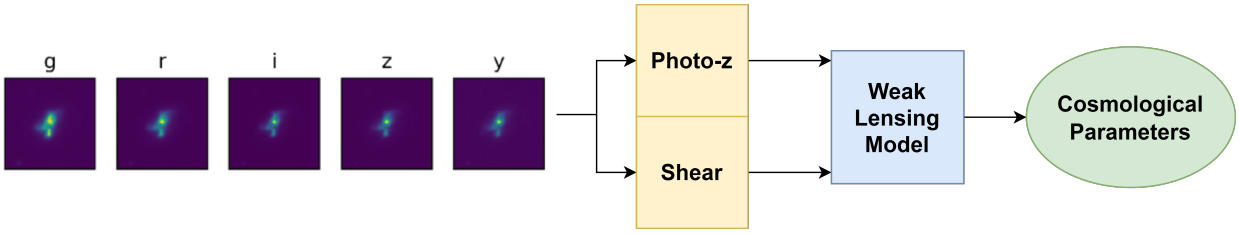

Figure 1.1 Components of a cosmological weak lensing pipeline. The analysis begins with galaxy images, which are used to produce photo-z and shear estimates. Weak lensing measurements obtained from the shear and photo-z estimates are compared to a theoretical weak lensing model to produce cosmological constraints.

Machine learning developments made over the last decade have positioned it as an ideal candidate for extracting scientific information from images provided by large scale surveys. Machine learning is a general class of algorithms that learn from data. Machine learning are ideal for situations where: 1) analytical models or likelihoods are not well-defined, 2) there

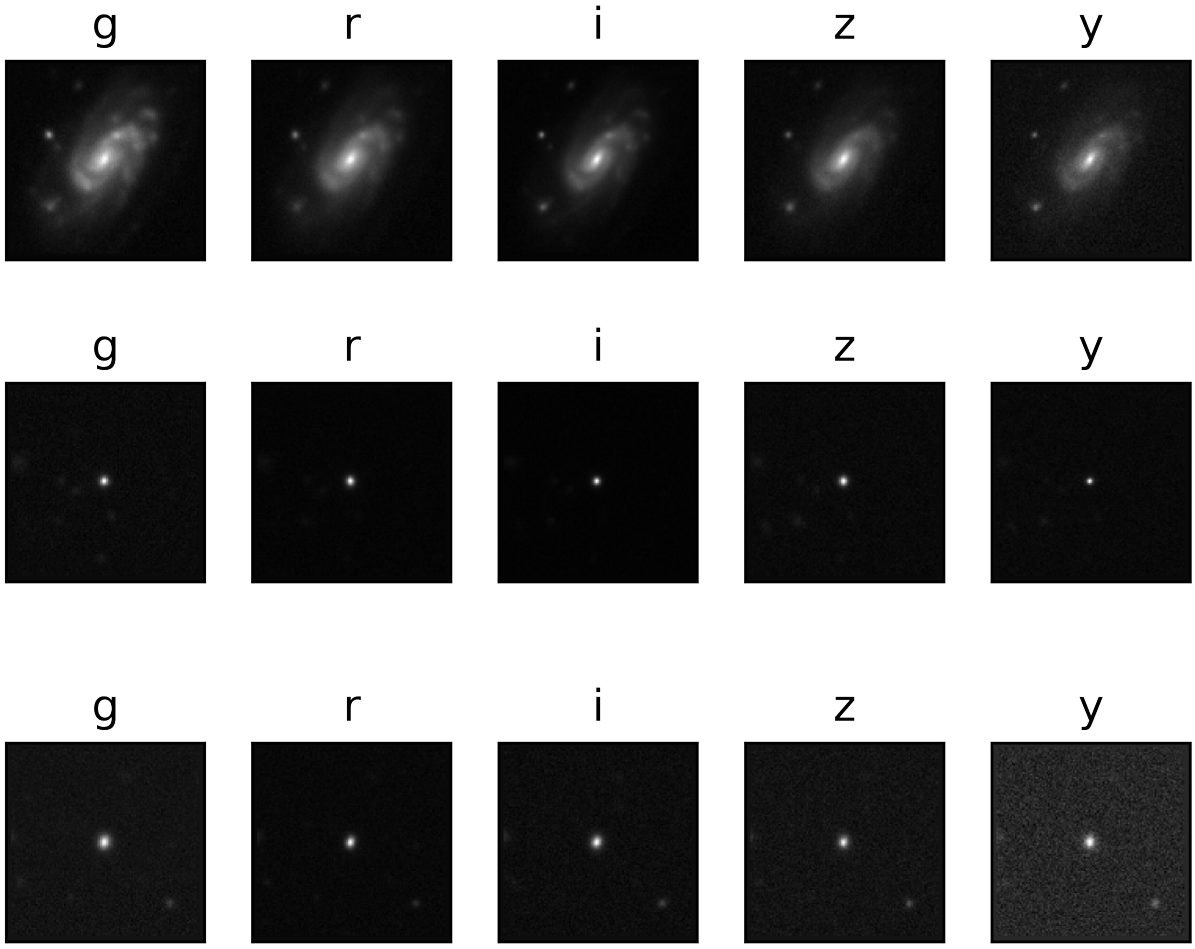

Figure 1.2 Example HSC galaxy images with grizy photometry for a low redshift galaxy at $z = 0.05$ (TOP), and a high redshift galaxy at $z = 3.92$ (MIDDLE), and another low redshift galaxy at $z = 0.14$. The similarity between the high redshift galaxy and the bottom low redshift galaxy highlights the difficulty of photo-z and shear estimation.

is an abundance of training data, and 3) one has access to high performance computing. Machine learning applications for photometric redshifts and shear is ideal because 1) galaxy images are not easily described with an analytical function, 2) precursor surveys to LSST (such as HSC) are providing hundreds of thousands of galaxy images with spectroscopic redshifts, and 3) efficient algorithms for image recognition and processing, as well as powerful GPUs, are now available.

For the analyses in this thesis I leverage recent developments in Bayesian neural networks and Bayesian convolutional neural networks – types of probabilistic neural networks (NNs) (Jospin et al., 2020) for photo-z and galaxy shape estimation. Previous works in machine learning for photo-z and shear estimation have been largely limited to point estimate predictions and do not produce accurate uncertainties until recently. Probabilistic neural networks, conceptualized in the 1990s (Specht, 1990), have previously been limited in their ability to process the size of data required for performing photo-z or shear estimation for large scale surveys because of the complexity of their computation. However, recent breakthroughs in conceptual understanding and computational capabilities (e.g. Filos et al. (2019); Dusenberry et al. (2020)) now make probabilistic deep learning possible for cosmology. Probabilistic deep learning with a BNN has many advantages compared to traditional neural networks, including better uncertainty representations, better point predictions, and better interpretability of neural networks because they can be viewed through the lens of probability theory. In this way one can draw upon decades of development in Bayesian inference analyses.

Current photo-z and shear models do not satisfy the LSST photo-z and shear science requirements (Collaboration et al., 2021). In this thesis I use the LSST science requirements as the basis for evaluating our models. We also include a discussion of alternative models that have been applied to the same data, when available, and introduce additional probabilistic metrics for evaluating the quality of uncertainty estimates.

Optimizing machine learning models for shear prediction in cosmological pipelines re-

quires the simulation of a galaxy dataset with 'known' shear values since it is not possible to measure this directly. In the absence of a realistic simulated galaxy image dataset, we need to rely on alternative estimation methods to obtain training labels for a machine learning model. Galaxy ellipticities, while also not possible to measure their intrinsic values directly, are $\sim 30 - 50$ times larger than galaxy shears and thus contain a larger signal-to-noise ratio and can better serve as a test for a model's ability to measure galaxy shape information given noisey training labels. In this work, I present a proof of concept for shear estimation utilizing galaxy images in machine learning models by estimating ellipticity as a proxy for shear.

The methods presented in this thesis serve as a framework for probabilistic deep learning image analyses of galaxies provided by large scale extragalactic surveys with the goal of advancing precision cosmological inference with weak lensing probes. This work is divided into four main sections:

1. Creating a dataset of galaxy images, photometry, and spectroscopic redshifts
2. Estimating photo-zs with a BNN using 5-band grizy photometry
3. Estimating photo-zs with a BCNN using 5-band grizy images
4. Estimating galaxy ellipticity with a BCNN and galaxy images

The unique aspects of my thesis are:

1. Largest publicly available machine-learning-ready galaxy image dataset for photo-z estimation: $\sim 300k$ galaxies from the Hyper Suprime-Cam survey containing five-band photometric images and known spectroscopic redshifts from $0 < z < 4$
2. One of the first, and best performing, probabilistic machine learning model (BNN) for photo-z estimation applied to data representative of LSST ($0 < z < 4$)

3. One of the first, and best performing, probabilistic image-based machine learning model (BCNN) for photo-z estimation applied to data representative of LSST ($0 < z < 4$)

4. One of the first applications of a probabilistic image-based machine learning model for ellipticity estimation (BCNN) applied to data representative of LSST ($0 < z < 4$)

CHAPTER 2

Building Galaxy Datasets

2.1 Introduction

One of the major questions in Physics and Astrophysics is the nature of dark matter and dark energy. Dark matter and dark energy represent over 95% of the energy density of the universe, but we do not know of their particle or field nature. One of the most promising approaches to investigate their nature is through observations of the Universe on cosmic scales. Large science surveys such as LSST Collaboration et al. (2021) and the Euclid mission Collaboration et al. (2022) aim to observe billions of galaxies to map their distribution throughout cosmic time to constrain models of dark matter and dark energy. These surveys are expected to produce orders of magnitude more data than we currently have.

In order to efficiently analyze and take advantage of the large datasets, astronomers have turned to machine learning methods. Machine learning models work well for problems where the likelihoods are difficult to calculate in conjunction with large datasets for training and sufficient computing resources. Extracting information from images can fall into this category. For example, it is crucial for the cosmology goals of LSST to be able to estimate redshifts from images. While spectroscopic measurements are the most reliable way to measure redshifts, spectroscopy takes too much time, expense, and is not viable for the billions of galaxies needed for to measure sensitive cosmological parameters (Newman &

Gruen, 2022). However, because it is not clear how to best analytically compute the redshift from large numbers of images; machine learning offers us a data-driven method to solve this problem.

The most important ingredient to data-driven methods is quality data. If training data used to build models contain errors, resulting models will be unable to predict results reliably. Creating valid, replicable datasets are often the most time consuming facet of machine learning processes (Terrizzano et al., 2015). The quantity and quality of the training data has the greatest impact on the performance of machine learning models. A major source of biases in the models also arises from the choice of training data gleaned from the larger datset. For example, training data that is not representative of the subsequent use of the model can lead to inaccurate predictions.

In this chapter, we present a new publicly available dataset of galaxy images and galaxy properties specifically built for machine learning applications. The dataset has been used for photometric redshift estimation, but can be also used for other science goals as well. This dataset makes use of publicly available survey data for images, photometry, and spectroscopic redshifts. Additional processing was also performed to measure morphological information for each galaxy image. We choose to use the HyperSuprime-Cam Survey (Aihara et al., 2018) to provide a sample of galaxies that are closer to what LSST will be able to observe. While the requirement for spectroscopic redshifts limits this dataset to brighter galaxies, it has a larger redshift range and sample size than other publicly available training datasets. This dataset is optimized for machine learning models by careful consideration of outliers and missing data, and provides the images and tabular data in a format that is amenable to drop into modern machine learning frameworks easily.

By releasing this dataset, we hope to provide 1) a source of a large amount of training data, 2) a consistent dataset for model comparisons, 3) a reduction in barriers to entry for machine learning in cosmology, 4) a fixed dataset enabling reproducibility in findings.

In Section 2.2, we describe the data sources and the construction of the GalaxiesML

dataset. Section 2.3, we present the structure and properties of the dataset.

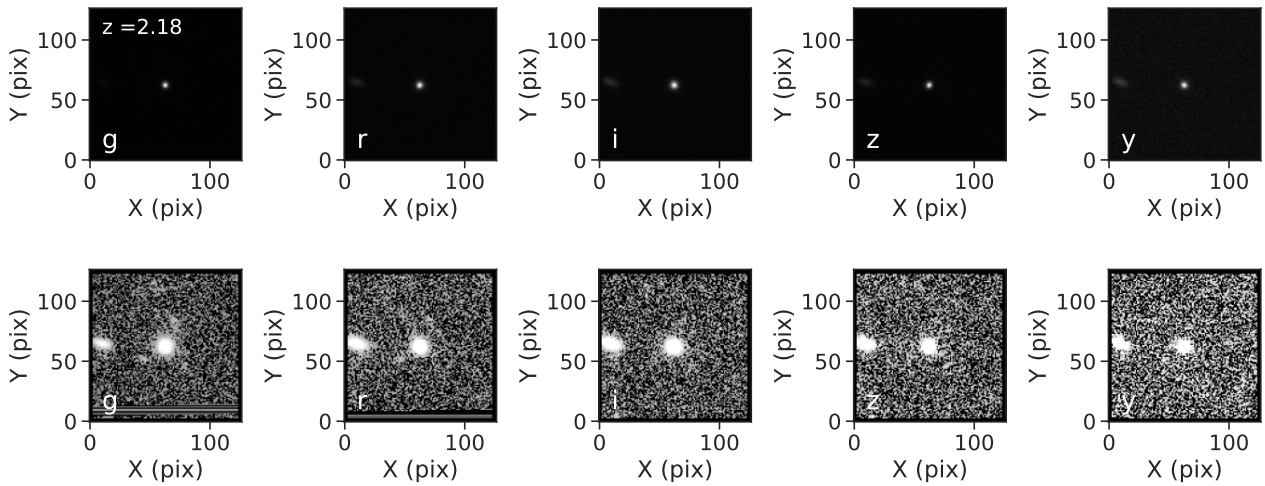

Figure 2.1 Example of a galaxy at $z = 2.14$ from the HSC Survey. The five filters ($grizy$) are in each column. The top row shows the image in a linear intensity scale while the bottom row shows the images in a logarithmic scale to show lower surface brightness features like other galaxies nearby.

2.2 Constructing the dataset

HSC is a wide-field optical camera with a FOV of 1.8 deg^2 on the Subaru Telescope. HSC PDR2 surveys more than 300 deg^2 in five optical filters ($grizy$). The median seeing in the i-band is 0.6".

The primary data sources for this work are from the HSC Survey Data Release 2 (Aihara et al., 2019) and the associated spectroscopic redshift database. The survey contains three layers: Wide, Deep, and UltraDeep. We use the PDR2-Wide layer which surveys more than 300 deg^2 in five optical filters ($grizy$). The HSC survey reaches a similar magnitude and depth as the 10-year LSST goal, which makes this survey a good precursor to train and test machine learning models. The HSC PDR2 database contains spectroscopic redshifts cross-matched by the team (within a projected distance of $< 0.5''$) to the HSC catalog using

publicly available spectroscopic redshift catalogs (Lilly et al., 2009; Bradshaw et al., 2013; McLure et al., 2012; Skelton et al., 2014; Momcheva et al., 2016; Le Fèvre et al., 2013; Garilli et al., 2014; Liske et al., 2015; Davis et al., 2003; Newman et al., 2013; Coil et al., 2011; Cool et al., 2013).

We assemble the dataset in 6 major stages:

1. Query and download from the HSC PDR2 and spectroscopic redshift databases
2. Apply additional data quality filters & remove duplicates and outliers
3. Download images and produce cutouts
4. Fit images to determine morphological information
5. Save the dataset into ML compatible formats

These stages are shown in the flow chart in Fig. 2.2 and outlined below.

2.2.1 Database Queries

We create a custom SQL query to select and download data from the HSC Survey PDR2 Archive (Aihara et al., 2019). The initial selection of galaxies is designed to include as many well-observed galaxies as possible. We use the following criteria for selecting galaxies from the PDR2 database:

- `grizy_cmodel_flux_flag = False` $z > 0$
- `grizy_pixelflags_edge` = False
- `grizy_pixelflags_interpolatedcenter` = False
- `grizy_pixelflags_saturatedcenter` = False, unique galaxy object ID
- `grizy_pixelflags_crcenter` = False,

- `grizy_pixelflags_bad` = False
- `grizy_sdsscentroid_flag` = False

For photometric redshift applications, the dataset needs a source of 'truth' for the redshift of each galaxy, so we also require that the objects have reliable spectroscopic redshift measurements by joining the tables with the spectroscopic redshift tables in the HSC database. We use the spectroscopic redshift information gathered by the HSC team as the ground truth for this sample. A match was made between the location of objects in the HSC survey with those from multiple spectroscopic surveys. In addition to joining with the photometry table, we also apply the following filters for the redshifts:

- $z > 0$
- $z \neq 9.9999$
- $0 < z_{err} < 1$
- `specz_flag_homogeneous = True`

We require that the galaxies be detected in all 5 imaging filters. The database query is reproduced in Appendix A. Overall, this initial sample has 801,246 objects.

2.2.2 Additional data quality filters & remove duplicates

We apply the following additional filters to the list of galaxies after extracting the objects from the database:

- Redshift range: $0.01 < z < 4.0$
- Spectroscopic redshift error: $\sigma_{specz} < 0.005/(1 + specz)$
- Magnitude range in all bands: $0 <$ {`band`}`_cmodel _mag` < 50

To build our final sample, we then remove duplicate objects from the sample. While each row has a unique object ID in the database, there can be multiple object IDs for the same physical source because there were multiple measurements of its photometry. About 70% of the entries do not refer to unique sources. We define duplicates as objects that have the same spectroscopic redshift identifier. Note that the spectroscopic redshifts were matched to the photometry using a distance of 0.5 arcseconds by the HSC team[1]. We also identify duplicate sources with the same HSC object ID, but different spectroscopic IDs. In the case of duplicates, we keep the first match and remove the others. After this stage, our final sample includes 286,401 sources.

2.2.3 Download and produce image cutouts

After obtaining the final sample of galaxies, we query the HSC PDR2 cutout service to download the images[2] (see 2.4.3). We submit queries at the RA and DEC for each band in batches of 100,000 galaxies at a time, with cutout sizes of $10'' \times 10''$. We download the `coadd` option for images and selected the `PDR2 Wide` option. The images are downloaded as FITS files.

2.2.4 Measurement of the Morphological Parameters

We also extracted typical morphological features from the galaxy images to aid in interpreting the images and models. The 127×127 pixel images were fit using Source Extractor (Bertin & Arnouts, 1996). Source Extractor is a tool that is often used to model galaxies in images. Source Extractor fits the pixel values of images using parameterized models of galaxies using an estimate of the point spread function. It can fit for multiple sources at once and produce a segmentation map that indicates which pixels belong to which source. We parameterized

[1] `https://hsc-release.mtk.nao.ac.jp/doc/index.php/dr1_specz/`

[2] `https://hsc-release.mtk.nao.ac.jp/das_cutout/pdr2/`

the morphology of the galaxies using several models:

- Elliptical model - fitting the sources as an ellipse with a semi-major axis, semi-minor axis, and orientation.
- Sersic model - fitting the flux distribution with a Sersic profile
- Isophotal, half-light, and Petrosian radius.

We also determine the number of galaxies in the image from the image segmentation map from Source Extractor. Source Extractor identifies sources using a detection threshold of pixel values above the background (called `DETECT_THRESH`), which we set to 3σ. We use the source position parameters `X_IMAGE` and `Y_IMAGE` closest to the center of the image as the galaxy that is associated with the spectroscopic redshift from the HSC catalog. We also utilize the position parameters to identify the number of other galaxies in a circle with a radius of 15, 10, and 5 pixels around the center of the image to quantify the number of nearby sources. The complete Source Extractor configuration file we use for each galaxy is reproduced in Appendix 2.4.3.

2.2.5 Save the dataset into ML compatible format

The imaging data is stored in the HDF5 file format. This file contains cutouts of each galaxy image in g, r, i, z, y stored in the `image` key of the HDF5 file as an $N_{gal} \times N_x \times N_y$ array. The decision to use HDF5 is due to its support for reading in only parts of the dataset at a time, and for its ease of use in machine learning frameworks. To create the HDF5, we download the images of each object in all five HSC imaging filters in the FITS format and combine the data into HDF5 files. The data are downloaded as cutouts from larger scale HSC images using the HSC cutout service[3]. We use an image radius query of `sh = 10arcseconds, sw = 10arcseconds`, which results in images of 20×20 arcseconds in spatial dimension. The

[3] `https://hsc-release.mtk.nao.ac.jp/das_cutout/pdr2/`

images have a plate scale of 0.168 arseconds per pixel (Aihara et al., 2019). We created two image sizes: 127x127 pix and 64x64 pix. Most machine learning methods require all images to have fixed sizes. Having multiple options for the image sizes allows one to test the effect of image sizes on model performance and to use wider variety of models. For each image size we create 3 HDF5 files for training (60%), validation (20%), and testing (20%) by randomly splitting the data. We perform the split for convenience and to more easily compare model performances with the same split.

2.3 Description of ML Dataset

The GalaxiesML dataset is a collection of tabular data, imaging data, and metadata. The tabular data is organized as a CSV file and also included in the HDF5 files with the images. In the HDF5, the images ($5 \times 127 \times 127$ or $5 \times 64 \times 64$) are under the `image` key, while the other tabular data are under the keys corresponding to their column name. The tabular data providing detailed information on the identification and characteristics of each galaxy sourced from the HSC and spectroscopic database. The tables also contain the extracted features including morphology.

The following columns are the list of extracted parameters using Source Extractor:

- {band}_central_image_pol_15px_rad: The number of detected objects within a 15-pixel-radius circle, centered at the middle of the image. Derived from Source Extractor segmentation file.
- {band}_central_image_pop_10px_rad: The number of detected objects within a 10-pixel-radius circle, centered at the middle of the image. Derived from Source Extractor segmentation file.
- {band}_central_image_pop_5px_rad: The number of detected objects within a 5-pixel-radius circle, centered at the middle of the image. Derived from Source Extractor

Table 2.1 GalaxiesML Column Definition - Galaxy Properties & Morphology Measurements

Column Name	Units	Description
object_id		object ID from the HSC survey. Unique ID in 64bit integer
coord	(deg, deg, deg)	Coordinate used in coneSearch(coord, RA, DEC, RADIUS)
ra	deg	RA (J2000.0) of the image center
dec	deg	DEC (J2000.0) of the image center
{band}_cmodel_mag	mag	- magnitude of the central galaxy in filter {band}
{band}_cmodel_magsigma	mag	uncertainty in the magnitude in filter {band}
skymap_id		location of the galaxy in internal survey position definition (tract, patch)
specz_name		name(s) of the galaxy in the spectroscopic survey(s)
specz_flag_homogeneous		Homogenized spec-z flag. (TRUE=secure, FALSE=insecure)
specz_mag_i	mag	i-band magnitude of the galaxy in the spectroscopic survey
specz_ra	deg	RA (J2000.0) of galaxy in spectroscopic survey
specz_dec	deg	DEC (J2000.0) of galaxy in spectroscopic survey
specz_redshift		spectroscopic redshift
specz_redshift_err		spectroscopic redshift uncertainty
{band}_central_image_pol_15px_rad		See Section 2.3
{band}_central_image_pop_10px_rad		See Section 2.3
{band}_central_image_pop_5px_rad		See Section 2.3
{band}_ellipticity	pixels	See Section 2.3
{band}_half_light_radius	pixels	See Section 2.3
{band}_isophotal_area	pixels	See Section 2.3
{band}_major_axis	pixels	See Section 2.3
{band}_minor_axis	pixels	See Section 2.3
{band}_peak_surface_brightness	mag/sq. arcsec	See Section 2.3
{band}_petro_rad	pixels	See Section 2.3
{band}_pos_angle	deg	See Section 2.3
{band}_sersic_index		See Section 2.3
{band}_total_galaxies		See Section 2.3

segmentation file.

- {band}_ellipticity: The ellipticity of the object, defined as 1 - B/A. Where B is the semi-minor axis of an object and A is the semi-major axi (pixels).
- {band}_half_light_radius: The radius of an object at which 50% of the flux is contained (pixels).
- {band}_isophotal_area: The total number of pixels of which a detected object is composed.
- {band}_major_axis: Major axis of the detected object (pixels).
- {band}_minor_axis: Minor axis of the detected object (pixels).
- {band}_peak_surface_brightness: The peak surface brightness above background of the object (magnitudes per square arcsecond).
- {band}_petro_rad: Petrosian radius of an object (pixels).
- {band}_pos_angle: Rotation of the major axis with respective to the x-axis of the image plane, counterclockwise (degrees).
- {band}_sersic_index: The Sérsic index of the object, which describes the shape of the object's light profile.
- {band}_total_galaxies: Total number of galaxies detected by Source Extractor in an image.

The data is available from Zenodo.

2.3.1 Properties of the dataset

In this section, we will describe properties of the dataset including: (1) the redshift distribution of objects, (2) the magnitude distribution, and (3) the color distribution. These properties are important factors to consider when training and developing machine learning models with these data. While we aim to build as large and comprehensive a dataset as possible with the HSC data, there are selection effects that can bias machine learning predictions. For example, predictions for galaxies at redshifts > 2.5 are challenging because of the relative lack of training data at those redshifts.

The redshift distribution is set by overlap between the HSC survey and spectroscopic redshift sample. The availability of spectroscopic redshifts is the limiting factor in the redshift distribution. The redshift distribution of the dataset has two main peaks at $z \sim 0.12$ and $z \sim 51$. The median of the redshift distribution is at ~ 0.49 and the 95% of the galaxies have $z < 1.8$. At redshifts $z < 1$, the colors of the galaxies tend to become redder with redshift (Fig. 2.4).

2.4 Appendix

2.4.1 HSC SQL Query

```
blueSELECT
    object_id
      , specz_redshift_err
      , specz_redshift
      , specz_mag_i
      , specz_name
      , specz_ra
      , specz_dec,
specz_flag_homogeneous,
```

```
  ra,
  bluedec,
  coord,
  skymap_id,
  g_cmodel_mag
      , r_cmodel_mag
          , i_cmodel_mag
              , z_cmodel_mag
                  , y_cmodel_mag


      , g_cmodel_magsigma
      , r_cmodel_magsigma
          , i_cmodel_magsigma
              , z_cmodel_magsigma
                  , y_cmodel_magsigma




blueFROM pdr2_wide.forced



blueLEFT blueJOIN pdr2_wide.forced2 blueUSING (object_id)
blueLEFT blueJOIN pdr2_wide.specz blueUSING (object_id)

    blueWHERE
```

```
blueNOT
g_sdsscentroid_flag
    blueAND blueNOT
r_sdsscentroid_flag
    blueAND blueNOT
i_sdsscentroid_flag
    blueAND blueNOT
z_sdsscentroid_flag
    blueAND blueNOT
y_sdsscentroid_flag

blueAND blueNOT
    g_pixelflags_interpolatedcenter
blueAND blueNOT
r_pixelflags_interpolatedcenter
blueAND blueNOT
i_pixelflags_interpolatedcenter
blueAND blueNOT
z_pixelflags_interpolatedcenter
blueAND blueNOT
y_pixelflags_interpolatedcenter
blueAND blueNOT
    g_pixelflags_saturatedcenter
blueAND blueNOT
r_pixelflags_saturatedcenter
blueAND blueNOT
i_pixelflags_saturatedcenter
blueAND blueNOT
z_pixelflags_saturatedcenter
```

```
blueAND blueNOT
y_pixelflags_saturatedcenter
blueAND blueNOT
g_cmodel_flag
blueAND blueNOT
r_cmodel_flag
blueAND blueNOT
i_cmodel_flag
blueAND blueNOT
z_cmodel_flag
blueAND blueNOT
y_cmodel_flag
blueAND blueNOT
g_pixelflags_edge
blueAND blueNOT
r_pixelflags_edge
blueAND blueNOT
i_pixelflags_edge
blueAND blueNOT
z_pixelflags_edge
blueAND blueNOT
y_pixelflags_edge
blueAND blueNOT
g_pixelflags_crcenter
blueAND blueNOT
r_pixelflags_crcenter
blueAND blueNOT
i_pixelflags_crcenter
blueAND blueNOT
```

```
    z_pixelflags_crcenter
    blueAND blueNOT
    y_pixelflags_crcenter
    blueAND blueNOT
    g_pixelflags_bad
    blueAND blueNOT
    r_pixelflags_bad
    blueAND blueNOT
    i_pixelflags_bad
    blueAND blueNOT
    z_pixelflags_bad
    blueAND blueNOT
    y_pixelflags_bad
    blueAND  specz_redshift > 0
    blueAND 0 < specz_redshift_err
    blueAND specz_redshift_err < 1
    blueAND specz_redshift < 9.999
```

2.4.2 HSC Image Query

Images were obtained by uploading the ra and dec positions in batches of 100,00 requests at a time to the HSC Image Cutout Service at: `https://hsc-release.mtk.nao.ac.jp/das_cutout/pdr2/manual.html#list-to-upload`. Here is a example of some of the lines from the positional query:

```
#? rerun    filter  ra  bluedec sw  sh
pdr2_wide   HSC-Y   31.73471487 -6.610750394    10arcseconds    10
   arcseconds
```

```
pdr2_wide   HSC-Y   31.19229445 -6.44807734 10arcseconds    10
   arcseconds
pdr2_wide   HSC-Y   31.20098659 -6.878016245    10arcseconds    10
   arcseconds
pdr2_wide   HSC-Y   31.09620369 -6.250176133    10arcseconds    10
   arcseconds
pdr2_wide   HSC-Y   31.16933934 -6.523789799    10arcseconds    10
   arcseconds
```

2.4.3 Source Extractor Configuration

We use the following parameter file running Source Extractor on each image:

```
NUMBER      1                   #Running object number
DETECT_THRESH 3
CATALOG_NAME output_image
CATALOG_TYPE ASCII_HEAD
CHECKIMAGE_TYPE   SEGMENTATION
PETRO_RADIUS 1
PETRO_TYPE AUTO
X_IMAGE 60
Y_IMAGE 60
XMIN_IMAGE 57.5
XMAX_IMAGE 62.5
YMIN_IMAGE 57.5
YMAX_IMAGE 62.5
ISOAREA_IMAGE 0.0
```

ISOAREA_WORLD 0.0

A 0.0

B 0.0

A_IMAGE

B_IMAGE

THETA_IMAGE 0.0

THETA_WORLD 0.0

MU_MAX 0.0

ELLIPTICITY 0.0

SERSIC 0.0

PHOT_TYPE SERSIC

PHOT_AUTOPARAMS 2.5, 3.5

PHOT_APERTURES 5

FLUX_RADIUS

SPHEROID_SERSICN 4.0

SPHEROID_RE 10.0

We use the following configuration file running Source Extractor on each image:

```
\begin{lstlisting}[basicstyle=\small]
# Default configuration file for SExtractor 2.25.0
# EB 2021-05-31
NUMBER      1
PETRO_TYPE AUTO
PETRO_THRESH 0.2
#------------------------------- Catalog
```

```
#------------------------------------

CATALOG_NAME     test.cat       # name of the output catalog
CATALOG_TYPE     ASCII_HEAD     # NONE,ASCII,ASCII_HEAD,
    ASCII_SKYCAT,
                                # ASCII_VOTABLE, FITS_1.0 or
                                    FITS_LDAC
PARAMETERS_NAME  default.param  # name of the file containing
    catalog contents

#-------------------------------- Extraction
    ----------------------------------

DETECT_TYPE      CCD            # CCD (linear) or PHOTO (with
    gamma correction)
DETECT_MINAREA   5              # min. # of pixels above threshold
DETECT_THRESH    1.5            # <sigmas> or <threshold>,<ZP> in
    mag.arcsec-2
ANALYSIS_THRESH  1.5            # <sigmas> or <threshold>,<ZP> in
    mag.arcsec-2

FILTER           Y              # apply filter for detection (Y or
     N)?
FILTER_NAME      gauss_1.5_3x3.conv   # name of the file
    containing the filter

DEBLEND_NTHRESH  32             # Number of deblending sub-
```

```
    thresholds
DEBLEND_MINCONT  0.005          # Minimum contrast parameter for
    deblending

CLEAN            Y              # Clean spurious detections? (Y or
    N)?
CLEAN_PARAM      1.0            # Cleaning efficiency

MASK_TYPE        CORRECT        # type of detection MASKing: can
    be one of
                                # NONE, BLANK or CORRECT

#------------------------------- Photometry
    ----------------------------------

PHOT_APERTURES   5              # MAG_APER aperture diameter(s) in
     pixels
PHOT_AUTOPARAMS  2.5, 3.5       # MAG_AUTO parameters:<Kron_fact
    >,<min_radius>
PHOT_PETROPARAMS 2.0, 3.5       # MAG_PETRO parameters:<
    Petrosian_fact>,
                                # <min_radius>

SATUR_LEVEL      50000.0        # level (in ADUs) at which arises
    saturation
SATUR_KEY        SATURATE       # keyword for saturation level (in
     ADUs)
```

```
MAG_ZEROPOINT      0.0             # magnitude zero-point
MAG_GAMMA          4.0             # gamma of emulsion (for
    photographic scans)
GAIN               0.0             # detector gain in e-/ADU
GAIN_KEY           GAIN            # keyword for detector gain in e-/
    ADU
PIXEL_SCALE        0.17            # size of pixel in arcsec(0=use
    FITS WCS info)

PHOT_TYPE SERSIC
SERSIC_FIT Y
FIT_PROFILE N
#------------------------------ Star/Galaxy Separation
    -------------------------------

SEEING_FWHM        1.2             # stellar FWHM in arcsec
STARNNW_NAME       default.nnw     # Neural-Network_Weight table
    filename

#------------------------------------ Background
    -----------------------------------

BACK_SIZE          64              # Background mesh: <size> or <
    width>,<height>
BACK_FILTERSIZE    3               # Background filter: <size> or <
    width>,<height>
```

```
BACKPHOTO_TYPE     GLOBAL           # can be GLOBAL or LOCAL

#------------------------------- Check Image
    ----------------------------------

#CHECKIMAGE_TYPE   -BACKGROUND     # can be NONE, BACKGROUND,
    BACKGROUND_RMS,
                                   # MINIBACKGROUND, MINIBACK_RMS, -
                                       BACKGROUND,
                                   # FILTERED, OBJECTS, -OBJECTS,
                                       SEGMENTATION,
                                   # or APERTURES
#CHECKIMAGE_NAME   check.fits      # Filename for the check-image

#--------------------- Memory (change with caution!)
    ------------------------------

MEMORY_OBJSTACK    3000            # number of objects in stack
MEMORY_PIXSTACK    300000          # number of pixels in stack
MEMORY_BUFSIZE     1024            # number of lines in buffer

#------------------------------- Miscellaneous
    ----------------------------------

VERBOSE_TYPE       NORMAL          # can be QUIET, NORMAL or FULL
HEADER_SUFFIX      .head           # Filename extension for
```

```
   additional headers
WRITE_XML          N                  # Write XML file (Y/N)?
XML_NAME           sex.xml            # Filename for XML output
DETECT_THRESH 3
NUMBER     1                          #Running object number
DETECT_THRESH 3
CATALOG_NAME output_image
CATALOG_TYPE ASCII_HEAD
CHECKIMAGE_TYPE   SEGMENTATION
ERRXY_IMAGE 5
```

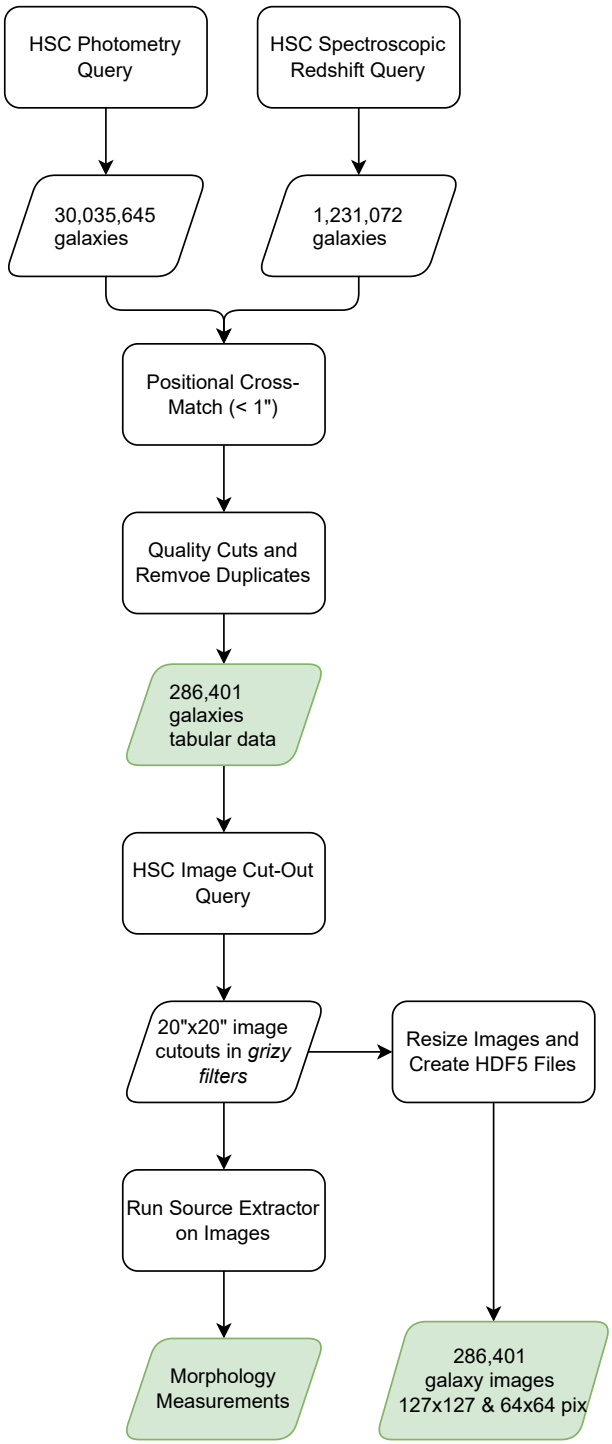

Figure 2.2 Flow chart showing the steps used in creating the GalaxiesML dataset. Rectangles represent processes and parallelograms are the products. The green parallelograms are the datasets that are part of the release.

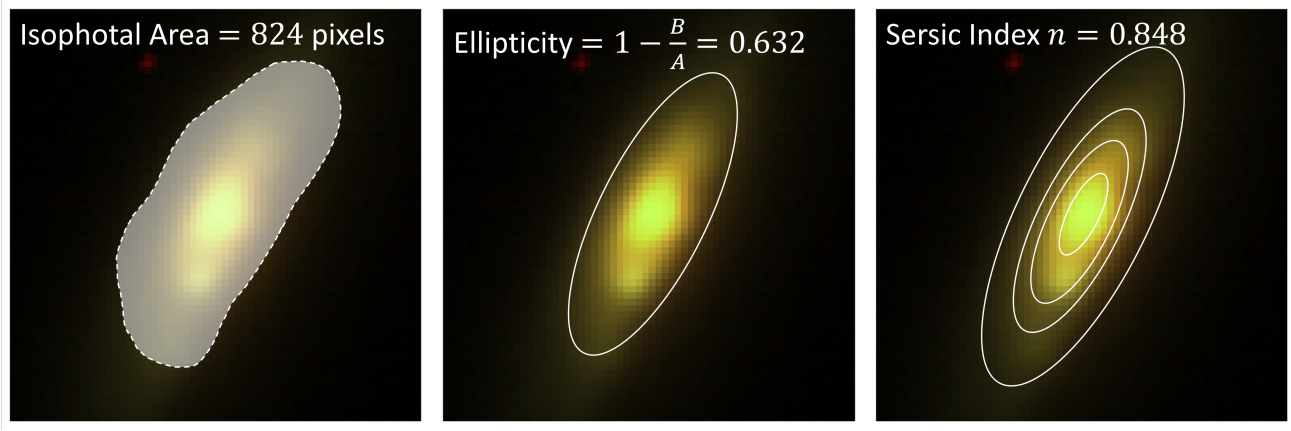

Figure 2.3 Example of the morphological parameters measured on a low redshift galaxy (Object ID 36416246018753893, $z = 0.0713$) using Source Extractor. **Left:** isophototal area, **center:** ellipticity, **right:** Sersic Index.

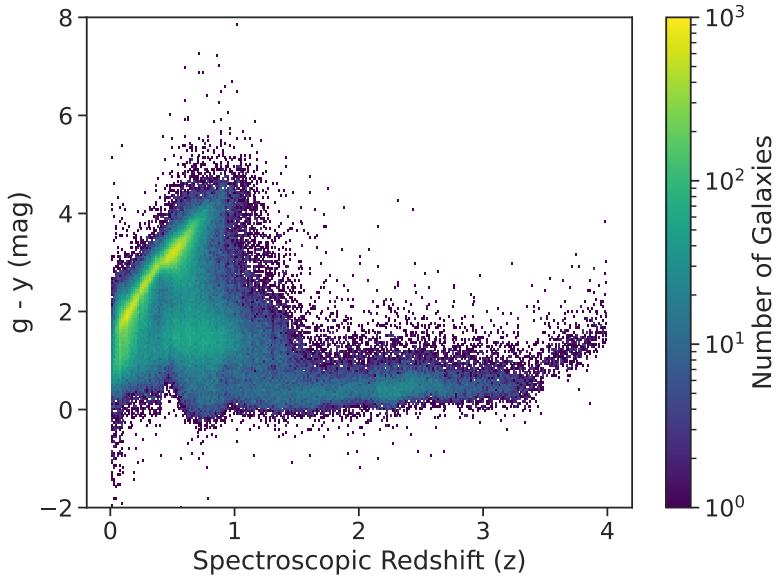

Figure 2.4 2D histogram of the distribution of *g*-*y* color as a function of redshift. The colorbar shows the number of galaxies in that bin. The galaxies grow redder as a function until about redshift $z > 1$, where the *g*-*y* color is more constant with redshift.

CHAPTER 3

Photometric Redshift Estimation with Galaxy Photometry

This thesis chapter originally appeared in the literature as Improving Photometric Redshift Estimation for Cosmology with LSST Using Bayesian Neural Networks

Evan Jones, Tuan Do, Bernie Boscoe, Jack Singal, Yujie Wan, and Zooey Nguyen, *The Astrophysical Journal, Volume 964, Number 2*

3.1 Abstract

We present results exploring the role that probabilistic deep learning models can play in cosmology from large-scale astronomical surveys through photometric redshift (photo-z) estimation. Photo-z uncertainty estimates are critical for the science goals of upcoming large-scale surveys such as LSST, however common machine learning methods typically provide only point estimates and lack uncertainties on predictions. We turn to Bayesian neural networks (BNNs) as a promising way to provide accurate predictions of redshift values with uncertainty estimates. We have compiled a galaxy data set from the Hyper Suprime-Cam Survey with grizy photometry, which is designed to be a smaller scale version of large surveys like LSST. We use this data set to investigate the performance of a neural network (NN) and a probabilistic BNN for photo-z estimation and evaluate their performance with respect to LSST photo-z science requirements. We also examine the utility of photo-z uncertainties

as a means to reduce catastrophic outlier estimates. The BNN outputs the estimate in the form of a Gaussian probability distribution. We use the mean and standard deviation as the redshift estimate and uncertainty. We find that the BNN can produce accurate uncertainties. Using a coverage test, we find excellent agreement with expectation – 67.2% of galaxies between $0 < 2.5$ have 1-σ uncertainties that cover the spectroscopic value. We also include a comparison to alternative machine learning models using the same data. We find the BNN meets two out of three of the LSST photo-z science requirements in the range $0 < z < 2.5$.

3.2 Introduction

Cosmological probes of dark matter and dark energy aim to measure the structure and evolution of the universe, and thus rely in part on accurately and precisely measuring galaxy redshifts of hundreds of millions of galaxies with well-constrained uncertainties. Obtaining accurate photometric redshift estimates and with well-constrained uncertainties is a major challenge. Spectroscopic redshift measurements are the most reliable method of obtaining redshift, however they are a time consuming measurement and therefore cannot be used on large scales. Unlike precise spectroscopic redshift measurements, photometric redshift estimation is subject to significant systematic errors that need to be minimized. Science goals such as using weak lensing cosmological probes are strongly affected by the number of photo-z outliers — those objects whose estimated photo-zs are far from the actual redshifts (e.g. Hearin et al. (2010); Jones et al. (2021b); Singal et al. (2022); Newman & Gruen (2022))

According to the LSST Science Requirements Document (SRD)[1], sufficiently accurate photo-z estimates for $\sim$ four billion galaxies are required to meet the LSST science goals for their main cosmological sample. Specifically, for the $i < 25$ flux-limited galaxy sample measured by LSST, one must achieve

[1]https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17

- number of galaxies $\approx 10^7$
- rms error < 0.2 (Equation 3 in Table 2)
- bias < 0.003 (Equation 4 in Table 2)
- 3σ catastrophic outliers $< 10\%$ total sample (Equation 2 in Table 2)

Currently, no published model satisfies the LSST photo-z science requirements up to $z = 3$ (Tanaka et al., 2018a; Schuldt et al., 2020b; Schmidt et al., 2020a). Additionally, methods for rejecting the majority of outliers and characterizing their effects on the predictions must be developed (Ivezic, 2018). Beyond the LSST metrics stated in the SRD, we consider additional probabilistic metrics for quantifying the quality of uncertainty estimates (see Table 2 – Malz & Hogg, 2020; Schmidt et al., 2020a; Jones et al., 2022a). The requirement thresholds for the probabilistic metrics are not as well quantified at this time as those for point metrics, but they allow us to compare the performance between different probabilistic models evaluated on the same data. Techniques for identifying photo-z outlier predictions in machine learning models have been investigated in Jones & Singal (2020); Wyatt & Singal (2020), and Singal et al. (2022).

Photo-z estimation techniques have traditionally been divided into two main approaches. Template fitting methods, such as Lephare (Ilbert et al., 2006; Arnouts et al., 1999), Mizuki (Nishizawa et al., 2020), and Bayesian Photometric Redshift (BPZ – Benítez, 2000), involve correlating the observed band photometry with model galaxy spectra and redshift, and possibly other model properties. Machine learning methods, such as artificial neural networks (e.g. ANNZ – Collister & Lahav, 2004), boosted decision trees (e.g. ARBORz – Gerdes et al., 2010), regression trees / random forests (Carrasco Kind & Brunner, 2013), support vector machines (SVMs – e.g. Wadadekar, 2005; Jones & Singal, 2017, 2020)), a Direct Empirical Photometric method (DEmP – Tanaka et al., 2018b), and others develop a mapping from input parameters to redshift with a training set of data in which the actual spectroscopic

redshifts are known, then apply the mappings to data for which the redshifts are to be estimated. Both have their drawbacks – template fitting methods require assumptions about intrinsic galaxy spectra or their redshift evolution, and empirical methods require the training set and evaluation set to significantly overlap in parameter space. As machine learning approaches for photo-z estimation have increased in capability and larger data sets have been observed over the past decade, galaxy images can be effectively used as inputs to utilize morphological information for photo-z estimation, unlike template-fitting approaches.

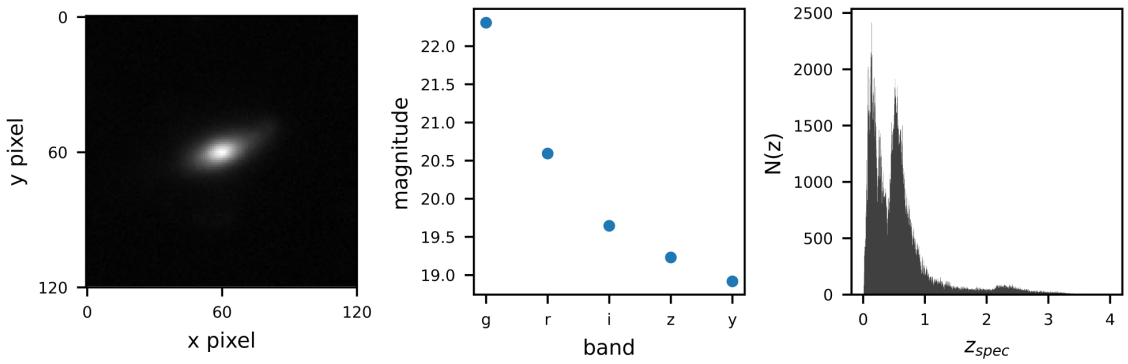

Figure 3.1 Left: Example of a galaxy ($z = 0.48$) image in the i-band. Middle: five-band photometry for the same galaxy. Right: $N(z)$ distribution for the data set discussed in §4.3.1 For the photo-z determinations in this work we use training, validation, and testing sets consisting of 229,120, 28,640, and 28,640 galaxies respectively.

There have been a number of works studying the application of neural networks for photo-z estimation (Firth et al., 2003; Collister & Lahav, 2004; Singal et al., 2011), while probabilistic NN techniques have had limited investigations until recently (Sadeh et al., 2016; Pasquet et al., 2019; Schuldt et al., 2020b; Zhou et al., 2022). Bayesian neural networks , a type of probabilistic NN (Jospin et al., 2020) are a promising approach that has not been well explored. Probabilistic neural networks, conceptualized in the 1990s (Specht, 1990), have previously been limited in their ability to process the size of data required for performing photo-z estimation for large-scale surveys, because of the complexity of their computation. However, recent breakthroughs in conceptual understanding and computational capabilities (e.g. Filos et al., 2019; Dusenberry et al., 2020) now make probabilistic deep learning possible

for cosmology. Probabilistic deep learning with a BNN has many advantages compared to traditional neural networks, including better uncertainty representations, better point predictions, and offers better interpretability of neural networks because they can be viewed through the lens of probability theory. In this way we can draw upon decades of development in Bayesian inference analyses.

We have three goals in this work: (1) develop a probabilistic ML model that can produce robust uncertainties for photometric redshifts, (2) assess the model with respect to LSST requirements and alternative photo-z estimation methods, and (3) investigate the use of photo-z uncertainties to identify likely outliers in photometric redshift predictions. For the analysis in this work, we have created the largest publicly available machine-learning-ready galaxy image data of $\sim 300k$ galaxies from the Hyper Suprime-Cam survey containing five-band photometric images and known spectroscopic redshifts from $0 < z < 4$. This data will be released in Do et al. 2024 (in prep). In §2 we discuss the data and network architecture. In §3 we state the results. In §4 and §5 we provide a discussion and conclusion.

3.3 Data and Methods

3.3.1 Data: Galaxy observations

For the analysis in this work we compile a data set intended to approximate the data produced by future large-scale deep surveys for photo-z estimation (Collaboration et al., 2021). We use the Hyper-Suprime Cam (HSC) Public Data Release 2 (Aihara et al., 2019), which is designed to reach similar depths as LSST but over a smaller portion of the sky. We choose the HSC survey because it mimics LSST in photometry and depth. Including photometry in infrared bands would improve photo-z estimates, but since LSST will provide observations in only optical bands (Ivezić et al., 2008), we will restrict our analysis to optical bands only. HSC is a wide-field optical camera with a FOV of 1.8 deg^2 on the Subaru Telescope. HSC PDR2 surveys more than 300 deg^2 in five optical filters ($grizy$). The median seeing in the

i-band is 0.6". This data set is presented in more detail in Do et al. 2024, in prep.

The final data set used in the analyses of this paper consists of $\sim$ 300k galaxies with 5-band grizy photometry and spectroscopic redshifts. Fig. 1 contains the N(z) distribution for the dataset and Fig. 2 contains grizy images for three example HSC galaxies. Spectro-zs were obtained by crossmatching galaxy photometry from HSC with the HSC collection of publicly available spectroscopic redshifts using galaxy sky positions (d $< 1''$) in Lilly et al. (2009), Bradshaw et al. (2013), McLure et al. (2012), Skelton et al. (2014), Momcheva et al. (2016), Le Fèvre et al. (2013), Garilli et al. (2014), Liske et al. (2015) Davis et al. (2003), Newman et al. (2013), Coil et al. (2011), Cool et al. (2013). We used data quality cuts similar to Nishizawa et al. (2020) and Schuldt et al. (2021) (see Table 1 and Do et al. 2024 (in prep) for a full list), which are intended to remove outlier photometric measurements and poorly measured spectroscopic redshifts. We also required detections in each band. The spectroscopic redshift values are treated as the ground truth for training and evaluation. In total, the data consists of 286,401 galaxies with broad-band grizy photometry and known spectroscopic redshifts. Our galaxy sample extends from $0.01 < z < 4$, however the majority of the sample lies between redshift of 0.01 and 2.5 with peaks at z 0.3 and z 0.6 (see N(z) in Fig. A.2). We use 80% of the galaxies for training, 10% for validation, and 10% for testing. The data used for training is available[2] from (Jones et al., 2021a). This dataset includes the photometry and spectroscopic redshifts. A future release will also include images (Do et al. in prep.).

3.3.2 Network architectures

We built two neural networks for this work – one is a fully connected neural network that produces single-valued redshift predictions and one is a Bayesian neural network that outputs Gaussian probability distributions. The NN and BNN models are visualized in Fig. 3.3. Both

[2] `https://zenodo.org/records/5528827`

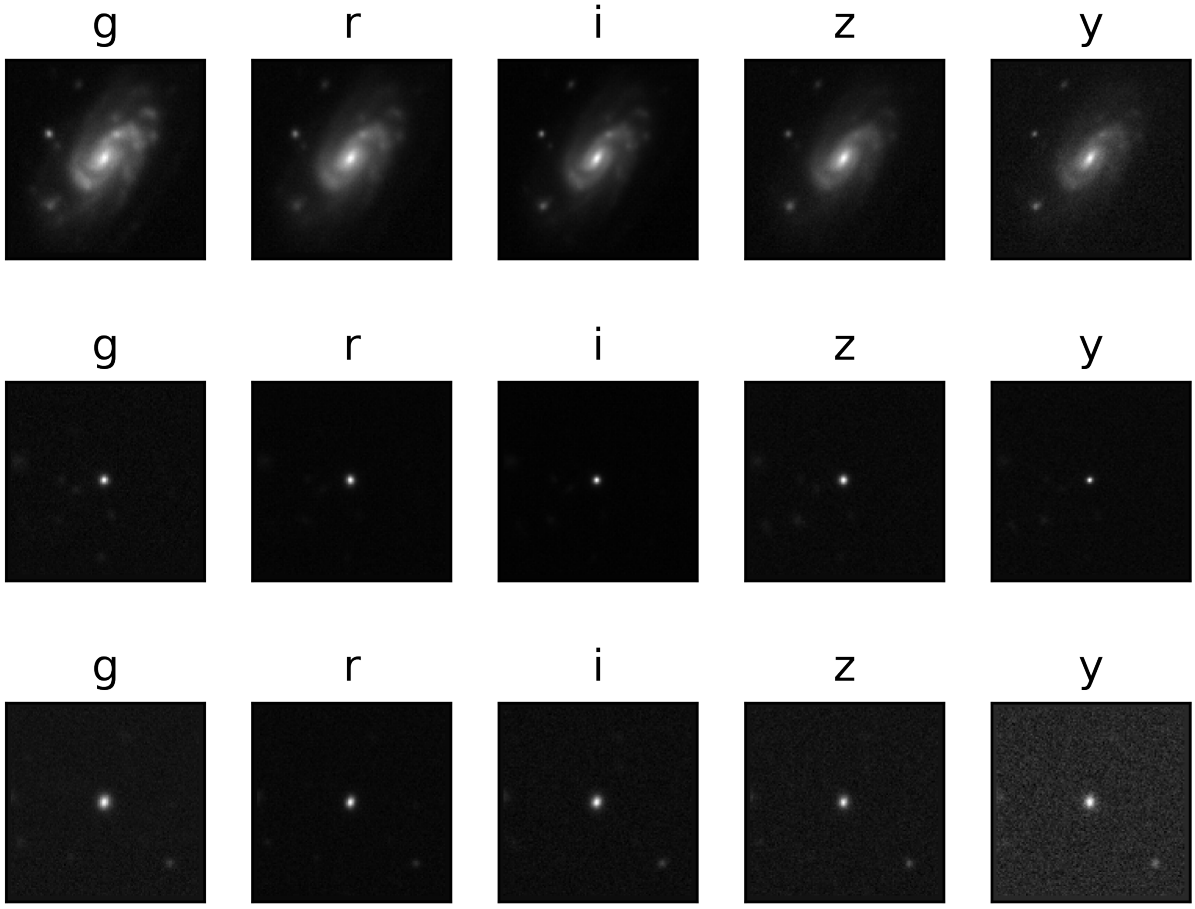

Figure 3.2 Example HSC galaxy images for the data set used in this work with *grizy* photometry for a low redshift galaxy at $z = 0.05$ (TOP), and a high redshift galaxy at $z = 3.92$ (MIDDLE), and another low redshift galaxy at $z = 0.14$ (BOTTOM). The similarity between the high redshift galaxy and the bottom low redshift galaxy highlights the difficulty of photo-z estimation.

the NN and the BNN are implemented in TensorFlow (Abadi et al., 2016) and have five input nodes for the five-band *grizy* photometry. We performed a parameter grid search to optimize for free parameters, such as the number of epochs, number of layers, number of nodes per layer, learning rate, loss function, activation function, and optimizer. Both the NN and BNN used for the final analysis in this work contain four hidden layers with 200 nodes per layer and utilize a rectified linear activation function. The networks also have a skip connection

Table 3.1 Quality cuts used to construct the data set.

photometry cuts	z_{spec} cuts
grizy_ cmodel_flux_flag = False	$z > 0$
grizy_ pixelflags_edge = False	$z \neq 9.9999$
grizy_ pixelflags_interpolatedcenter = False	$0 < z_{err} < 1$
grizy_ pixelflags_saturatedcenter = False	unique galaxy object ID
grizy_ pixelflags_crcenter = False	specz_flag_homogeneous = True
grizy_ pixelflags_bad = False	
grizy_ sdsscentroid_flag = False	

between the input nodes and the final layer. The NN has an output node to produce a single point estimate photo-z prediction while the BNN has a final output node that produces a mean and standard deviation assuming a Gaussian distribution for each photo-z prediction. For the BNN we use a negative log likelihood loss function with RMS error as the metric. We choose the negative log-likelihood loss function for the BNN because it has been shown to be more effective than MAE for probabilistic NNs (Lakshminarayanan et al., 2016). The NN uses a mean absolute error loss function, and we also consider a custom loss function (Nishizawa et al., 2020) defined in equation 6 of Table 3.2. The NN and BNN use the Adam optimizer and have learning rates of 0.0005 and 0.001, respectively. We train using an AMD Ryzen Threadripper PRO 3955WX with 16-Cores and NVIDIA RTX A6000 GPU. Training and evaluation runtimes are typically under 30 minutes.

3.3.3 Other ML models

We use three other common ML models in order to compare to the neural network performance: (1) a support SVM classification model, (2) a random forest regression (RF) model, and (3) a gradient boosted tree regression model. For RF models we utilize the Scikit-Learn implementation (Pedregosa et al., 2011) [3]. For the SVM we use SPIDERz (Jones & Singal, 2017, 2020), which implements support vector classification on classes of redshift bins of

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

width $z = 0.1$ spanning $0 < z < 4$. The RF model uses the RandomForestRegressor package to produce photo-z estimates. We also use the default hyperparameters with the RF, with the exception of using 200 trees in the forest. We use the XGboost software package for the gradient boosted tree model (the XGBRegressor library) with default hyperparameters. We perform a broad hyperparameter grid search for each model, however the performance boost over default parameters is not significant.

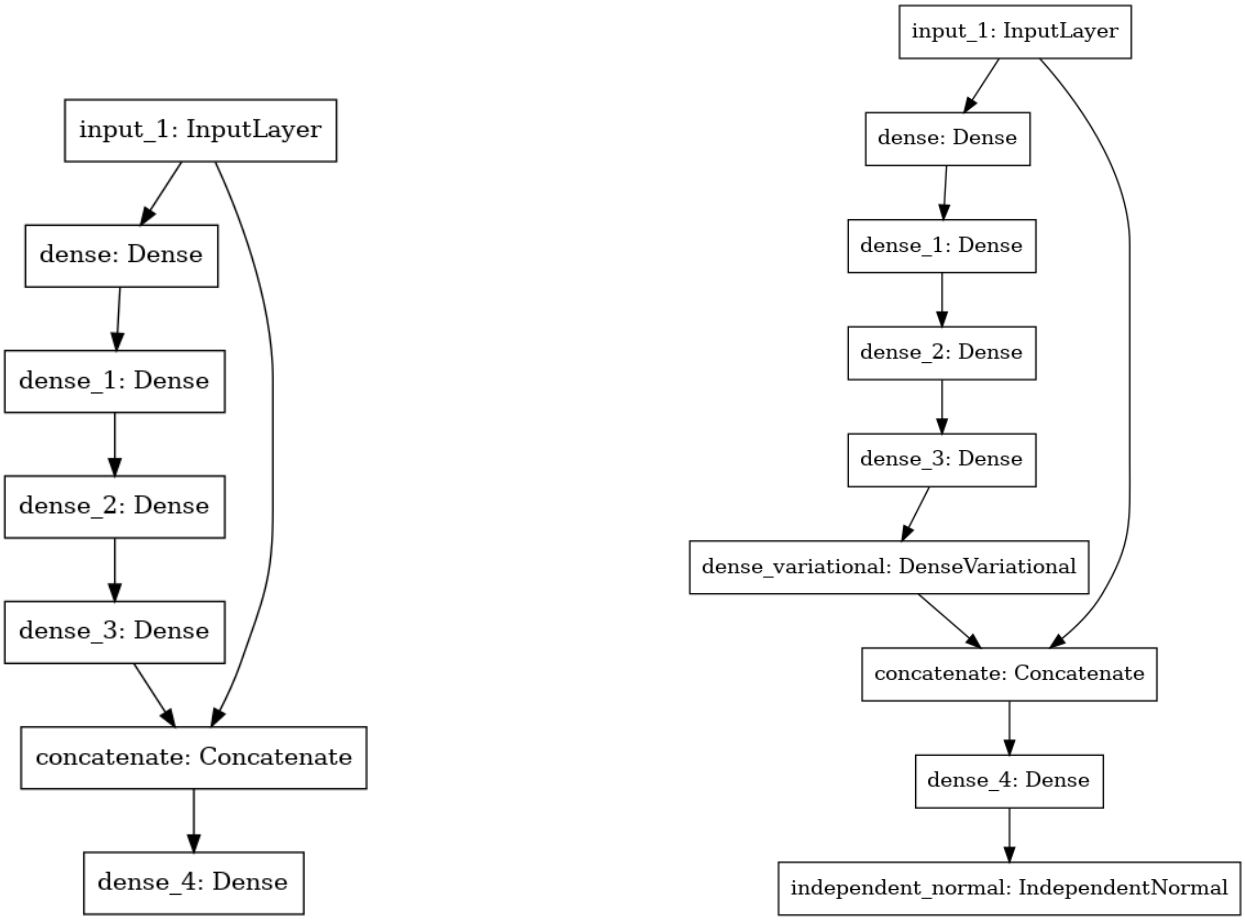

Figure 3.3 Left: NN architecture. Right: BNN architecture. The inputs for both networks are five-band photometry in the g,r,i,z,y filters. The output for the NN is a single point photo-z estimate while the output for the BNN is a photo-z PDF, which we sample to obtain a photo-z estimate. We assume Gaussianity in the creation of the photo-z PDF, so a photo-z uncertainty is produced by the standard deviation of the PDF.

3.4 Photo-z metrics

Photo-z uncertainties are propagated to measurement uncertainties on dark matter and dark energy. Therefore, our choice of metrics to evaluate the photo-z determinations in this work include chiefly the photo-z metrics used in the LSST science requirements document (RMS error (Eq. 3), Bias (Eq. 4), and 3σ Outliers (Eq. 7)) which are calculated to provide the necessary precision to constrain important cosmological quantities. Specifically, for the purpose of constraining dark matter and dark energy we require photo-z RMS error (< 0.2), Bias (< 0.003), and 3σ Outliers ($< 10\%$). In addition, we also include in our analysis a number of point metrics that are commonly used in the photo-z literature (Outlier (Eq. 1), Catastrophic Outlier (Eq. 2), Scatter (Eq. 5), and Loss (Eq. 6)) for the purpose of comparison to other models, as well as additional probabilistic metrics to evaluate the photo-z uncertainties produced by the BNN. To measure model performance we evaluate predictions using the metrics in Table 2, which are separated into non-probabilistic and probabilistic categories. These metrics describe different ways to characterize the photometric redshift performance averaged over all predictions. Ideally, photo-z measurements should be accurate out to the redshift limit of LSST observations (z = 3.4 is where galaxies begin dropping out of the g band), however the main redshift range of focus is $0.3 < z < 3.0$. In this redshift range, LSST aims to measure the comoving distance as a function of redshift to an accuracy of 1-2%. In order to achieve this goal, LSST must obtain (1) a sufficiently large sample of galaxies ($\sim$ four billion) and (2) sufficiently accurate photo-z measurements for these galaxies as defined by the aforementioned requirements. In addition to meeting photo-z science requirements, the LSST team also requires 'methods for rejecting the majority of those outliers, and for characterizing their effects on the sample'.

We note that science missions of LSST and Euclid for which photo-zs are necessary divide the redshift ranges of interest into several discrete tomographic redshift bins ($0 < z < 1.5$ divided into four bins of $z = 0.3$ in weak lensing analyses). The photo-z science requirements

must be achieved on average throughout each tomographic redshift bin, rather than on average throughout the entire sample. This means that a full evaluation of a particular photo-z method must include an evaluation of important metrics as a function of redshift, rather than averaging across the entire photo-z sample. This distinction is particularly important for evaluating model performance of high redshift regions ($z > 1.0$), which contain significantly fewer galaxies than low redshift regions (see Fig. A.2), and are thus more challenging for any photo-z method to accurately produce photo-zs.

3.4.0.1 Point Metrics

We use the conventional definition for photometric redshift outliers and catastrophic outliers in Eqs. 1 and 2, where z_{phot} and z_{spec} are the estimated photo-z and actual (spectroscopically determined) redshift of the galaxy. The RMS photo-z error is given by a standard definition in Eq. 3, where n_{gals} is the number of galaxies in the evaluation testing set and Σ_{gals} represents a sum over those galaxies. Bias and scatter are defined in Eqs. 4 and 5. We follow Tanaka et al. (2018b) and define a loss function in Eq. 6 to characterize the point estimate photo-z accuracy with a single number, where we use $\gamma = 0.15$.

3.4.0.2 Probability metrics

We propose coverage as a key metric for assessing the performance of the BNN (see Eq. 8). Coverage is typically used to assess whether confidence intervals are accurate. In this case, we define coverage as the fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If the coverage is over 68%, then the estimated uncertainties are on average too large. Similarly, if the cover is below 68%, the estimated uncertainties are on average too small.

Error in the bulk photo-z distribution width for the evaluation set can be difficult to

Table 3.2 Metrics used to assess model performance.

Point Metrics		Probabilistic Metrics	
Outlier	$O : \frac{\lvert z_{phot}-z_{spec}\rvert}{1+z_{spec}} > .15$ (1)	3σ Outlier:	$\lvert z_{phot} - z_{spec}\rvert > 3z_\sigma$ (7)
Catastrophic Outlier	$O_c : \lvert z_{phot} - z_{spec}\rvert > 1.0$ (2)		
RMS error	$\sqrt{\frac{1}{n_{gals}}\Sigma_{gals}\left(\frac{z_{phot}-z_{spec}}{1+z_{spec}}\right)^2}$ (3)	Coverage	$\sum_i^{n_{gals}} \frac{(\bar{z}_{pdf,i} - z_{spec,i}) < z_{\sigma,i}}{n_{gals}}$ (8)
Bias	$b = \frac{z_{phot}-z_{spec}}{1+z_{spec}}$ (4)		
Scatter	$\text{Median}(\lvert\Delta z - \text{Median}(\Delta z_i)\rvert)$ (5)	PIT:	$\int_{-\infty}^{z_{spec}} p(z)dz$ (9)
Loss	$L(\Delta z) = 1 - \frac{1}{1+(\frac{\Delta z}{\gamma})^2}$ (6)		

distinguish between uncertainties associated with galaxy bias or uncertainties in the mean redshift of photo-z tomographic bins. The Probability Integral Transform (PIT) is a photo-z metric that can detect systematic error in the photo-z distribution width for galaxy samples with known spectroscopic redshifts (Malz & Hogg, 2020; Malz, 2021). The PIT value for a single galaxy is defined in Eq. 9 in Table 3.2, where $p(z)$ is the predicted photo-z PDF. A histogram of PIT values for a galaxy sample should be uniform for an accurate collection of $p(z)$ samples. Ideally, the PIT histogram is flat across all redshift bins. If the PIT histogram peaks at the center, the $p(z)$ collection is too broad. If the PIT histogram peaks at high and low PIT values, the $p(z)$ samples are too narrow. For a comparison of several probabilistic photo-z methods, see Schmidt et al. (2020b).

3.4.1 Leveraging BNN for Outlier Identification

We propose a method for utilizing the photo-z uncertainties z_σ produced by the BNN to preemptively flag photo-z predictions with high uncertainties as potential poor predictions. The method is simple: all galaxies with a photo-z uncertainty greater than the specified z_σ cutoff value are flagged as potential outlier or catastrophic outlier candidates and removed from the evaluation sample. Figs. 3.4, 3.5, and 3.6 depict performance improvements with example σ_z removal values for a variety of performance metrics including the LSST photo-z requirements. An acceptable balance needs to be achieved between the number of galaxies

correctly flagged as poor predictions versus the number of non-outlier galaxies removed for a given z_σ cutoff value. Other outlier removal strategies have previously been explored in Jones & Singal (2020), Wyatt & Singal (2020), and Singal et al. (2022).

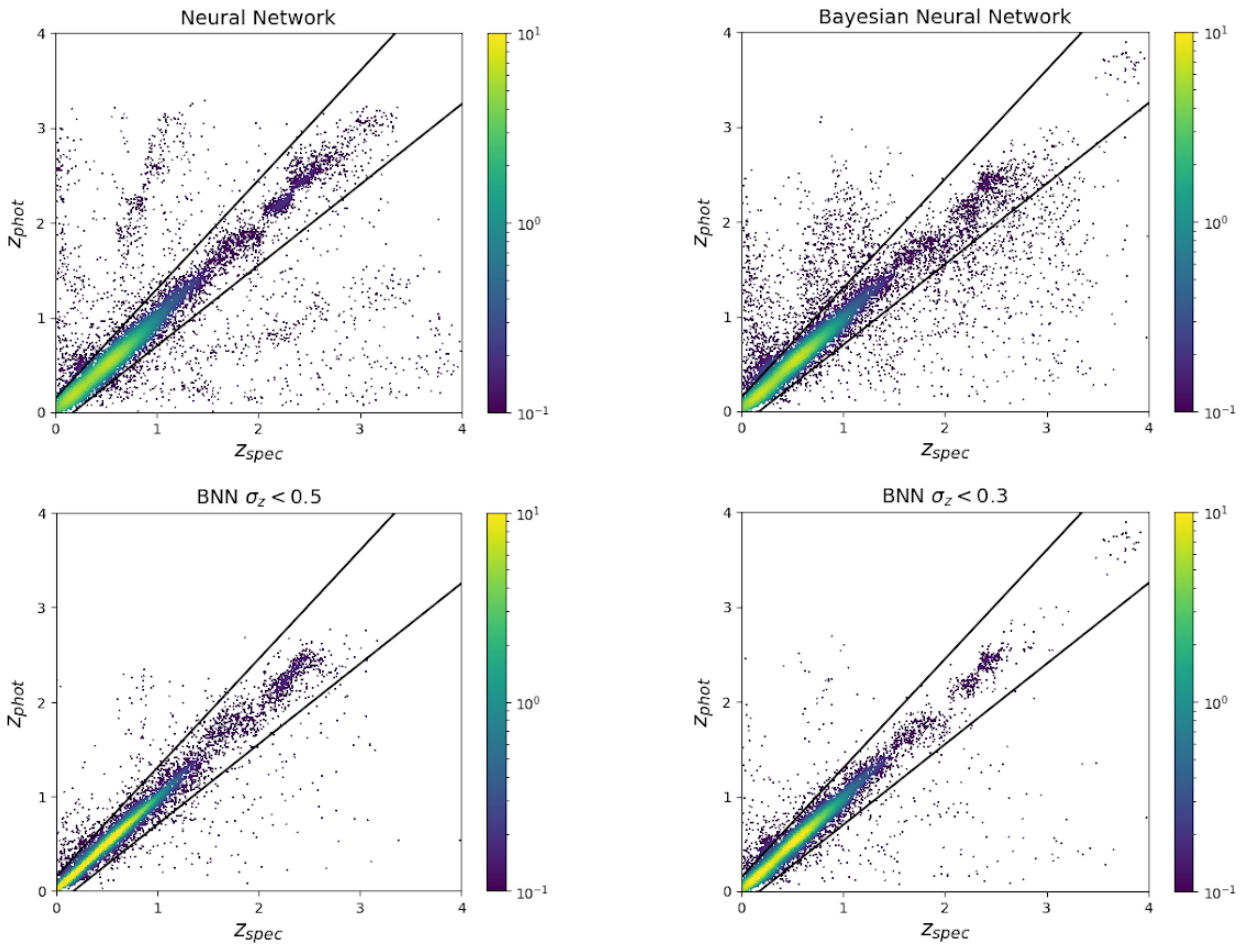

Figure 3.4 Visualization of NN (top left) and BNN (top right) performance compared to the BNN with outlier removal criteria examples $\sigma_z = 0.5$ (bottom left) and $\sigma_z = 0.3$ (bottom right).

With the data used in this work we find a significant reduction in the number of catastrophic outliers and outliers by sacrificing a minimal number of non-outlier predictions; for example, we find that by removing all galaxies in the evaluation sample with a photo-z uncertainty $\sigma_z > 0.3$, the RMS error was reduced by 57.6%, outliers were reduced by 70.1%,

and catastrophic outliers were reduced by 80.43% – at the cost of removing only 11% of the evaluation set. See Fig. 4.15 for the N(z) distribution of removed galaxies for example cases of $\sigma_z > 0.3$ and $\sigma_z > 0.5$.

3.5 Results

The BNN generally satisfies LSST photo-z science requirements in the range of $0.3 < z < 1.5$ (redshift range for weak lensing analyses – see Fig. 3.8) and performs as well or better than the 6 common alternative methods investigated in this study (see Table 3 and Figs. 3.7 and 3.8). We compare the BNN and NN to a support vector machine (Cortes & Vapnik, 1995), a random forest (Breiman, 2001), and a gradient boosting model, XGBoost (Chen & Guestrin, 2016), using the same data discussed in §4.3.1. We also form a comparison to photometric redshift predictions measurements from the HSC team (Nishizawa et al., 2020), which used the template-fitting model Mizuki and empirical method DEmP (Hsieh & Yee, 2014). To form a comparison, we relied on the photometric redshifts produced by the HSC team. Mizuki and DEmP were trained and evaluated on a slightly larger data set of 300k galaxies by (Nishizawa et al., 2020), but the majority of galaxies overlap with the data set introduced in this work. We crossmatched the HSC data set with the object IDs of our data to obtain a pre-evaluated sample of ∼60 thousand galaxies. Another photo-z investigation performed by Schuldt et al. (2021) utilized HSC imaging data and obtained a precision of $\Delta_z = |z_{phot} - z_{spec}| = 0.12$ with a convolutional neural network averaged over all galaxies in the redshift range $0 < z < 4$. We obtain $\Delta_z = 0.0031$ for the NN and $\Delta_z = 0.0032$ for the BNN averaged over all galaxies in our data set in this range. We note that a perfect comparison between photo-z models requires identical training, validation, and evaluation data sets. While the photo-z models from Schuldt et al. (2020a) and the HSC team compared in this work utilized largely the same data that was used in this work, there are some differences between their data and the data used in this investigation, which

introduces additional uncertainty in the comparison made between results.

We note that training a Bayesian neural network does not deterministically produce weights on the same data. The weights in the variational layers are sampled from a Gaussian distribution. The results presented here are representative of a typical training run with the BNN model presented in this work. However, there can be variations of several percents in outlier rates and other metrics depending on the training run. There can also be variations in the final loss achieved at the end of training. We find that the accuracy of a particular training run is correlated to the final loss value.

3.5.1 Using BNN Uncertainties to Identify Outliers

The BNN with the outlier removal method discussed in §3.1 stands out as the overall best performing model for the majority of photo-z performance metrics considered in this work, achieving the lowest percentage of outliers, catastrophic outliers, and RMS error. The outlier removal method described in §3.1 is visualized in Figs. 3.4,3.5,3.6, 3.8, and 4.15. Notably, Fig. 3.5 shows the performance of the NN and BNN with respect to LSST photo-z science requirements. The utilization of the photo-z uncertainties produced by the BNN to remove poor predictions significantly reduces RMS error. The BNN satisfies the LSST photo-z science requirements with respect to RMS error and 3σ outlier fraction across $0 < z < 2.5$, however the bias requirement is only partially met in the range $0.3 < z < 1.2$. We note that the BNN bias deviation from the acceptable range is confluent with the drop-off of the galaxy population in the N(z) distribution in Fig. A.2.

3.5.2 Bayesian Neural Network Photo-z Uncertainty Estimates

We find that the BNN produces accurate uncertainties as defined by the probabilistic metrics. The quality of the uncertainties produced by the BNN are visualized in Figs 3.5, 3.6, and 3.10. The BNN 3σ outlier fraction is shown in Fig. 3.5, which indicates that uncertainties are

generally well-estimated on average across the redshift range $0 < z < 2.5$. It is notable that the BNN performs best with respect to the σ outlier fraction when no galaxies with large uncertainties are removed. The PIT histogram produced for a sample determination with the BNN is shown in Fig. 3.10. The PIT histogram is generally flat, as is desired, however the slight bump in the middle indicates that the photo-z PDFs tend to be overly broad. For a comparison of PITs produced by other probabilistic photo-z methods (performed on different data) see Schmidt et al. (2020b). The BNN uncertainty coverage of the sample is provided in Fig. 3.6, showing acceptable agreement with the target 68% confidence interval up to the target redshift interval for weak lensing applications $0.3 < z < 1.5$, indicating the uncertainties of photo-z estimates for this galaxy population are accurately defined.

The results from evaluating the NN and BNN models on the evaluation set are available at `https://zenodo.org/doi/10.5281/zenodo.10145347`.

3.5.3 Investigating the effect of non-representative training data

The distribution of brightness of the training sample is peaked at brighter magnitudes compared to the full HSC photometric sample because of the need for spectroscopy. We investigated how this bias might affect our results by re-sampling the testing dataset to have a magnitude distribution closer to original HSC dataset. We find that the performance on this re-sample is similar or slightly worst by about 1 to 2% depending on the metric. See Appendix A for more details.

3.6 Discussion

Future large-scale astronomical surveys will provide high quality observations of billions of celestial objects that will be used to investigate the mysterious and unknown nature of dark matter and dark energy. LSST will play a crucial part in this investigation; we model our analysis here with respect to the photo-z science requirements provided by the LSST team.

Table 3.3 Comparison of the performance results with each model discussed in §2. We use the data discussed in §4.3.1 to train and evaluate a NN, BNN, a SVM SPIDERz (Jones & Singal, 2017), a random forest (Breiman, 2001), and a gradient boosting model XGBoost (Chen & Guestrin, 2016). We also include a comparison to the template-fitting model, Mizuki, and empirical method, DEmP (Hsieh & Yee, 2014), that were evaluated on a larger, overlapping data set in (Nishizawa et al., 2020). To form a comparison to Mizuki and DEmP in this work, we crossmatched the larger data set with the object IDs of our data discussed in §4.3.1 to obtain a pre-evaluated sample of 60 thousand galaxies.

Network	O	O_c	O_b	RMS	$\|b\|$	Scatter	$L(\Delta z)$
BNN	0.079	0.023	0.023	0.174	0.013	0.026	0.105
BNN ($\sigma_z < 0.5$)	0.034	0.0071	0.025	0.0854	0.002	0.029	0.066
BNN ($\sigma_z < 0.3$)	0.0236	0.0045	0.017	0.0738	0.002	0.022	0.056
NN	0.059	0.029	-	0.174	0.0001	0.026	0.089
Mizuki	0.274	0.102	-	0.307	0.011	0.055	0.289
DEmP	0.250	0.092	-	0.277	0.003	0.040	0.258
RF	0.092	0.006	-	0.088	0.001	0.012	0.065
XGBoost	0.105	0.022	-	0.149	0.002	0.033	0.144
SPIDERz	0.090	0.051	-	0.199	0.002	0.044	0.135
LSST Req.	-	-	-	< 0.2	< 0.003	$<$ -	-

Bayesian Neural Networks have been used in the past for photo-z estimation (e.g. Zhou et al., 2022; Schuldt et al., 2020a; Jones et al., 2021b), however this is the first BNN (following our prototype in Jones et al. (2022a)) applied to photometry observations of a representative dataset similar to what we will obtain from future large scale surveys like LSST. This work is among the first that evaluates the photo-z model performance with respect to the LSST science requirements as a function of redshift.

The BNN largely satisfies LSST science requirements in the redshift range of interest for LSST weak lensing surveys ($0.3 < z < 1.5$), and outperforms alternative models on the same data, however the BNN does not fully satisfy the bias requirement. We believe the BNN model can be further optimized for these requirements. Compared to the NN model, the BNN has the advantage of producing uncertainties for each prediction, which are both required for precision cosmology and can be used to eliminate galaxies with large uncertainties from the data sample. We note that both the NN and BNN models generally

perform worse at higher redshifts, which is due in large part to the reduced signal to noise for distant dim sources and also the disproportionate number of high redshift sources (z $>$ 2.5) compared to low redshift sources (see Fig. A.2 and also the discussion in Wyatt & Singal (2020)).

The BNN model introduced here is an improved version of the model we introduced in a previous work (Jones et al., 2022a). The uncertainty estimates produced in the BNN model discussed in this work are significantly improved from the previous model – due in large part to optimizing the learning rate during training and modifying the network architecture. The previous model network contained four variational layers, which we adjusted to contain three dense layers and one variational layer. We find that the coverage for the architecture with all variational layers produces coverage that is generally 10% larger than expectation (uncertainties too large). Using a single variational now produces more accurate coverage.

The BNN uncertainty coverage of the sample provided in Fig. 3.10 shows excellent agreement with the target 68% confidence interval up to $z = 1.5$, indicating the uncertainties of photo-z estimates for this galaxy population are accurately defined. Beyond $z = 1.5$, the coverage oscillates around the target %68 level. A likely explanation for the reduced quality of galaxy uncertainties beyond $z = 1.5$ is the lack of data samples at this redshift range (see Fig. A.2) compared to lower redshifts. Another possible factor affecting photo-z uncertainties may result from a disparity between the complexity present in the band magnitudes compared to the BNN model; we use five photometric band fluxes paired with a single spectroscopic redshift per galaxy for training. In a future work we will apply galaxy photometric images to a Bayesian convolutional neural network, which is likely to contribute more useful information than the five photometric measurements per galaxy.

Another benefit of using a BNN for photo-z estimation is the use of the photo-z uncertainty z_σ to preemptively flag photo-z predictions with high uncertainties as potential poor predictions. An acceptable balance needs to be achieved between the number of galaxies correctly flagged as poor predictions versus the number of non-outlier galaxies removed for

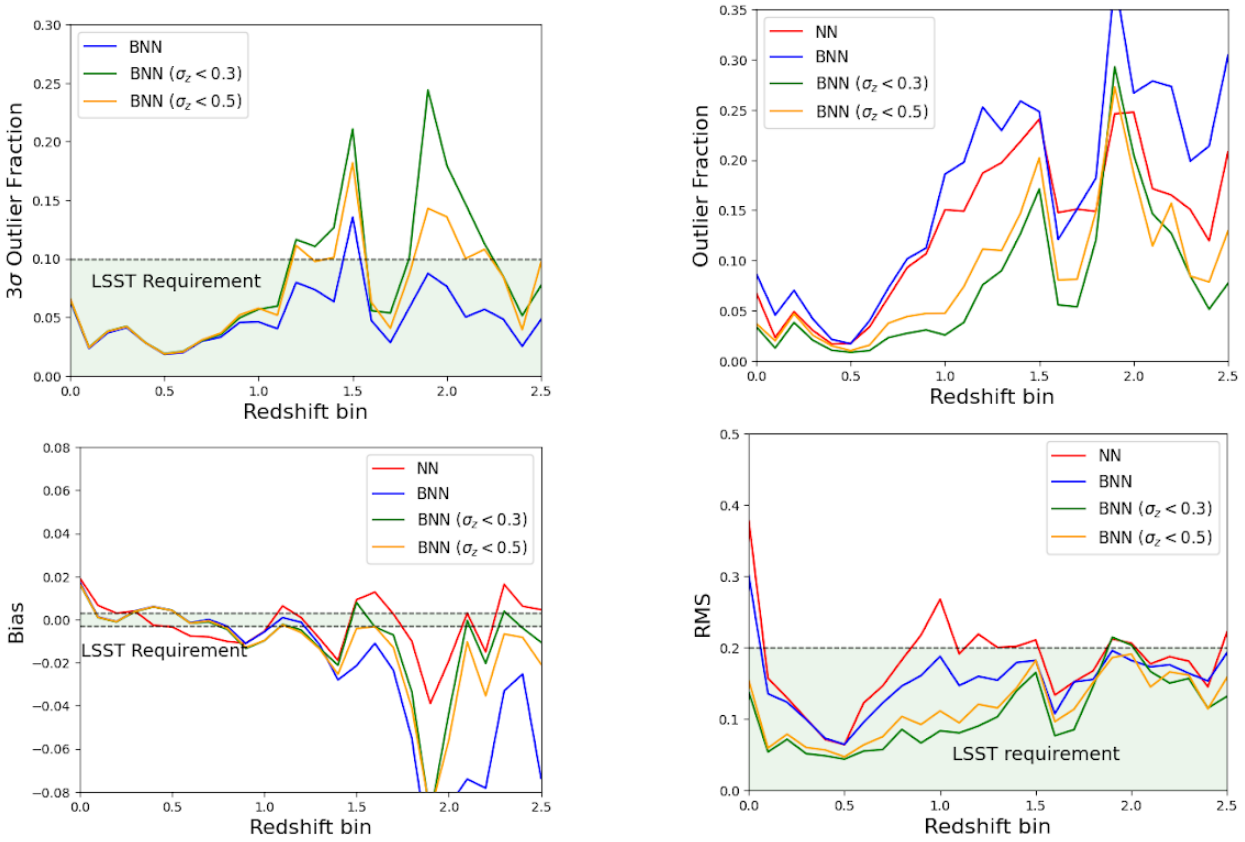

Figure 3.5 BNN and NN performance with respect to LSST photo-z requirements. We note that the 3σ outlier fraction can only be calculated with the BNN because the metric requires photo-z uncertainties so we additionally include the standard outlier fraction for the NN and BNN for comparison. The plots reflect results with 80% of galaxies for training, 10% for validation, and 10% for evaluation. We include only those results in the redshift range $0 < z < 2.5$ because the $N(z)$ distribution of the data set degrades significantly at higher redshifts (see Fig. A.2) and would likely significantly improve given sufficient training data.

a given z_σ cutoff value. With the data used in this work we find a significant reduction in the number of catastrophic outliers and outliers can be achieved by sacrificing a relatively small number of non-outlier predictions; for example, we find that by removing all galaxies in the evaluation sample with a photo-z uncertainty $\sigma_z > 0.3$, the RMS error was reduced by 57.6%, outliers were reduced by 70.1%, and catastrophic outliers were reduced by 80.43% – at the cost of removing only 11% of the evaluation set.

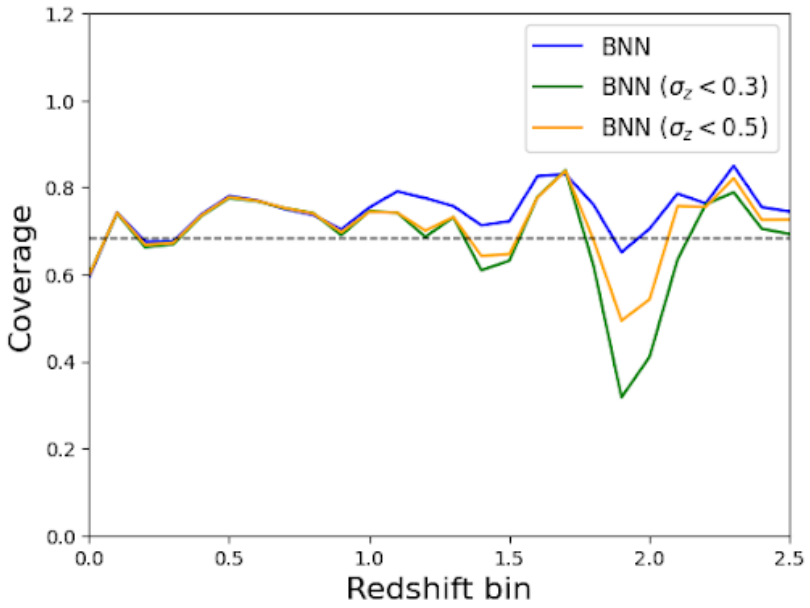

Figure 3.6 The fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies.

3.7 Conclusion

In preparation of the coming influx of data from large scale surveys like LSST, it is important to prepare photo-z estimation models in advance. Such models must provide both accurate photo-z predictions and reliable photo-z uncertainties, which are required for using photo-z predictions in subsequent cosmological analyses. The quality of photo-z models should be assessed using data that is representative of data from future large scale surveys, and principally evaluated using the scientific requirements provided by those surveys.

This work introduces a BNN model for photometric redshift estimation. We apply the BNN to data from the Hyper Suprime Cam survey, which is designed to reflect the data we will soon receive from large scale surveys such as LSST. We evaluate the BNN with respect to LSST science requirements and compare the results to alternative photo-z estimation tools

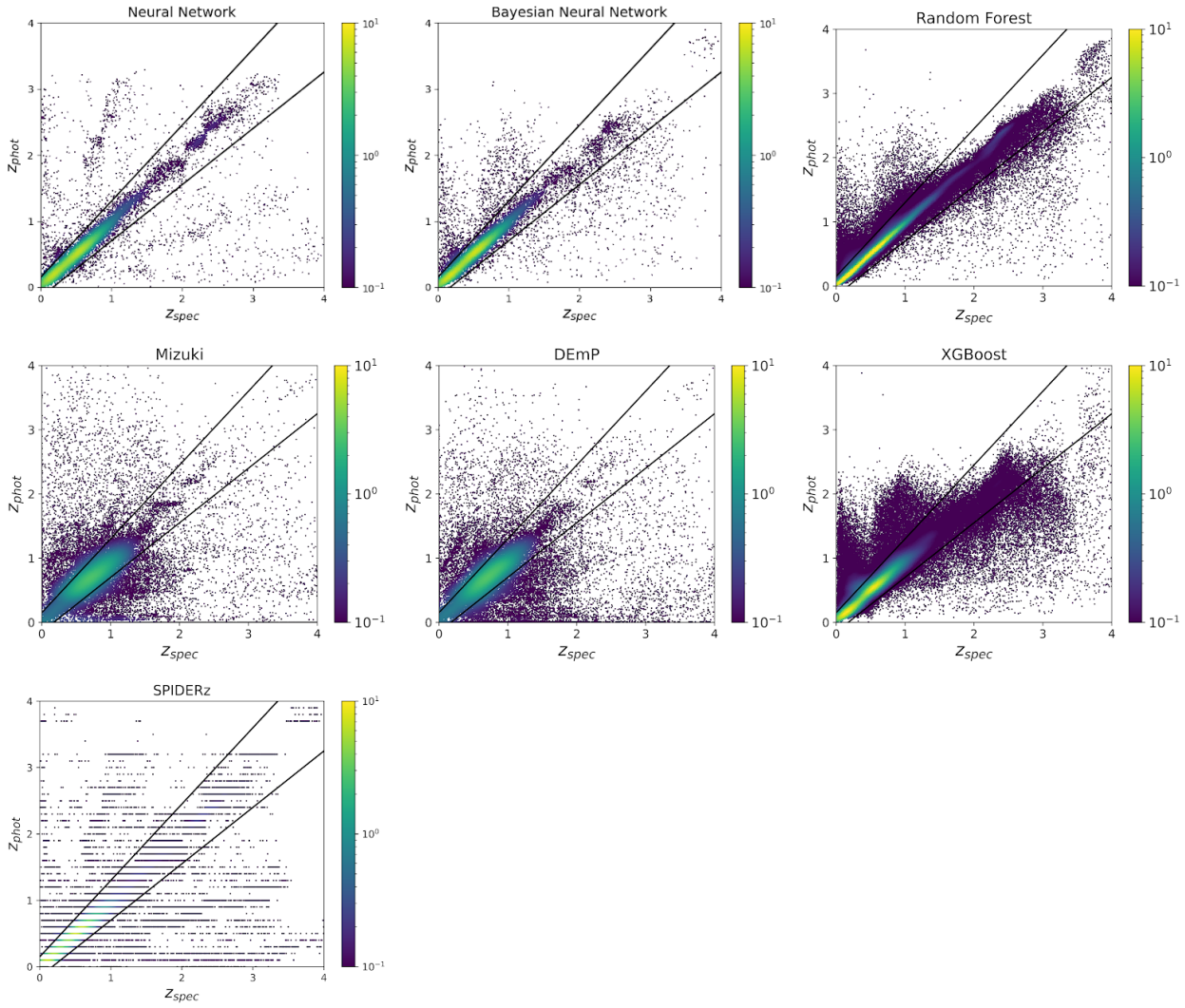

Figure 3.7 Visualization of predicted photo-zs versus measured spectroscopic redshifts by the models discussed in §2. The results of these determinations are quantified in Table 3. The colorbars indicate the density of evaluation data points as computed with a Gaussian kernel-density estimation.

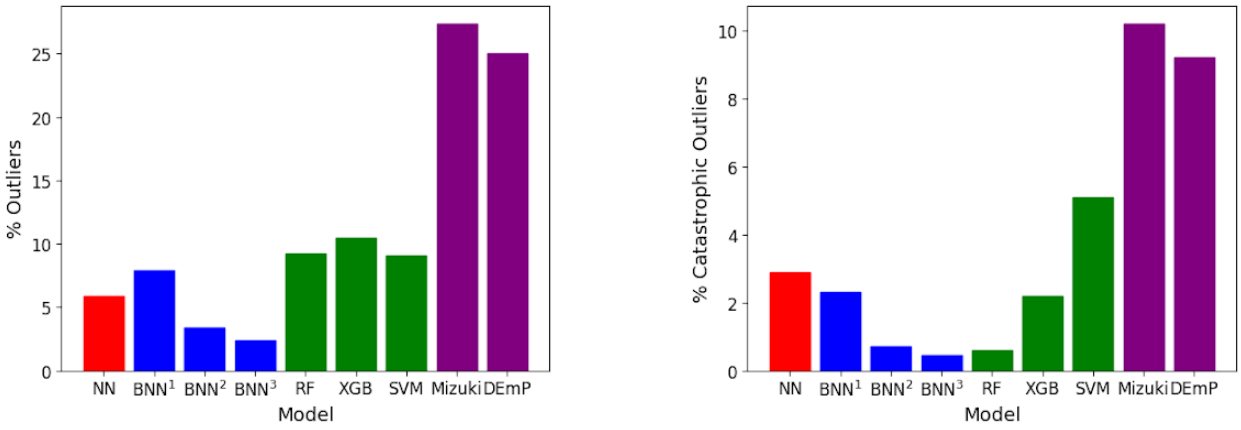

Figure 3.8 Comparison of the percentage of outliers (Eqn 1) and catastrophic outliers (Eqn 2) achieved with each model. BNN^1 refers to the default BNN results with no galaxies removed based on a z_σ criteria. BNN^2 and BNN^3 refer to results obtained after removing all galaxies from the evaluation set containing photo-z uncertainties greater than 0.5 and 0.3, respectively. We use the data discussed in §4.3.1 to train and evaluate a NN, BNN, a SVM SPIDERz (Jones & Singal, 2017), a random forest (Breiman, 2001), and a gradient boosting model XGBoost (Chen & Guestrin, 2016). We also include a comparison to the template-fitting model, Mizuki, and empirical method, DEmP (Hsieh & Yee, 2014), that were evaluated on a larger, overlapping data set in (Nishizawa et al., 2020). To form a comparison to Mizuki and DEmP in this work, we crossmatched the larger data set with the object IDs of our data discussed in §4.3.1 to obtain a pre-evaluated sample of 60 thousand galaxies.

including a fully connected neural network, random forest, support vector machine (Jones & Singal, 2017), XGBoost, Mizuki, and DEmp (Aihara et al., 2018, 2019). We find that the BNN meets two of the three LSST photo-z requirements in the redshift range considered for weak lensing cosmological probes ($0.3 < z < 1.5$) and provides superior photo-z estimations to the other models.

A key attribute of the BNN model is the production of photo-z uncertainties, which are needed for using photo-z results in cosmological analyses. We find that the BNN produces accurate uncertainties. Using a coverage test, we find excellent agreement with expectation – 68.5% of galaxies between $0 < 2.5$ have 1-σ uncertainties that cover the spectroscopic value. In addition, the BNN photo-z uncertainties can be used to flag likely outlier or catastrophic outlier estimates with high success.

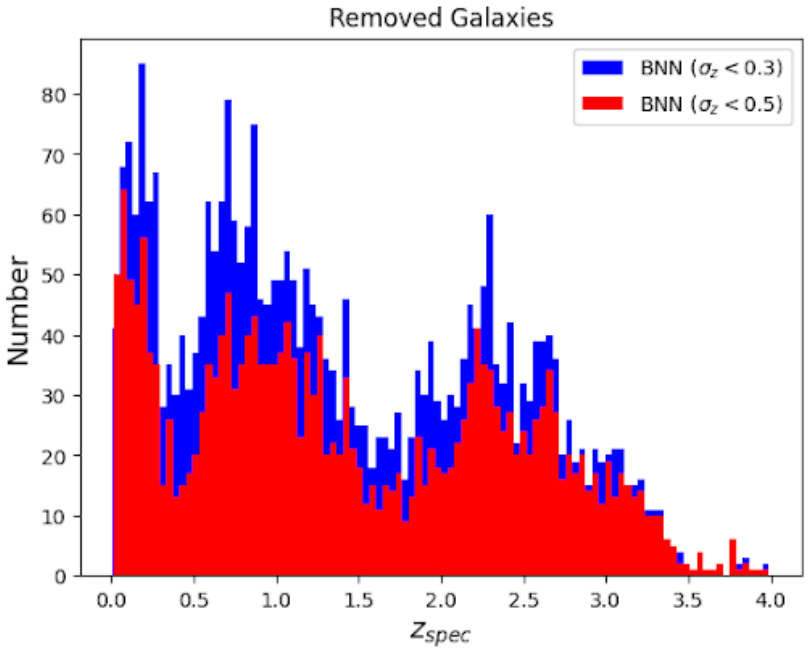

Figure 3.9 Histogram of photo-z uncertainties produced by the BNN that exceed 0.3 and 0.5. By removing all galaxies in the evaluation sample with a photo-z uncertainty $\sigma_z < 0.3$, outliers were reduced by 70.1%, and catastrophic outliers were reduced by 80.43% – at the cost of removing 11% of the evaluation set. Using a photo-z uncertainty cutoff of 0.5 reduces the number of outliers by 70.1% and catastrophic outliers by 67.8% at the cost of removing 7.67% of the evaluation set.

This analysis is subject to the potential sources of bias that affect most photometric redshift estimation studies. For example, spectroscopic redshift observations are biased toward high luminosity galaxies, particularly at higher redshift ranges ($z > 1.5$), which may not be fully representative of galaxy populations at a specific redshift range. Another source of bias in this analysis is the underrepresented galaxy population in the N(z) distribution beyond z = 1.5. Both of these potential sources of bias can be alleviated with improved spectroscopic samples in future galaxy surveys.

We will continue this analysis by applying the BNN method to galaxy images via a Bayesian convolutional neural network in a forthcoming paper.

We are grateful for the financial support for this work from the Sloan Foundation.

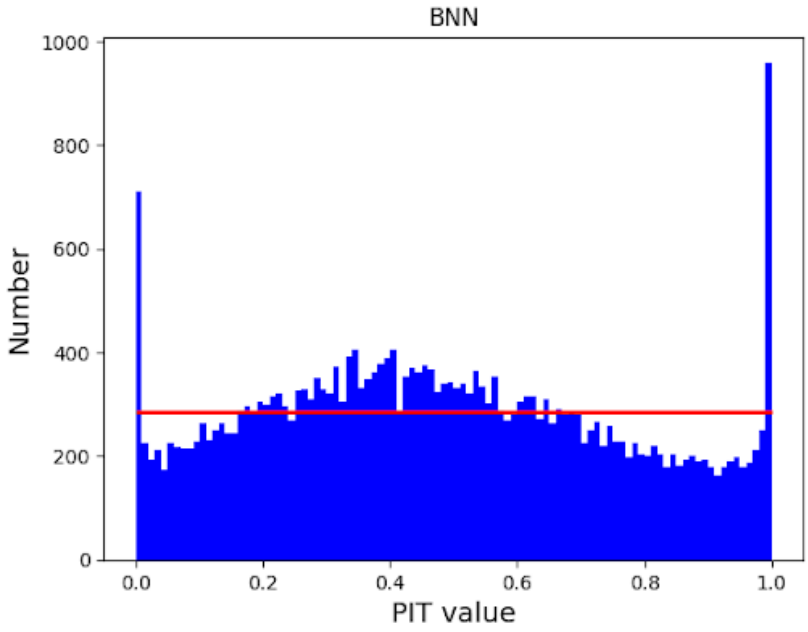

Figure 3.10 PIT histogram of the photo-z PDF produced by the Bayesian Neural Network. The red horizontal line indicates the ideal PIT histogram distribution: if the PIT histogram peaks at the center, the photo-z PDFs are generally too broad, and if the PIT histogram peaks at high and low PIT values, the PDF samples are too narrow or contain a large amount of catastrophic outliers.

3.8 Appendix

3.8.1 Addressing potential biases in the dataset

Because our selection of data for training and evaluation relied on only those galaxies for which spectroscopic redshifts are available, the magnitude distribution is biased compared to a the bulk photometric sample from HSC (see Fig. 11). In order to address this, we have performed an additional analysis with the NN and BNN models using a re-sampled testing set that mimics the magnitude distribution of the bulk HSC photometry sample (Fig. 4.16). We use the g-band to re-sample our testing dataset to reproduce the overall HSC g-band distribution. Since the color of the resampled dataset is not enforced, the resampled distribution in the *rizy* filters are slightly different than the main HSC distribution. However,

these distributions are all much closer than the initial spectrosopic sample. The re-sampling process reduced our testing dataset size from 28,640 to 8,517 −− a 70.3% decrease. The distribution of redshifts for the resampled testing data is similar to the original (Fig. 3.12).

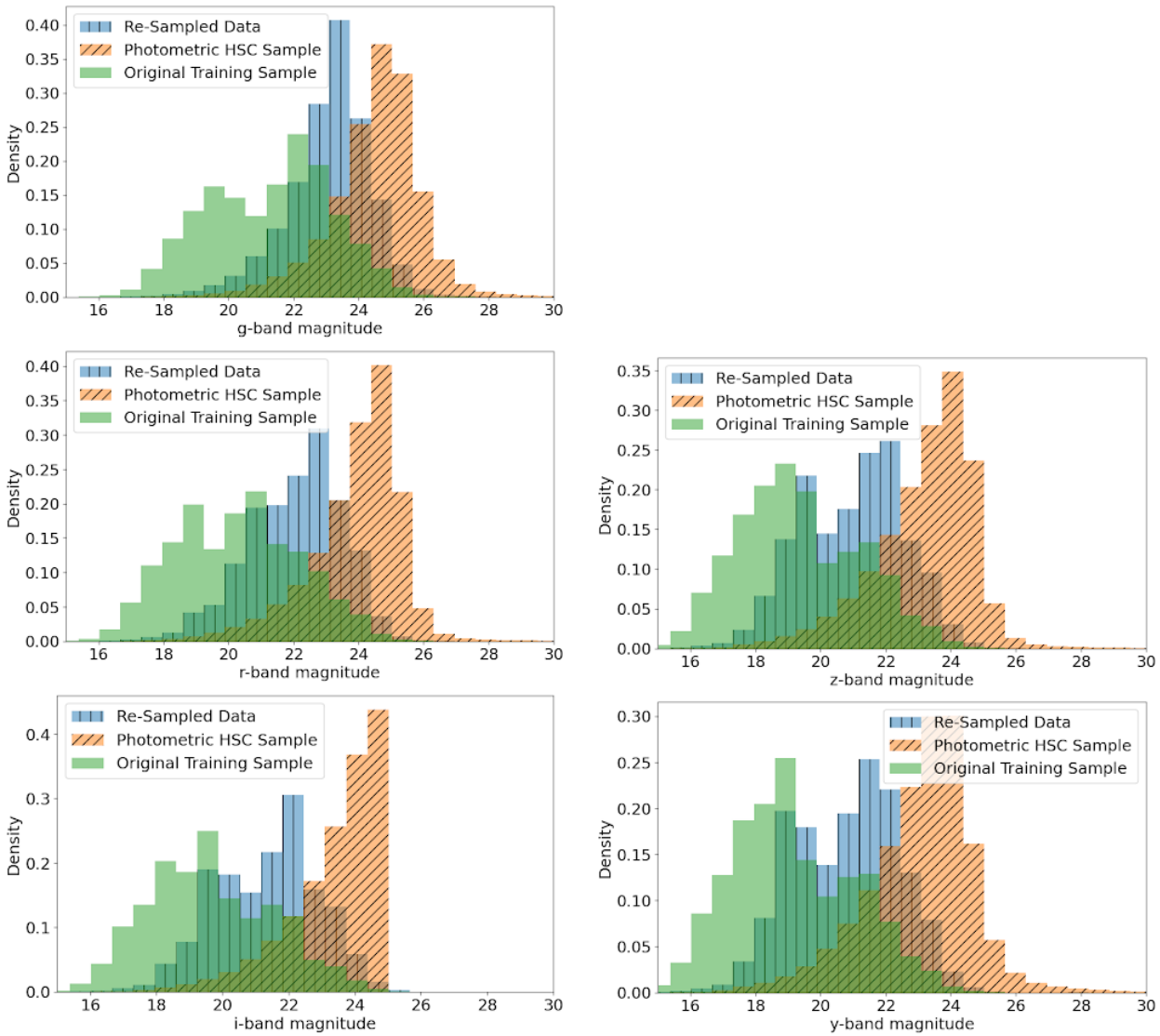

Figure 3.11 Visualisation of the grizy bands before and after the data is re-sampled to approximate the bulk HSC photometry sample.

Overall, the model does not perform significantly different on the resampled testing dataset (Fig. 3.13,3.14,3.15,3.16,4.17). Overall, the resampled testing data performs about

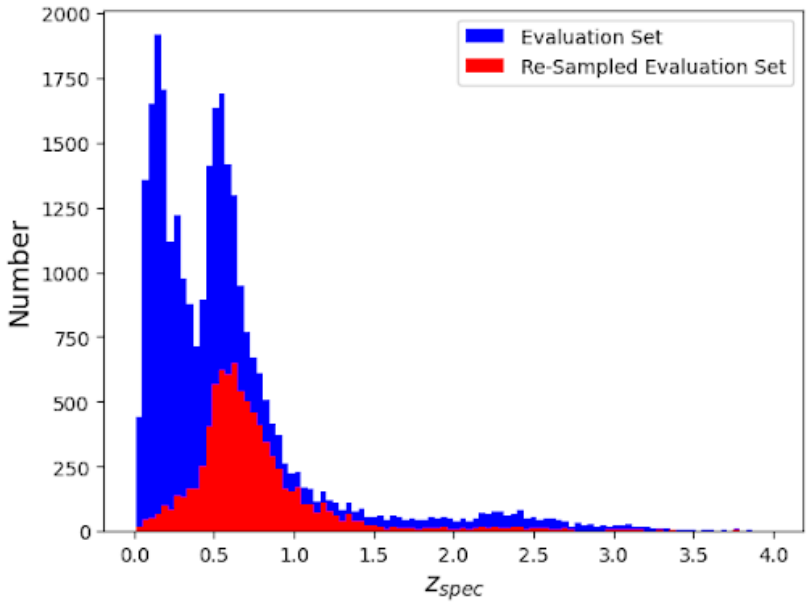

Figure 3.12 N(z) distributions of the original evaluation set discussed in §4.3.1 and the re-sampled evaluation set.

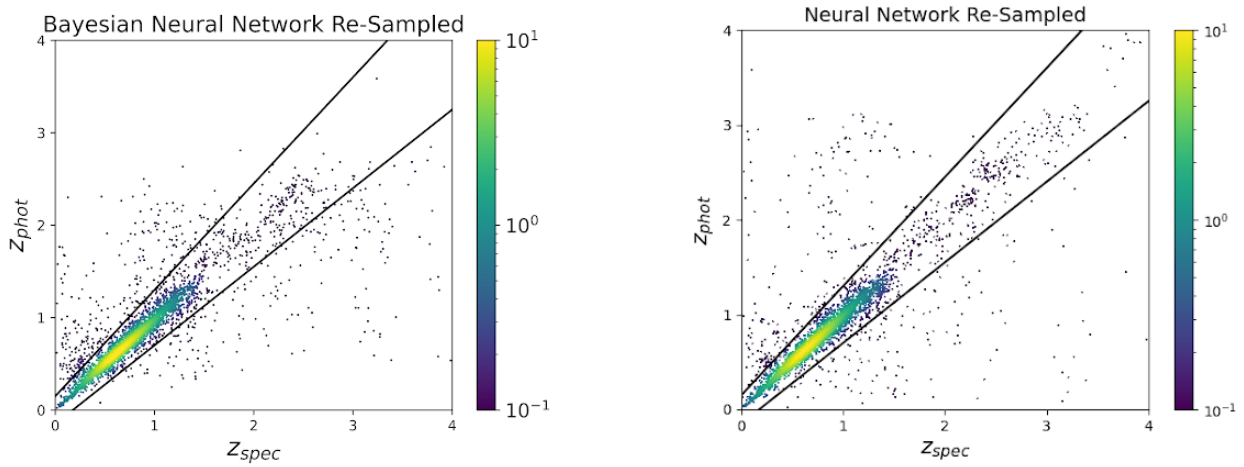

Figure 3.13 Visualization of the NN and BNN results using an evaluation set that is re-sampled to more closely approximate the bulk HSC photometry. The models are trained on the original data discussed in §4.3.1.

$1-2\%$ worse than the original testing dataset. The coverage of the re-sampled data is not signficantly affected either (Fig. 3.14). In addition, the PIT histogram (Fig. 3.16) indicates

that the photo-z PDFs produced in the re-sampled testing set closely resemble the photo-z PDFs produced in the original evaluation sample shown in in Fig. 4.15. Table 4.6 quantifies the performance metrics shown in Table 4.5.

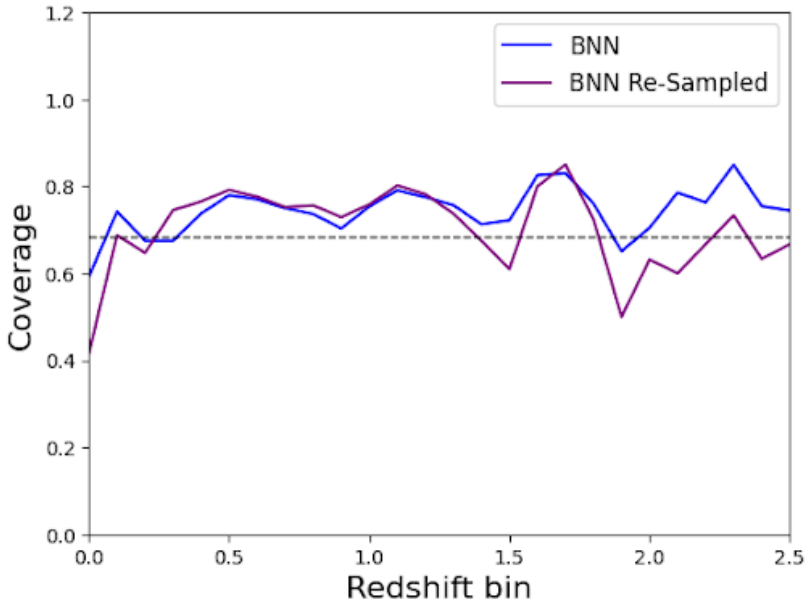

Figure 3.14 Comparison of the photo-z uncertainty coverage present in the original evaluation set compared to the re-sampled evaluation sample. Coverage is defined as the fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies.

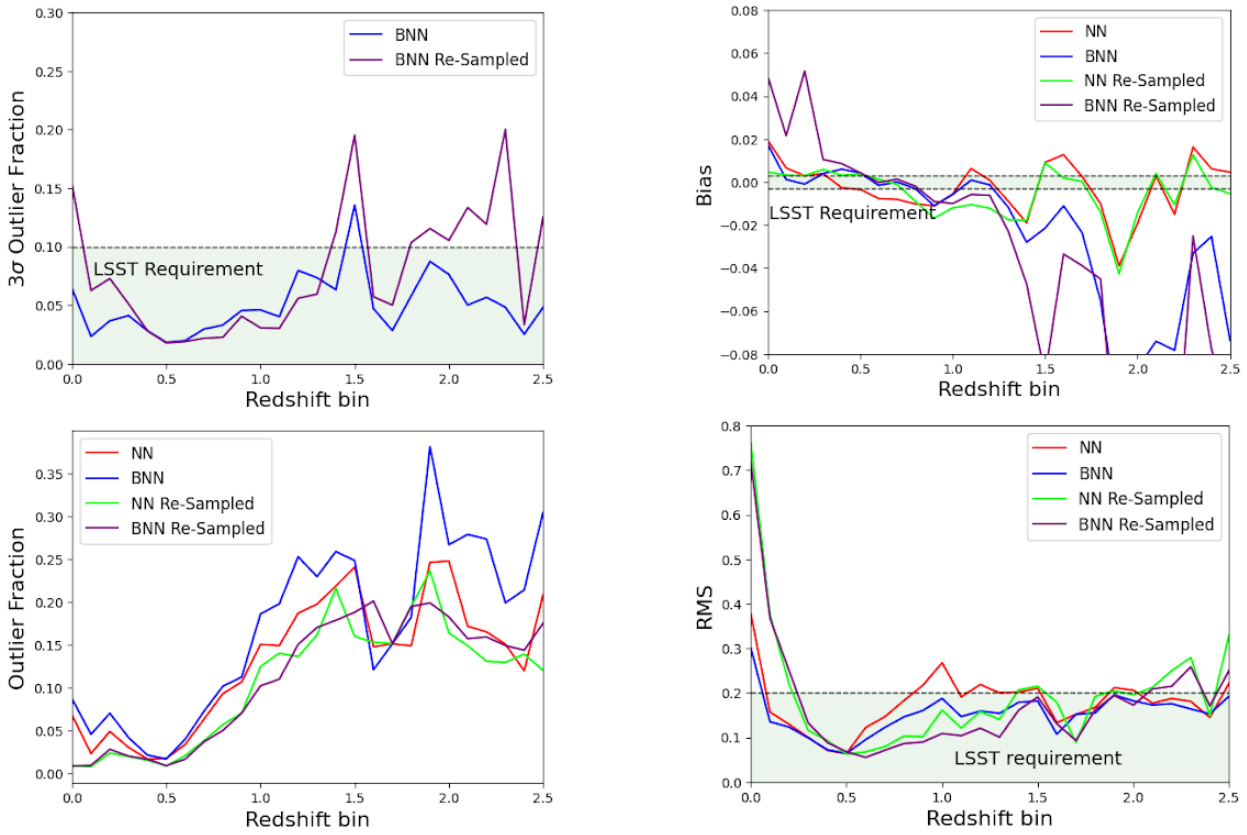

Figure 3.15 BNN and NN performance with respect to LSST photo-z requirements using an evaluation set with photometry that is re-sampled to approximate the bulk HSC photometry. We note that the 3σ outlier fraction can only be calculated with the BNN because the metric requires photo-z uncertainties so we additionally include the standard outlier fraction for the NN and BNN for comparison. The plots reflect results with 80% of galaxies for training, 10% for validation, and 10% for evaluation. We include only those results in the redshift range $0 < z < 2.5$ because the $N(z)$ distribution of the data set degrades significantly at higher redshifts (see Fig. A.2) and would likely significantly improve given sufficient training data.

Network	O	O_c	O_b	RMS	$\|b\|$	Scatter	$L(\Delta z)$
BNN	0.079	0.023	0.0233	0.174	0.013	0.026	0.1054
BNN re-sampled	0.08	0.017	0.026	0.134	0.0047	0.028	0.1082
NN	0.059	0.029	0.174	0.0001	0.026	0.089	0.089
NN re-sampled	0.067	0.021	0.14	-0.0029	0.027	0.097	0.097

Table 3.4 Comparison of the performance results with the NN and BNN with the original evaluation data set discussed in §4.3.1 and the re-sampled evaluation set to approximate the bulk HSC photometry. We use the data discussed in §4.3.1 to train. The re-scaling process reduces the initial evaluation set size from 28,640 to 8,517 −− a 70.3% decrease.

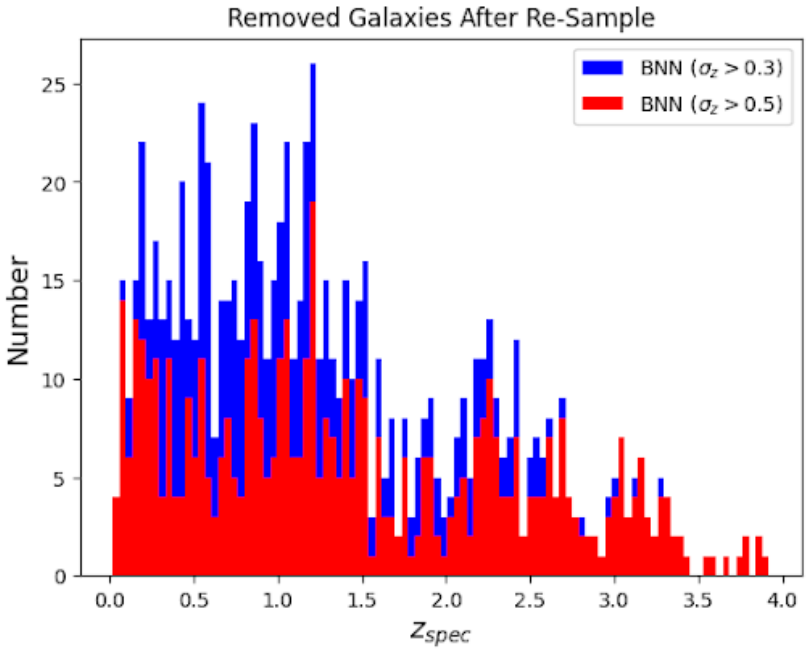

Figure 3.16 Histogram of photo-z uncertainties produced by the BNN that exceed 0.3 and 0.5 using the re-sampled dataset. By removing all galaxies in the evaluation sample with a photo-z uncertainty $\sigma_z < 0.3$, outliers were reduced by 70.1%, and catastrophic outliers were reduced by 80.43% – at the cost of removing 11% of the evaluation set. Using a photo-z uncertainty cutoff of 0.5 reduces the number of outliers by 70.1% and catastrophic outliers by 67.8% at the cost of removing 7.67% of the evaluation set.

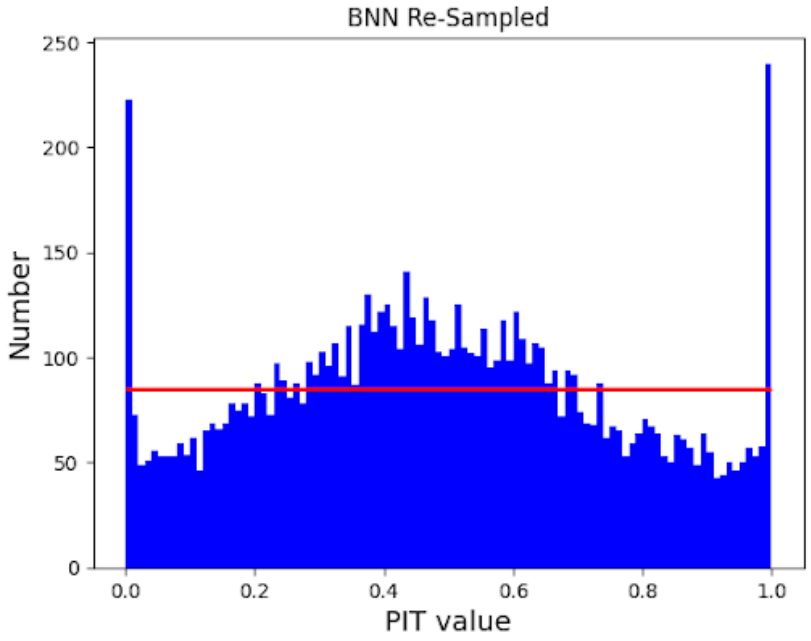

Figure 3.17 PIT histogram of the photo-z PDF produced by the Bayesian Neural Network using an evaluation set with photometry that is re-sampled to approximate the HSC bulk photometry. The red horizontal line indicates the ideal PIT histogram distribution: if the PIT histogram peaks at the center, the photo-z PDFs are generally too broad, and if the PIT histogram peaks at high and low PIT values, the PDF samples are too narrow or contain a large amount of catastrophic outliers.

CHAPTER 4

Redshift Prediction with Images for Cosmology using a Bayesian Convolutional Neural Network with Conformal Predictions

This thesis chapter has been accepted in the literature as Redshift Prediction with Images for Cosmology using a Bayesian Convolutional Neural Network with Conformal Predictions

Evan Jones, Tuan Do, Bernie Boscoe, Jack Singal, Yujie Wan, and Zooey Nguyen, *The Astrophysical Journal, accepted*

4.1 Abstract

In the emerging era of big data astrophysics, large-scale extragalactic surveys will soon provide high quality imaging for billions of celestial objects to answer major questions in astrophysics such as the nature of dark matter and dark energy. Precision cosmology with surveys requires accurate photometric redshift estimation with well-constrained uncertainties as inputs for weak lensing models to measure cosmological parameters. Machine learning methods have shown promise in optimizing the information gain from galaxy images in photo-z estimation, however many of these methods are limited in their ability to estimate accurate uncertainties. In this work we present one of the first applications of Bayesian convolutional neural networks for photo-z estimation and uncertainties. In addition, we use conformal mapping to calibrate the photo-z uncertainties to achieve good statistical coverage. We use

the public GalaxiesML dataset of $\sim 300k$ galaxies from the Hyper Suprime-Cam survey containing five-band photometric images and known spectroscopic redshifts from $0 < z < 4$. We find that the performance is much improved when using images compared to photometry, with the BCNN achieving 0.098 rms error, a standard outlier rate of 3.9%, 3σ outlier rate of 4.5%, and a bias of 0.0007. The performance drops significantly beyond $z > 1.5$ due to relative lack of training data beyond those redshifts. This investigation demonstrates the power of using images directly and we advocate that future photo-z analysis of large scale surveys include galaxy images.

4.2 Introduction

Dark matter and dark energy comprise $\sim 95\%$ of the energy density of the universe, but their natures are largely unknown. To investigate dark matter and dark energy, large-scale extragalactic surveys such as the Large Scale Survey of Space and Time (LSST – e.g. Ivezić et al., 2008) and Euclid (e.g. Collaboration et al., 2022) will soon provide observations of billions of galaxies. Cosmological probes of dark matter and dark energy aim to measure the structure and evolution of the universe, and thus rely in part on precise measurements of the redshifts of hundreds of millions of galaxies and accurate uncertainties.

Spectroscopic redshift measurements are the most reliable method of obtaining redshifts, but are too time consuming and therefore not a suitable solution for obtaining the number of redshifts required for cosmological measurements. Photometric redshift estimation can provide redshifts for billions of galaxies, however photo-z estimates are subject to significant systematic errors because the spectral information of a galaxy is sampled with only a limited number of imaging bands. These systematic errors can manifest as outlier predictions that are far from their true redshift, biases in the distribution of redshift predictions, and large scatter in redshift predictions (e.g. Newman & Gruen (2022)). These systematics strongly affect science goals such as weak lensing inferences of cosmological parameters since photo-

z uncertainties will be propagated into models constraining cosmological quantities. Any photo-z model developed for the potential application to these science missions must produce uncertainties on photo-z predictions.

According to the LSST Science Requirements Document (SRD)[1], sufficiently accurate photo-z estimates for $\sim$ four billion galaxies are required to meet the LSST science goals for their main cosmological sample. Specifically, for the $i < 25$ flux-limited galaxy sample measured by LSST, one must achieve

- number of galaxies $\approx 10^7$
- rms error < 0.2 (Equation 3 in Table 4)
- bias < 0.003 (Equation 4 in Table 4)
- 3σ catastrophic outliers $< 10\%$ total sample (Equation 9 in Table 4)

Currently, no published image-based model satisfies the LSST photo-z science requirements up to $z = 3$ (Tanaka et al., 2018a; Schuldt et al., 2020b; Schmidt et al., 2020a). Additionally, methods for rejecting the majority of outliers and characterizing their effects on the predictions must be developed (Ivezic, 2018). Beyond the LSST metrics stated in the SRD, we consider additional probabilistic metrics for quantifying the quality of uncertainty estimates (Malz & Hogg, 2020; Schmidt et al., 2020a; Jones et al., 2022a). The requirement thresholds for the probabilistic metrics are not as well quantified at this time as those for point metrics, but they allow us to compare the performance between different probabilistic models evaluated on the same data. Techniques for identifying photo-z outlier predictions in machine learning models have been investigated in Jones & Singal (2020), Wyatt & Singal (2020), Singal et al. (2022), and Jones et al. (2023).

There have been a handful of works over the last several years investigating the use of convolutional neural networks (CNNs) for photo-z estimation. Pasquet et al. (2019) was

[1]https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17

one of the first studies investigating the use of a CNN for photo-z estimation, wherein they applied a CNN for classification of redshift bins for a sample of low redshift SDSS galaxies in $0 < z < 0.4$. Treyer et al. (2023) similarly performed an analysis using a CNN on SDSS image data using 370k galaxies primary concentrated between $0 < z < 0.3$. Lin et al. (2022) applied a CNN to a SDSS and CFHTS sample between $0 < z < 0.3$. Ait-Ouahmed et al. (2023) applied a CNN to a sample from SDSS, CFHTS, and HSC that contains redshifts up to $z = 4$, but their primary analysis was limited to $0 < z < 1.6$. Schuldt et al. (2020a) applied a CNN to HSC imaging data throughout the range $0 < z < 4$.

In this work, we present a new method for photometric redshift estimation using a probabilistic Bayesian convolutional neural network model, which predicts both the redshifts and uncertainties that are necessary for constraining cosmological parameters. We train and apply our model to a dataset that extends to $z = 4.0$ to more accurately reflect the conditions in which these models might be used for surveys like LSST. This work is a continuation of our previous non-image based Bayesian neural network (BNN) model for photo-z predictions (Jones et al., 2022b, 2023).

We have three goals in this work: (1) develop a probabilistic image-based ML model that can produce robust uncertainties for photometric redshifts, (2) assess this model and other photo-z methods with respect to LSST requirements, and (3) investigate the use of photo-z uncertainties to identify likely outliers in photometric redshift predictions. For the analysis in this work, we use one of the largest publicly available machine-learning-ready galaxy image data sets[2] of $\sim 300k$ galaxies from the Hyper Suprime-Cam survey containing five-band photometric images and known spectroscopic redshifts from $0 < z < 4$ (Do et al. 2024, in prep). In §2 we discuss the data and network architecture; in §3 we discuss the conformal prediction analysis, in §4 we discuss the photo-z metrics, in §5 we state the results, and in §6 and §7 we provide a discussion and conclusion.

[2]https://doi.org/10.5281/zenodo.5528827

4.3 Data and Methods

4.3.1 Data: Galaxy observations

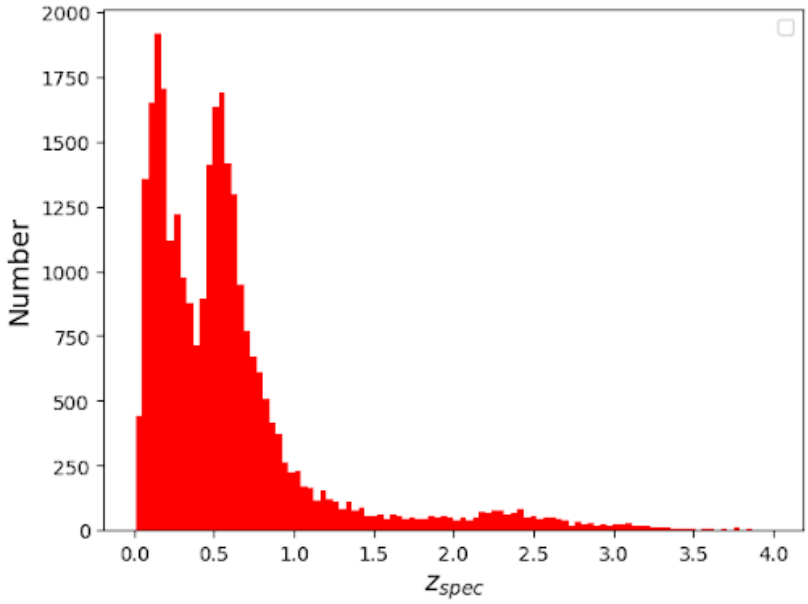

Figure 4.1 $N(z)$ distribution for the data set discussed in §4.3.1. For the photo-z determinations in this work we use training, validation, and testing sets consisting of 229,120, 28,640, and 28,640 galaxies respectively.

Images are a crucial part of the analysis in this work because they allow us to include full pixel-level information in the machine learning models. Because of the larger frequency of mergers at higher redshifts and the general evolution of galaxies with time, it is a reasonable hypothesis that these physical processes will change the appearance of galaxies as a function of redshift. For this reason a key component of our work is to use galaxy images as input into our photo-z network – a feature which has only become possible in recent years, because of improvements in deep learning models and availability of large datasets representative of future large scale surveys.

For this analysis, we use the GalaxiesML dataset discussed in Do et al. 2024 (in prep) for training and performance evaluation. This dataset is intended to approximate the data

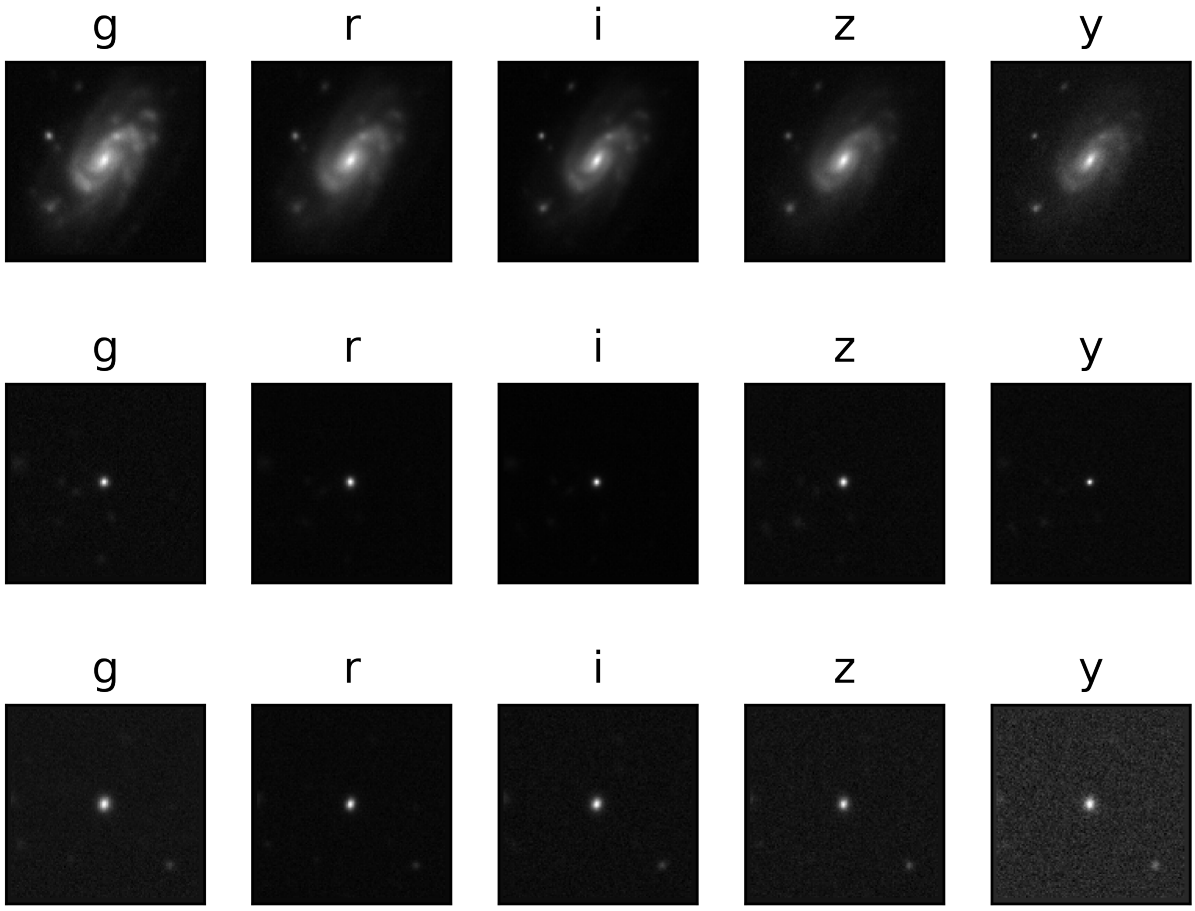

Figure 4.2 Example HSC galaxy images for the data set used in this work with *grizy* photometry for a low redshift galaxy at $z = 0.05$ (TOP), and a high redshift galaxy at $z = 3.92$ (MIDDLE), and another low redshift galaxy at $z = 0.14$ (BOTTOM). The similarity between the high redshift galaxy and the bottom low redshift galaxy highlights the difficulty of photo-z estimation.

produced by future large-scale deep surveys for photo-z estimation (Collaboration et al., 2021). GalaxiesML uses the Suprime Cam (HSC) Public Data Release 2 (PDR2) (Aihara et al., 2019), which is designed to reach similar depths as LSST but over a smaller portion of the sky. We choose the HSC survey because it mimics LSST in photometry and depth.

The final data set used in the analyses of this paper consists of ∼286,401 galaxies with 5-band grizy photometry and spectroscopic redshifts. Fig. A.2 contains the $N(z)$ distribution

Table 4.1 Quality cuts used to construct the data set.

photometry cuts	z_{spec} cuts
grizy_ cmodel_flux_flag = False	$z > 0$
grizy_ pixelflags_edge = False	$z \neq 9.9999$
grizy_ pixelflags_interpolatedcenter = False	$0 < z_{err} < 1$
grizy_ pixelflags_saturatedcenter = False	unique galaxy object ID
grizy_ pixelflags_crcenter = False	specz_flag_homogeneous = True
grizy_ pixelflags_bad = False	
grizy_ sdsscentroid_flag = False	

for the dataset and Fig. A.1 shows g, r, i, z, and y band images for three example HSC galaxies. Spectro-zs were obtained by crossmatching galaxy photometry from HSC with the HSC collection of publicly available spectroscopic redshifts using galaxy sky positions (d $< 1''$) in Lilly et al. (2009), Bradshaw et al. (2013), McLure et al. (2012), Skelton et al. (2014), Momcheva et al. (2016), Le Fèvre et al. (2013), Garilli et al. (2014), Liske et al. (2015) Davis et al. (2003), Newman et al. (2013), Coil et al. (2011), Cool et al. (2013). We used data quality cuts similar to Nishizawa et al. (2020) and Schuldt et al. (2021) (see Table 1 and Do et al. 2024 (in prep) for a full list), which are intended to remove outlier photometric measurements and poorly measured spectroscopic redshifts. We also required detections in each band. The spectroscopic redshift values are treated as the ground truth for training and evaluation. The galaxy sample extends from $0.01 < z < 4$, however the majority of the sample lies between redshift of 0.01 and 2.5 with peaks at z 0.3 and z 0.6 (see N(z) in Fig. A.2). We use 70% of the galaxies for training, 10% for validation, 10% for parameterisation with conformal mapping, and 10% for testing. The data used for training is available[3] from Do et al. 2024 (in prep).

[3] `https://zenodo.org/records/5528827`

4.3.2 Network architectures

We built two image-based neural networks for this work – one is a CNN that produces single-valued redshift predictions and one is a BCNN that outputs redshift probability distributions. Fig 3. depicts the differences between discrete and probabilistic neural network models for photo-z estimation. The CNN and BCNN models employed in this work are visualized in Fig 5. Both the CNN and the BCNN are implemented in TensorFlow (Abadi et al., 2016) and use 5-band *grizy* photometric images and magnitudes as input. To optimize the hyperparameters, we performed a grid search over the number of epochs, number of layers, number of nodes per layer, learning rate, loss function, activation function, and optimizer (see 4.3.3). The CNN has an output node to produce a single point estimate photo-z prediction while the BCNN has a final output node that produces a mean and standard deviation assuming a Gaussian distribution for each photo-z prediction. For the BCNN we use a negative log likelihood loss function with RMS error as the metric. We choose the negative log-likelihood loss function for the BCNN because it has been shown to be more effective than MAE for probabilistic NNs (Lakshminarayanan et al., 2016) The CNN uses a mean absolute error loss function (we also consider a custom loss function (Nishizawa et al., 2020) defined in equation 7 of Table 4). The CNN and BCNN use the Adam optimizer and have learning rates of 0.0005 and 0.001, respectively. A full description of the hyperparameters used for both models are provided in tables 2 and 3. We train using an AMD Ryzen Threadripper PRO 3955WX with 16-Cores and NVIDIA RTX A6000 GPU. Training runtimes are typically over 24 hours for 200 epochs for the final models and evaluation runtimes are on the scale of minutes.

4.3.3 Building CNN and BCNN architectures and hyperparameter tuning

There are a number of important distinctions between the CNN and BCNN that affect the way in which each type of model should be optimized. We find the quality of a CNN model

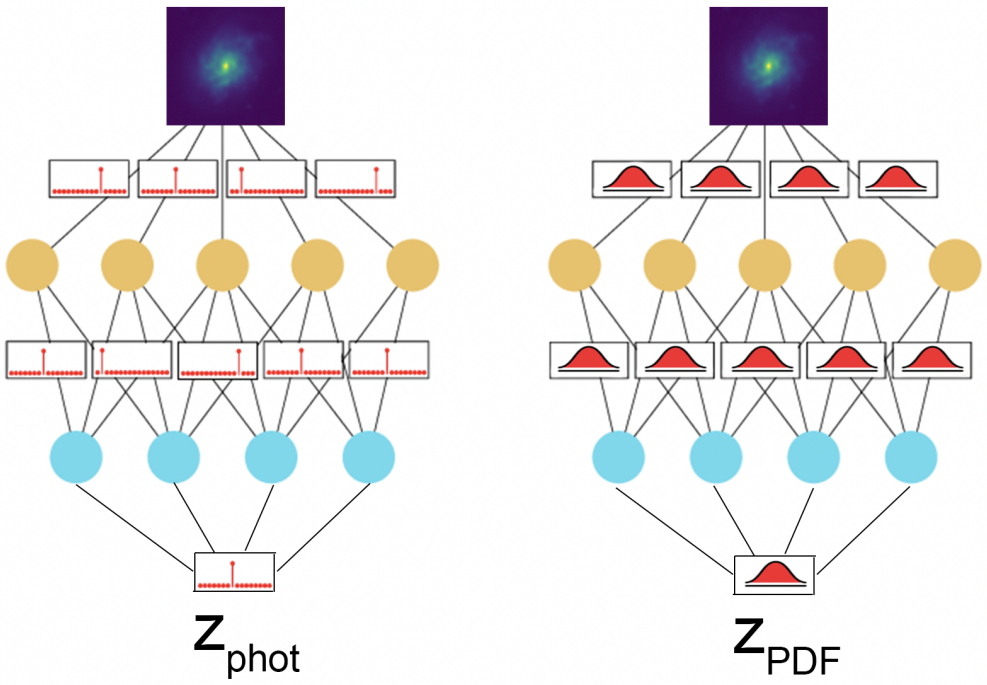

Figure 4.3 Comparison of simplified non-probabilistic (LEFT) and probabilistic (RIGHT) neural network models. The non-probabilistic model optimizes for discrete weights in each node (yellow and blue dots), whereas the probabilistic model optimizes for probability distributions over weights. Similarly, the non-probabilistic model produces a discrete photo-z prediction for each galaxy, while the probabilistic model produces a photo-z PDF for each galaxy.

is less variable to small adjustments in hyperparameter values or changes to individual layers compared to a BCNN, which tends to quickly degrade in performance or produce a divergent loss. Our CNN model investigation for photo-z estimation almost immediately produced results indicative of genuine learning between the image inputs and spectroscopic output. The BCNN required significantly more model tweaking and hyperparameter tuning to achieve similar results. We also found that transforming successful CNN networks into BCNNs by introducing additional variational layers at the end of the network provided lower loss and more accurate uncertainties than building a new network consisting of variational layers throughout. For the analysis in this work we transformed two distinct CNN architectures to obtain two BCNN architectures. One CNN network was created using the NN model

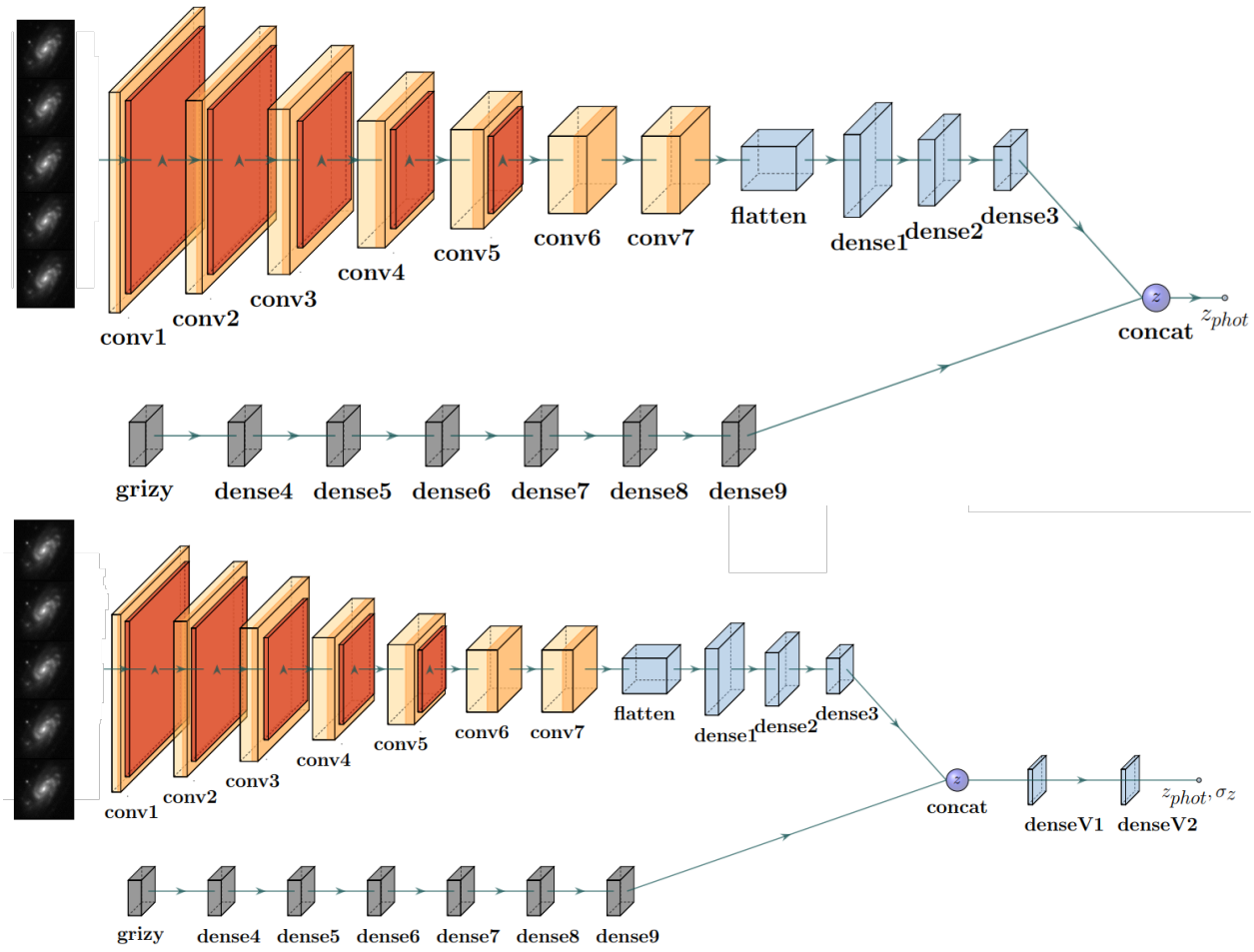

Figure 4.4 TOP: CNN architecture. BOTTOM: BCNN architecture. The inputs for both networks are five-band galaxy images and photometry in the g,r,i,z,y filters. The light orange boxes represent convolutional layers and the dark orange boxes represent maxpooling layers. 'denseV' refers to denseVariational layers. The output for the CNN is a single point photo-z estimate while the output for the BCNN is a photo-z PDF. We assume Gaussianity in the creation of the photo-z PDF, so a photo-z uncertainty is produced by the standard deviation of the PDF.

developed in (Jones et al., 2022a) and the other CNN network was developed using VGGNet (Simonyan & Zisserman, 2015). We found that transforming a NN to obtain a CNN was indeed a powerful way to produce a well-performing CNN. Additionally, we found that transforming the CNN into a BCNN by adding additional variational layers at the end of the network was a straightforward way to produce a well-performing BCNN. Ultimately we found the CNN and BCNN models inspired by VGGNet to be superior to the CNN and BCNN that were produced by transforming a NN model, so the results discussed in this work were obtained with the VGGNet-inspired models (Fig. 5.8).

We performed a hyperparameter grid search that iterated over a number of hyperparameters:

- # layers
- types of layers
- # nodes per layer
- learning rate
- loss function
- activation functions
- # epochs
- image pixel scaling
- batch sizes
- kernel sizes

A high-performing model requires a delicate balance between all hyperparameters. The most sensitive hyperparameters were the learning rate, type of layer, and the loss func-

Hyperparameter	**Value**
Number of Layers	24
Types of Layers	Convolutional, Pooling, Dropout, Flatten, Dense
Number of Nodes per Layer	32-512
Learning Rate	0.0001
Loss Function	RMSE
Activation Functions	ReLU, tanh
Number of Epochs	200
Image Pixel Scaling	0-1
Batch Size	256
Kernel Sizes	3x3
Pool Sizes	2x2

Table 4.2 CNN hyperparameters

tion. Compared to the CNN, the BCNN has additional hyperparameters that influence the probabilistic portions of the network:

- # variational layers
- default scale of posterior mean field normal distribution
- gaussian initializer mean and stddev

. The hyperparameter values that were optimized for the CNN do not necessarily translate to optimal values for the hyperparmeters of the BCNN. BCNNs have higher variance than CNNs during model hyperparameter tweaking, generally. A major factor in this relationship is that the weights learned during the training process are different for a CNN and BCNN. For the CNN, weights are discrete and thus the output is deterministic. For the BCNN, weights learned in the training process are actually uncertainty distributions (see Fig. 4.3).

We found that simply replacing dense layers with probabilistic convolutional or variational layers did not translate into an effective model. When all dense layers are variational layers, the estimated uncertainties had poor statistical coverage with much higher uncertainties than desired. We found that a probabilistic model generally performs best when the

Hyperparameter	**Value**
Number of Layers	26
Types of Layers	Convolutional, Pooling, Dropout, Flatten, Dense, Dense Variational
Num. of Nodes per Layer	32-512
Learning Rate	0.0001
Loss Function	Negative Log Likelihood
Activation Functions	ReLU, tanh
Number of Epochs	200
Image Pixel Scaling	0-1
Batch Sizes	256
Kernel Sizes	3x3
Pool Sizes	2x2
Number of Variational Layers	2
Gaussian Initializer	Mean: 0, Stddev: 0.1

Table 4.3 BCNN hyperparameters

bulk of the model is non-probabilistic and the final one or two layers are probabilistic.

Due to their probabilistic nature, BNNs do not produce the same prediction given the same data nor does the model result in the same weights each time it is trained. The weights in the variational layers are sampled from a Gaussian distribution, resulting in variations in the predictions. The results presented here are representative of a typical training run with the BCNN model presented in this work. However, there can be variations of several percents in outlier rates and other metrics depending on the training run. There can also be variations in the final loss achieved at the end of training.

Another notable finding during model optimisation in this analysis is that the CNN model almost immediately began producing acceptable photo-z predictions using galaxy images alone, while most variations of the BCNN required both galaxy images and galaxy photometry as inputs in order to achieve similar performance. One explanation for this observation is that images contain all of the information present in the magnitudes, but the architecture is not sufficient for extracting that information. BCNN has additional degrees

of freedom it needs to fit compared to CNN, so it required significantly more tweaking during the architecture construction and hyper parameter optimisation process.

4.3.4 The impact of photometry

While we use both photometry and images in our final model, we find that including the photometry only has a small impact on the performance of our final model (1-2% differences in LSST metrics (bias, RMS, 3sigma outlier)). We find that in the early stages of model optimisation of image based networks for photo-z estimation, utilizing galaxy photometry in combination with galaxy images boosts overall model performance and significantly reduces the probability of a loss function diverging during the training process with a probabilistic model. As the model is systematically optimized through a hyperparameter grid search and the loss function is further improved, we find that the use of galaxy photometry is less impactful. Our final CNN and BCNN models presented in this work utilized both photometry and images because we still found a 1-2 % boost in performance with respect to the metrics specified in Table 4.

4.4 Conformal Prediction

To ensure good statistical coverage of the BCNN, we use conformal prediction to rescale the predicted uncertainties. Conformal prediction is a promising method for uncertainty quantification that is agnostic to the method of photo-z prediction and does not need to assume a probability distribution (Papadopoulos et al., 2002; Vovk, 2012; Lei & Wasserman, 2014). It works by including an extra calibration step after a model is trained to create credible intervals and maintain exact statistical coverage. Given a required credible interval (ie. 90% coverage), this calibration step allows us to determine how a given prediction score maps to the range of predicted values that has that credible interval. For Bayesian models like the BCNN, we can calibrate how to scale the predicted variance to ensure exact coverage

(Hoff, 2021). In addition, this method can add uncertainty quantification to networks (like CNNs using quantile loss) that previously only supported point predictions by mapping between prediction scores and statistical coverage (Angelopoulos & Bates, 2022).

Because a dominant source of training uncertainty results from the inconsistent population of training galaxies as a function of redshift, we apply the conformal prediction analysis to individual redshift bins themselves rather than the entire dataset. We implement a binned approach to conformal prediction where we use a calibration dataset with known spectroscopic values and separate galaxies into spectroscopic redshift bins of $z = 0.1$. In each bin we calculate a nonconformity score

$$S = \frac{|z_{phot} - z_{spec}|}{\sigma_z}$$

and calculate the quantile of nonconformity scores using the desired coverage of 0.683

$$q = Q(S, 0.683)$$

which are used to scale the uncertainties associated with evaluation set galaxies. We divide evaluation set galaxies into bins based on their photometric redshifts (i.e 40 bins if galaxies have a redshift range from $0 < z < 4$), and we scale their uncertainties by the corresponding quantile scaling that was calculated from the calibration dataset spectropic redshifts.

$$\sigma_{z,f} = Q_i(S, \alpha) * \sigma_{z,i}$$

where $Q_i(S, \alpha)$ is the quantile scaling parameter calculated from the calibration data set in bin i, $\sigma_{z,i}$ is the photometric redshift uncertainty produced by the BCNN for an evaluation galaxy in bin i, and $\sigma_{z,f}$ is the final photometric redshift uncertainty for a galaxy in bin i.

4.5 Other photo-z ML models for comparison

We use five other common ML models discussed in (Jones et al., 2023) in order to compare to the CNN and BCNN performance: (1) a neural network from Jones et al. (2023), (2) a BNN from Jones et al. (2023), (3) a support vector machine (SVM) classification model, (4) a random forest regression (RF) model, and (5) a gradient boosted tree regression model. We perform a hyperparameter grid search for each model. See (Jones et al., 2023) for a larger discussion of each model. The non-neural network models were chosen because they are commonly used in photo-z predictions in the past [cite] and serves as a good baseline for comparison.

4.6 Photo-z metrics

A chief goal of this work is to prepare for the upcoming cosmological experiments from data release from large scale surveys like LSST in order to optimally extract scientific information from the data and use those insights to constrain cosmological parameters. Therefore, our choice of metrics to evaluate the photo-z predictions in this work is focused on the scientific requirements as set out by the LSST science requirements document Collaboration et al. (2021) and discussed in §1. As highlighted there, the three main requirements for photo-z measurements for the purpose of constraining dark matter and dark energy are: RMS error (< 0.2, Eq. 3), Bias (< 0.003, Eq. 4), and 3σ Outliers ($< 10\%$, Eq. 7).

We also include in our analysis a number of point metrics that are commonly used in the photo-z literature (Outlier (Eq. 1), Catastrophic Outlier (Eq. 2), Scatter (Eq. 5), and Loss (Eq. 6)) for the purpose of comparison to other models, as well as additional probabilistic metrics to evaluate the photo-z uncertainties produced by the BCNN (Section 4.6.1). The RMS photo-z error is given by a standard definition in Eq. 3, where n_{gals} is the number of galaxies in the evaluation testing set and Σ_{gals} represents a sum over those galaxies. Bias and scatter are defined in Eqs. 4 and 5. We follow Tanaka et al. (2018b) and define a loss

function in Eq. 6 to characterize the point estimate photo-z accuracy with a single number, where we use $\gamma = 0.15$.

Ideally, photo-z measurements should be accurate out to the redshift limit of LSST observations ($\sim z = 3.4$ is where galaxies begin dropping out of the g band), however the main redshift range of focus is $0.3 < z < 3.0$ and upcoming weak lensing analyses focus on the range $0.3 < z < 1.5$. In the range $0.3 < z < 3.0$, LSST aims to measure the comoving distance as a function of redshift to an accuracy of 1-2%. In order to achieve this goal, LSST must obtain (1) a sufficiently large sample of galaxies ($\sim$ four billion) and (2) sufficiently accurate photo-z measurements for these galaxies as defined by the aforementioned requirements. In addition to meeting photo-z science requirements, the LSST team also requires 'methods for rejecting the majority of those outliers, and for characterizing their effects on the sample'.

We therefore evaluate our models as a function of redshift. In weak lensing and other cosmological analyses, science requirements for photo-z estimates must be achieved on average throughout each tomographic redshift bin, rather than on average throughout the entire sample. This means that a full evaluation of a particular photo-z method must include an evaluation of important metrics as a function of redshift, rather than averaging across the entire photo-z sample. This distinction is particularly important for evaluating model performance of high redshift regions ($z > 1.0$), which contain significantly fewer galaxies than low redshift regions (see Fig. A.2), and are thus more challenging for any photo-z method to accurately produce photo-zs.

We also include additional metrics that quantify outliers in multiple ways. We use the conventional definition for photometric redshift outliers and catastrophic outliers in Eqs. 1 and 2, where z_{phot} and z_{spec} are the estimated photo-z and actual (spectroscopically determined) redshift of the galaxy.

Table 4.4 Metrics used to assess model performance.

Point Metrics		Probabilistic Metrics	
Outlier	$O : \frac{\lvert z_{phot}-z_{spec}\rvert}{1+z_{spec}} > .15$ (1)	3σ Outlier:	$\lvert z_{phot} - z_{spec}\rvert > 3z_{\sigma}$ (7)
Catastrophic Outlier	$O_c : \lvert z_{phot} - z_{spec}\rvert > 1.0$ (2)		
RMS error	$\sqrt{\frac{1}{n_{gals}}\Sigma_{gals}\left(\frac{z_{phot}-z_{spec}}{1+z_{spec}}\right)^2}$ (3)	Coverage	$\sum_{i}^{n_{gals}} \frac{(\bar{z}_{pdf,i} - z_{spec,i}) < z_{\sigma,i}}{n_{gals}}$ (8)
Bias	$b = \frac{z_{phot}-z_{spec}}{1+z_{spec}}$ (4)		
Scatter	$\text{Median}(\lvert\Delta z - \text{Median}(\Delta z_i)\rvert)$ (5)	PIT:	$\int_{-\infty}^{z_{spec}} p(z)dz$ (9)
Loss	$L(\Delta z) = 1 - \frac{1}{1+(\frac{\Delta z}{\gamma})^2}$ (6)		

4.6.1 Probabilistic Metrics

We use coverage as a key metric for assessing the performance of the BCNN (see Eq. 8). Coverage is typically used to assess whether confidence intervals are accurate. In this case, we define coverage as the fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If the coverage is over 68%, then the estimated uncertainties are on average too large. Similarly, if the cover is below 68%, the estimated uncertainties are on average too small. Fig. 4.9 depicts the photo-z statistical coverage for the evaluation set before and after conformal prediction is applied to the photo-z uncertainties.

Error in the bulk photo-z distribution width for the evaluation set can be difficult to distinguish between uncertainties associated with galaxy bias or uncertainties in the mean redshift of photo-z tomographic bins. The Probability Integral Transform (PIT) is a photo-z metric that can detect systematic error in the photo-z distribution width for galaxy samples with known spectroscopic redshifts (Malz & Hogg, 2020; Malz, 2021). The PIT value for a single galaxy is defined in Eq. 9 in Table 4, where $p(z)$ is the predicted photo-z PDF. A histogram of PIT values for a galaxy sample should be uniform for an accurate collection of $p(z)$ samples. Ideally, the PIT histogram is flat across all redshift bins. If the PIT histogram peaks at the center, the $p(z)$ collection is too broad. If the PIT histogram peaks at high and low PIT values, the $p(z)$ samples are too narrow. For a comparison of several probabilistic photo-z methods, see Schmidt et al. (2020b).

Figure 4.5 The fraction of galaxies that have a spectro-z within their 68% confidence interval before and after conformal prediction analysis. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies. The BCNN demonstrates accurate coverage throughout the redshift range after conformal prediction analysis.

4.7 Results

We find that the image-based CNN and BCNN have less scatter and fewer outliers and catastrophic outliers than the non-image based models as a function of spectroscopic redshift. Fig. 4.6 shows the predicted compared to the spectroscopic redshifts for the 9 models considered in this work evaluated on the $\sim 28,000$ galaxies in the testing dataset. The NN, BNN, and CNN have generally less scatter and smaller number of outliers in their predictions across the redshift range of the test sample compared to the other 5 models we consider. The NN and BNN models using only photometry have larger scatter and more outliers than compared to the CNN and BCNN models using images (see below for a quantitative comparisons). The non-probabilistic NN and CNN models have predictions that are systematically higher than the spectroscopic redshift for a small group of galaxies with spectroscopic $z \sim 1$. The BNN and BCNN do not not appear to have a similar systematic error in their prediction. The random forest model has the smallest RMS scatter at $z < 1.5$ (even compared to the CNN and BCNN models), but there are a large number of outliers, especially at higher redshifts. The random forest and gradient boosting models also systematically predict lower redshifts for galaxies with spectroscopic $z > 1.5$.

Of the neural network models, the image-based CNN and BCNN perform significantly better with respect to the fraction of outliers and RMS error (see Fig. 4.7). For this comparison, we use the BCNN model that filters out all sources with predictions have have redshift uncertainty σ_z <0.3 (See Section 4.7.1 for more about this choice.) The NN and

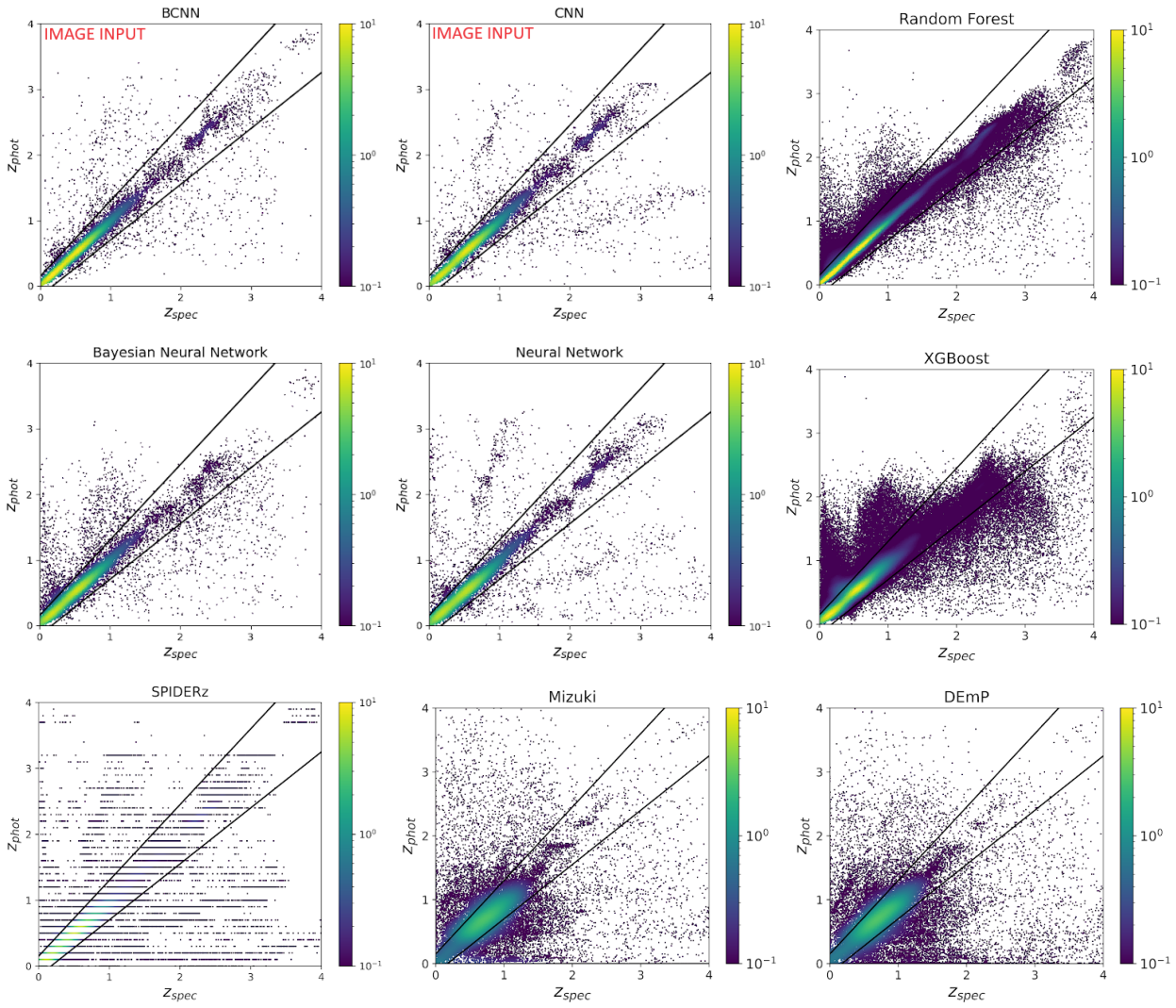

Figure 4.6 Visualization of predicted photo-zs versus measured spectroscopic redshifts by the models discussed in §2. The results of these determinations are quantified in Table 5. The colorbars indicate the density of evaluation data points as computed with a Gaussian kernel-density estimation.

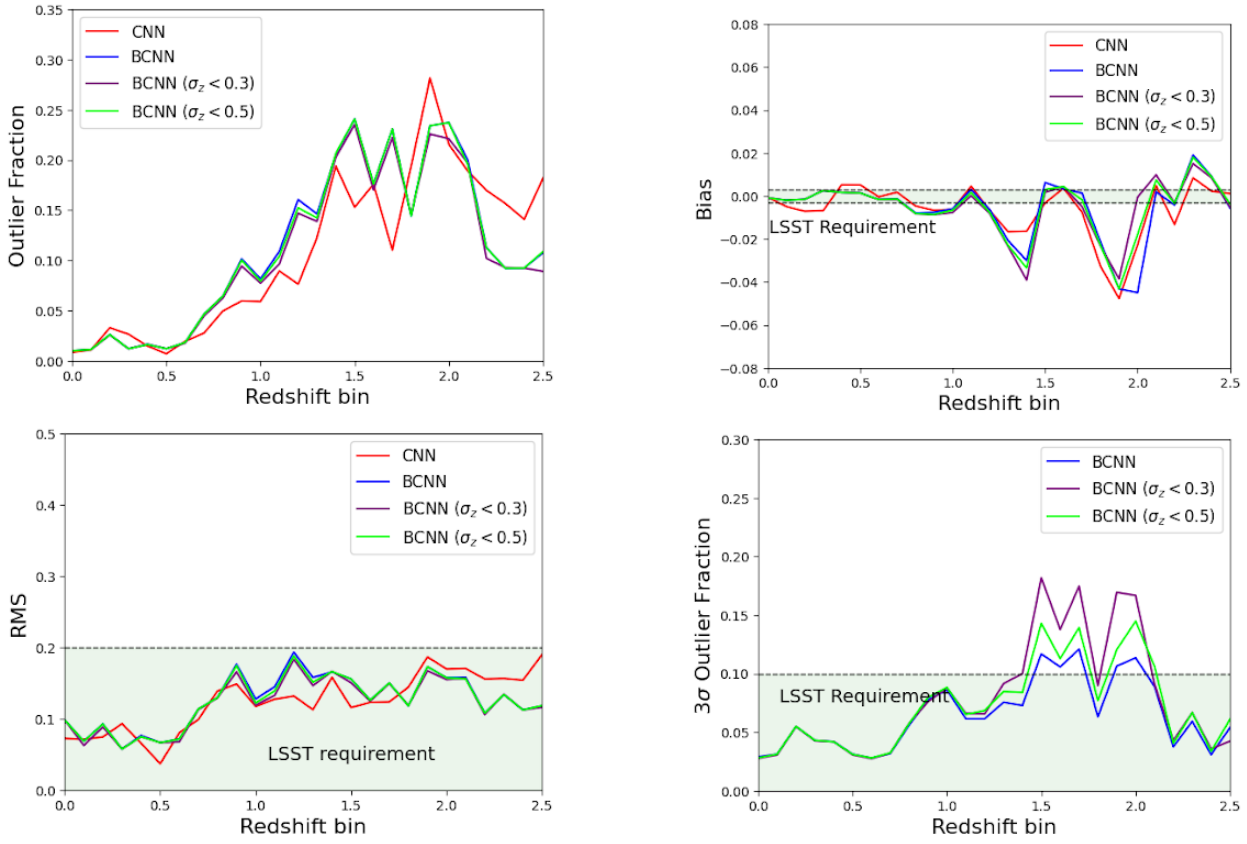

Figure 4.7 BCNN and CNN performance with respect to LSST photo-z requirements. We note that the 3σ outlier fraction can only be calculated with the BCNN because the metric requires photo-z uncertainties so we additionally include the standard outlier fraction for the CNN and BCNN for comparison. The plots reflect results with 70% of galaxies for training, 10% for validation, 10% for parameterisation of the conformal predictions, and 10% for evaluation. We include only those results in the redshift range $0 < z < 2.5$ because the $N(z)$ distribution of the data set degrades significantly at higher redshifts (see Fig. A.2) and would likely significantly improve given sufficient training data.

BNN have a factor of 2 to 4 greater number of outliers than the BCNN for $z < 1.5$. In this same redshift range, RMS error is a factor of 2 to 3 worse for the NN and BNN compared to the BCNN. All the neural network models, regardless of images or photometry, perform roughly the same with respect to bias. The 3σ outlier fraction is about 0.2 to 0.4 times lower for the BNN compared the BCNN even though the absolute outlier fraction is lower. We attribute this to the fact that the uncertainties are smaller for the BCNN models, which

increases the likelihood that sources will be identified as 3σ outliers.

The BCNN model achieves the LSST science requirements for photo-z estimates for RMS $z < 2.5$, 3σ outlier fraction for $z < 1.5$, and bias for $z < 1.1$. The CNN and BCNN models have comparable performance for the non-probabilistic metrics and generally outperform the non-image based models across all metrics (see Table 5 and Figs. 4.6 and 4.7). The BCNN generally satisfies LSST photo-z science requirements in the range of $0.3 < z < 1.5$ (redshift range for weak lensing analyses – see Fig. 4.7) and performs as well or better than the 8 alternative methods investigated in this study (see Table 5 and Figs. 4.6 and 4.7). The most difficult science requirement metric to satisfy is bias < 0.003, which is only met up to $z < 1.1$ with the BCNN. On average the performance of the BCNN with respect to bias is fairly constant for $0.1 < z < 1.1$, with an average bias of -0.00054 in this range. We believe that increasing the relative population of the training sample beyond $z = 1.1$ will improve the performance at larger redshifts. The RMS scatter satisfies the LSST requirement of 0.2 for $z < 2.5$. The scatter is about 0.08 for $z < 0.5$ and increases to about 0.15 for $z > 0.5$. The BCNN gives us uncertainty predictions, which allows us to identify 3 σ outliers. The fraction of 3σ outliers is about 5% for $z < 0.5$ and below 10% for $z < 1.5$. The raw redshift and uncertainty estimates from evaluating the CNN and BCNN models on the evaluation set are available at `https://zenodo.org/doi/10.5281/zenodo.10145347`.

The strong dependence in the model performance with redshift is likely due to the distribution of training data that is biased towards lower redshift samples (see Fig. A.2). This type of imbalanced training is a known problem with machine learning models. If the training data is unrepresentative of the true galaxies from large scale surveys, then the results may be biased. For example, the spectroscopic sample is brighter than the imaging sample of galaxies. To examine this effect, we resample the evaluation dataset so that it better resembles the distribution of brightness of the HSC imaging survey. We find the resampled dataset reduces the performance of the model between 5 to 30%, depending on the metric for $0.1 < z < 1.5$. The outlier fraction is the most affected, increasing from 3.9% to 5.99%

for the BCNN and increasing from 4.1% to 6.2% for the CNN in the resampled data set. The RMS increases from 0.098 to 0.1087 for the BCNN and increases from 0.0996 to 0.124 for the CNN in the resampled data set. The absolute bias is improved from 0.0007 to 0.0003 for the BCNN, but worsened for the CNN from 0.001 to 0.0029 in the resampled data set. See Appendix A for more details.

4.7.1 Leveraging Photo-z Uncertainties for Outlier Identification and Improving Performance

An important advantage of the BCNN is that the uncertainties produced by the model can be used as an indicator of potentially poor predictions. This method was proposed in Jones et al. (2023) and demonstrated for BNN models. In this method, all galaxies with a photo-z uncertainty greater than the specified σ_z cutoff value are flagged as potential outlier or catastrophic outlier candidates. Depending on the goal, one can consider removing these sources. An acceptable balance needs to be achieved between the number of galaxies correctly flagged as poor predictions versus the number of non-outlier galaxies removed for a given σ_z cutoff value.

We find a significant reduction in the number of outliers and catastrophic outliers by sacrificing a minimal number of non-outlier predictions using the uncertainty estimates of the BCNN. We tested different σ_z cutoff values between 0.2 and 2 to determine how this filtering affects the photo-z metrics. We find that a cutoff of $\sigma_z > 0.3$ to be a good compromise in removing many outliers while removing a minimal number of non-outliers. By selecting a photo-z uncertainty cutoff of $\sigma_z > 0.3$, the RMS error was reduced by 32.5%, outliers were reduced by 44%, and catastrophic outliers were reduced by 55.6% – at the cost of removing only 4% of the evaluation set. See Fig. 4.12 for the N(z) distribution of removed galaxies for example cases of $\sigma_z > 0.3$ and $\sigma_z > 0.5$. This result is similar to the results in Jones et al. (2023) using a BNN on the same data.

The BCNN with the outlier removal method stands out as the overall best performing

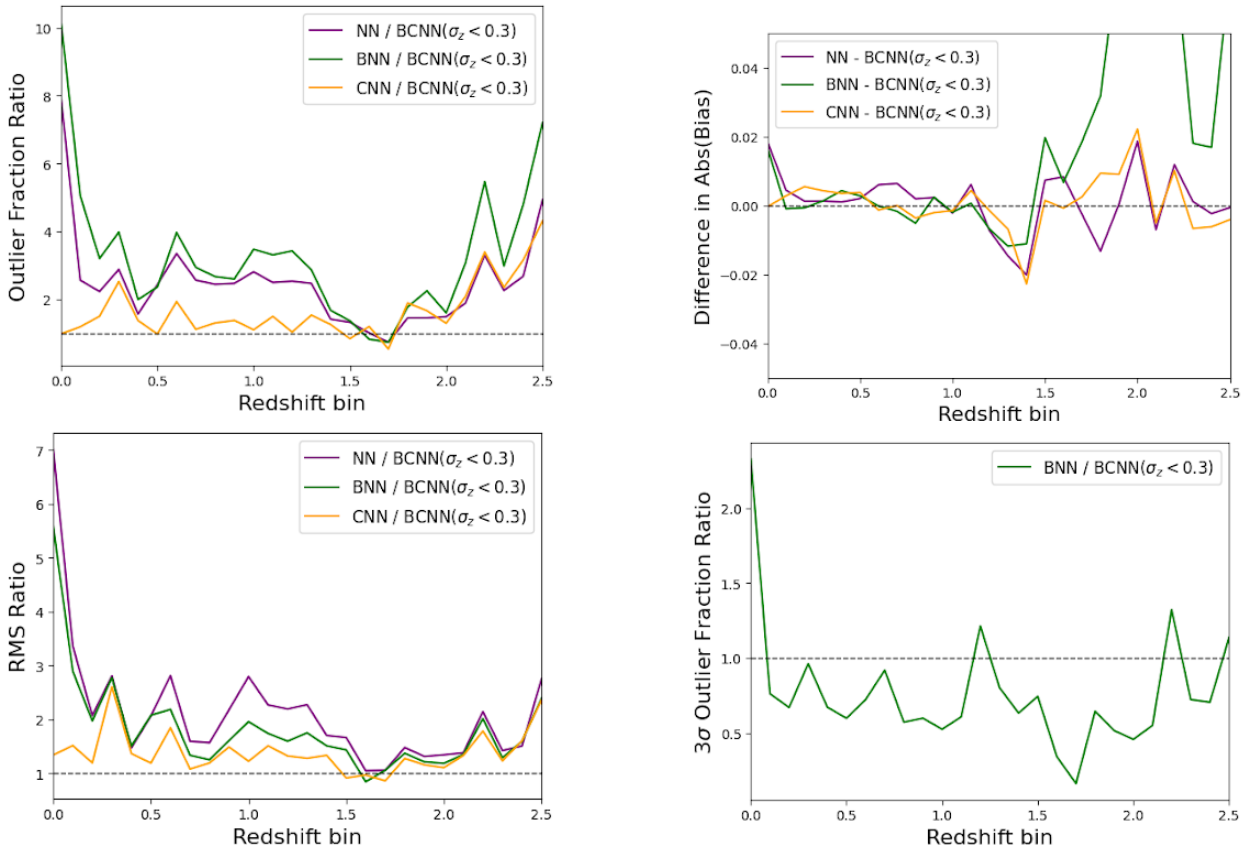

Figure 4.8 A comparison of the performance of all four neural network models relative to the BCNN (after removing all galaxies where $\sigma_z > 0.3$) using the LSST science requirement metrics and the conventional outlier fraction. For the 3σ outlier fraction, we could only include the BNN and BCNN models because the NN and CNN are non-probabilistic.

model for the majority of photo-z performance metrics considered in this work, achieving the lowest percentage of outliers, Bayesian outliers, and RMS error. Performance improvements with example σ_z removal values for a variety of performance metrics, including the LSST photo-z requirements, are visualized in Figs. 4.7, 4.9, 4.10, and 4.11.

4.7.2 Bayesian Convolutional Neural Network Photo-z Uncertainty Estimates

We find that the BCNN produces accurate uncertainties as defined by the probabilistic metrics. The quality of the uncertainties produced by the BCNN are visualized in Figs. 4.7,

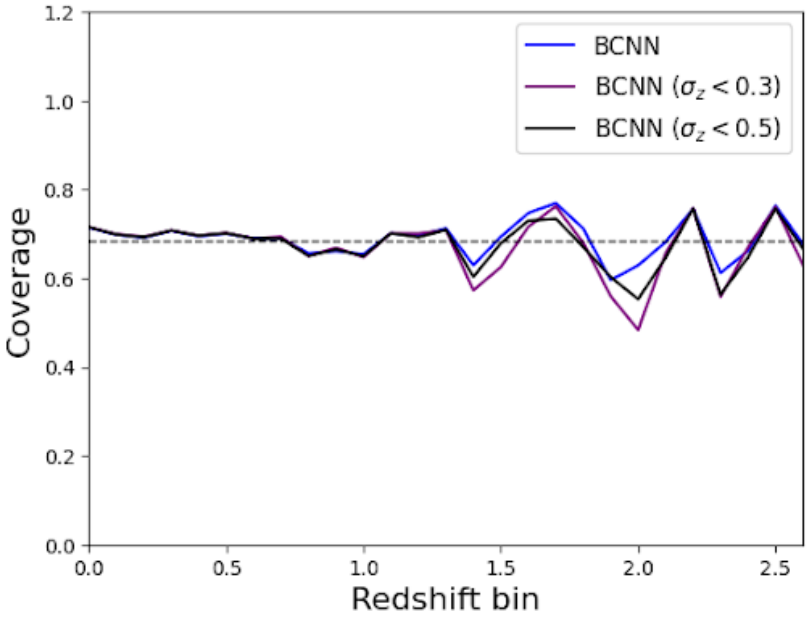

Figure 4.9 The fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies. The BCNN demonstrates accurate coverage throughout the redshift range.

4.9, 4.10, and 4.11. The BCNN 3σ outlier fraction is shown in Fig. 4.7, which indicates that uncertainties are generally well-estimated on average across the redshift range $0 < z < 2.5$. It is notable that the BCNN performs best with respect to the σ outlier fraction when no galaxies with large uncertainties are removed. The BCNN uncertainty coverage of the sample is provided in Fig. 4.6.1, showing acceptable agreement with the target 68% confidence interval up to the target redshift interval for weak lensing applications $0.3 < z < 1.5$, indicating the uncertainties of photo-z estimates for this galaxy population are accurately defined. The PIT histogram produced for a sample determination with the BCNN is shown in Fig. 4.13. The PIT histogram is generally flat, as is desired, however the slight bump in the middle indicates that some of the photo-z PDFs tend to be overly broad, while the peaks at the edges of the distributions indicate that some of the photo-z PDFs are overly narrow.

For a comparison of PITs produced by other probabilistic photo-z methods (performed on different data) see Schmidt et al. (2020b).

4.8 Discussion

In this work we find that using images to predict photometric redshifts can satisfy the LSST science requirement metrics up to $z < 1.5$ and perform significantly better than non-image methods. This is important because while the computational cost is much higher for images than for photometry, the images offer additional critical information for accurate redshift predictions. Incorporating images using models like a BCNN in photometric pipelines for large scale surveys has the potential to provide significant scientific gains. For example, the outlier fraction using the image based BCNN is 2-4 times better than the best performing non-image model on the same data.

We will now compare the results here to select results obtained in other works with other datasets. We note that a perfect comparison between photo-z models requires identical training, validation, and evaluation data sets. In the following comparisons, we aim to compare photo-z metrics over similar redshift ranges, but a more definitive comparison would need to train the other models with our specific dataset.

The BCNN model in this work can outperform previous CNN models and has the advantage of providing uncertainties. The work with the most similar data and model is the photo-z investigation performed by Schuldt et al. (2021) which utilized HSC imaging data and obtained a precision of $\Delta_z = |z_{phot} - z_{spec}| = 0.12$ with a convolutional neural network averaged over all galaxies in the redshift range $0 < z < 4$. We obtain $\Delta_z = 0.0031$ for the CNN and $\Delta_z = 0.0032$ for the BCNN averaged over all galaxies in our data set in this range. While the photo-z models from Schuldt et al. (2020a) and the HSC team utilized largely the same galaxy set that was used in this work, there are some differences between their data and the data used in this investigation, which introduces additional uncertainty in the

comparison made between results.

Recent image based photo-z investigations using CNN have been performed on low redshift (z < 0.5) Sloan Digital Sky Survey (SDSS) data in Pasquet et al. (2019) and Treyer et al. (2023). Both studies used SDSS *ugriz* images to train a CNN model to predict photometric redshifts. Pasquet et al. (2019) used 100,000 galaxies ($z < 0.3$) for training while Treyer et al. (2023) used 370,000 galaxies ($z < 0.5$). These galaxies were brighter than $r < 20$ mag. These studies found that CNNs have good performance over this redshift range in terms of bias and scatter. For example, Pasquet et al. (2019) achieve a bias of 0.0001 over $z < 0.32$, which meets the LSST photo-z science requirement threshold for bias but it does not satisfy the redshift range requirement. A direct comparision to our study is difficult due to the differences in the datasets and the redshift range sampled. Our study uses training data over a significantly larger redshift range ($0 < z < 4$) and includes galaxies 4 magnitudes fainter ($r < 24$ mag). However, our conclusions are consistent with their results in that images are a very promising way to improve photo-z estimates and can be used for science.

Our work shows that conformal prediction can be a powerful tool for improving the statistical coverage of photo-z uncertainties produced by BCNNs. The BCNNs tended to produce uncertainties that were too small, however conformal prediction provided a simple way to rescale the uncertainties to achieve the target statistical coverage without assumptions about the probability distribution of the predictions. We found the statistical coverage to work better when conformal prediction is applied separately to different redshift bins to achieve good statistical coverage for all redshifts. The reason for this is that the imbalance in the distribution of galaxies at different redshifts likely means that the model has different uncertainties at different redshifts. We note that the non-image based BNN in our previous work Jones et al. (2022a) achieved excellent statistical coverage without the use of conformal prediction.

The improved uncertainty estimates can also be used as a way to identify outliers in the redshift predictions (especially catastrophic outliers). The reduction of outliers is a

key objective outlined in the LSST photo-z science requirements document. When the uncertainty in the prediction is large, it is likely a reflection that input data is out of sample compared to the training that the model used to learn. Compared to non-probabilistic models like the CNN, the BCNN uncertainties provide additional insight into how the model interprets the data, which can be used to filter out poor predictions. Other outlier removal strategies that also rely on estimates of the probability information have previously been explored in Jones et al. (2023), Jones & Singal (2020), Wyatt & Singal (2020), Singal et al. (2022), and Pasquet et al. (2019).

4.9 Conclusion

In this work we present a probabilistic image-based photo-z estimation method utilizing data representative of future large scale surveys. We use conformal prediction to improve the uncertainties produced by the BCNN and evaluate the performance of the model relative to a CNN and 8 non-image based photo-z methods using LSST photo-z science requirement metrics. We also present results from utilizing the photo-z uncertainties to remove outlier predictions. We find that the BCNN produces uncertainties with excellent statistical coverage. We also find the BCNN performs significantly better than non-image based models and can satisfy the LSST science requirements in the redshift range of weak-lensing surveys (up to $z = 1.5$). We believe that these results indicate that image-based photo-z methods have potentially surpassed the performance abilities of current non-image based photo-z methods and therefore the development of image-based photo-z methods should be a priority for future large scale survey science missions. As we quickly approach the time period in which data from large scale surveys such as LSST are readily available, we hope the BCNN model and the techniques deployed in this work for improving photo-z uncertainties and predictions can serve as a useful framework for providing accurate photo-z predictions with well constrained uncertainties.

Table 4.5 Comparison of the performance results with each model discussed in §2. We use the data discussed in §4.3.1 to train and evaluate a NN, BNN, a support vector machine, (Cortes & Vapnik, 1995), a random forest (Breiman, 2001), and a gradient boosting model XGBoost (Chen & Guestrin, 2016). We also include a comparison to the template-fitting model, Mizuki, and empirical method, DEmP (Hsieh & Yee, 2014), that were evaluated on a larger, overlapping data set in (Nishizawa et al., 2020). To form a comparison to Mizuki and DEmP in this work, we crossmatched the larger data set with the object IDs of our data discussed in §4.3.1 to obtain a pre-evaluated sample of 60 thousand galaxies.

Network	O	O_c	O_b	RMS	$\|b\|$	Scatter	$L(\Delta z)$
CNN	0.041	0.0182	-	0.0996	0.001	0.018	0.061
BCNN	0.039	0.0135	0.021	0.098	0.0007	0.0162	0.058
BCNN($\sigma_z < 0.5$)	0.0275	0.008	0.0181	0.0758	0.0007	0.0158	0.0477
BCNN($\sigma_z < 0.3$)	0.02184	0.006	0.0153	0.0662	0.0007	0.0154	0.0421
BNN	0.079	0.023	0.023	0.174	0.013	0.026	0.105
BNN ($\sigma_z < 0.5$)	0.034	0.0071	0.025	0.0854	0.002	0.029	0.066
BNN ($\sigma_z < 0.3$)	0.0236	0.0045	0.017	0.0738	0.002	0.022	0.056
NN	0.059	0.029	-	0.174	0.0001	0.026	0.089
Mizuki	0.274	0.102	-	0.307	0.011	0.055	0.289
DEmP	0.250	0.092	-	0.277	0.003	0.040	0.258
RF	0.092	0.006	-	0.088	0.001	0.012	0.065
XGBoost	0.105	0.022	-	0.149	0.002	0.033	0.144
SPIDERz	0.090	0.051	-	0.199	0.002	0.044	0.135
LSST Req.	-	-	-	< 0.2	< 0.003	$<$ -	-

4.10 Acknowledgements

We thank Quanquan Gu for suggestions to use conformal predictions to improve uncertainty quantifications and Alex Malz for helpful discussions. Support for this work was provided by the Sloan Foundation and the UCLA Society of Hellman Fellows.

4.11 Appendix

4.11.1 Assessing redshift distribution biases in the dataset

We note there are multiple biases affecting modern photometric redshift training data. One such bias affecting the quality of spectroscopic redshifts results from the fact that the dominant source of ground truth for galaxy redshifts today are emission line galaxies; to improve results extending photo-z methods to large scale survey data, we need more spectroscopic data from absorption line galaxies. Additionally, the magnitude distribution of the data discussed in §2 affected by the requirement that all galaxies must have a measured spectroscopic redshift value, which imposes a bias toward bright galaxies compared to the bulk photometric sample from HSC. In order to address this, we have performed an additional analysis with the CNN and BCNN models using a re-sampled testing set that mimics the magnitude distribution of the bulk HSC photometry sample (Fig. 4.16). We use the g-band to re-sample our testing dataset to reproduce the overall HSC g-band distribution. Because our initial dataset size is limited to $\sim 300k$ galaxies, enforcing a strict re-sampling approach will diminish the re-sampled dataset size below an acceptable range for the purpose of the analysis in this work. Therefore, we perform a relaxed re-sampling method. The re-sampling process reduced our testing dataset size from 28,640 to 8,517 $--$ a 70.3% decrease. The distribution of redshifts for the resampled testing data is similar to the original (Fig. 4.17).

Overall, the BCNN and CNN models perform worse on the re-sampled data than the original testing data, but they still produce very accurate photo-z predictions (Figs. 4.18,4.19,4.20).

The coverage of the re-sampled data is not significantly affected (Fig. 4.19).Table 4.6 quantifies the performance metrics shown in Table 4.5.

Network	O	O_c	O_b	RMS	$\|b\|$	Scatter	$L(\Delta z)$
BCNN	0.039	0.0135	0.021	0.098	0.0007	0.0162	0.058
BCNN Re-Sampled	0.0599	0.0196	-	0.1087	0.0003	0.0276	0.090
CNN	0.041	0.0182	-	0.0996	0.001	0.018	0.061
CNN Re-Sampled	0.062	0.024		0.124	0.0029	0.0291	0.0896

Table 4.6 Comparison of the performance results with the CNN and BCNN with the original evaluation data set discussed in §4.3.1 and the re-sampled evaluation set to approximate the bulk HSC photometry. We use the data discussed in §4.3.1 to train. The re-scaling process reduces the initial evaluation set size from 28,640 to 8,517 −− a 70.3% decrease.

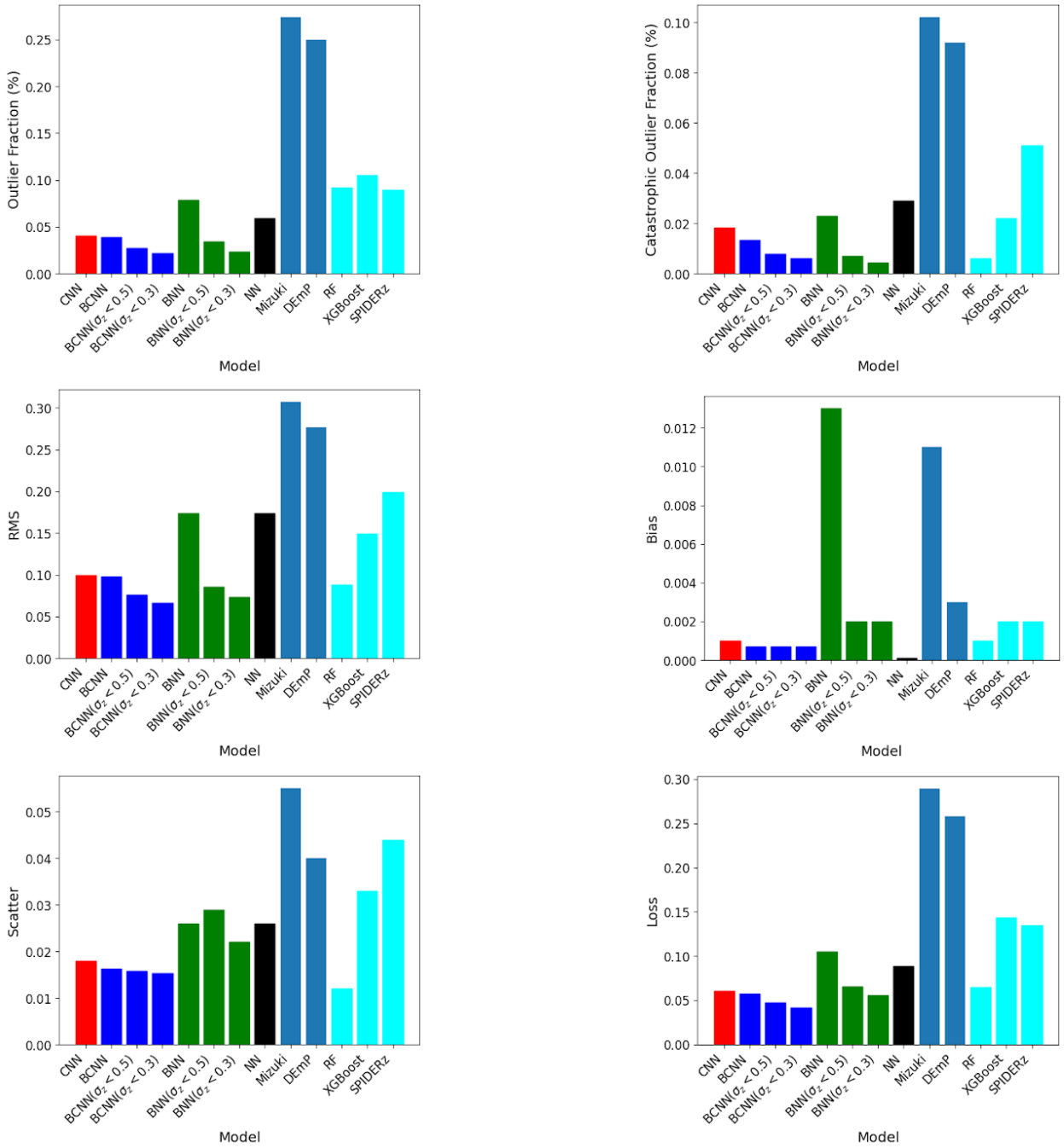

Figure 4.10 Comparison of the metric results achieved with each model. We use the data discussed in §4.3.1 to train and evaluate a CNN, BCNN, NN, BNN, the SPIDERz SVM (Jones & Singal, 2017), a random forest (Breiman, 2001), and a gradient boosting model XGBoost (Chen & Guestrin, 2016). We also include a comparison to the template-fitting model, Mizuki, and empirical method, DEmP (Hsieh & Yee, 2014).

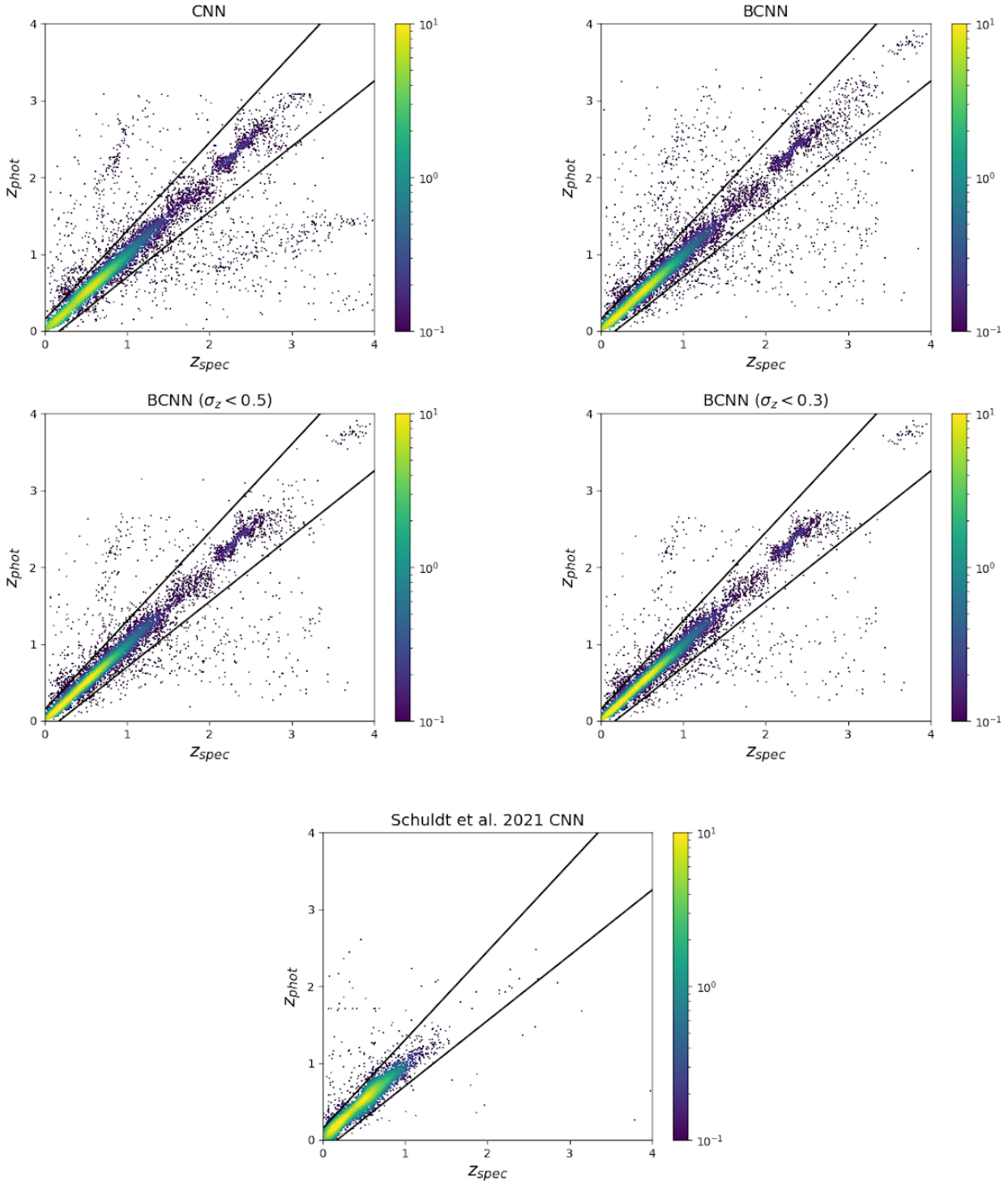

Figure 4.11 Visualization of CNN (top left) and BCNN (top right) performance compared to the BCNN with outlier removal criteria examples $\sigma_z = 0.5$ (bottom left) and $\sigma_z = 0.3$ (bottom right). We also include the results from Schuldt et al. (2020a) on overlapping data.

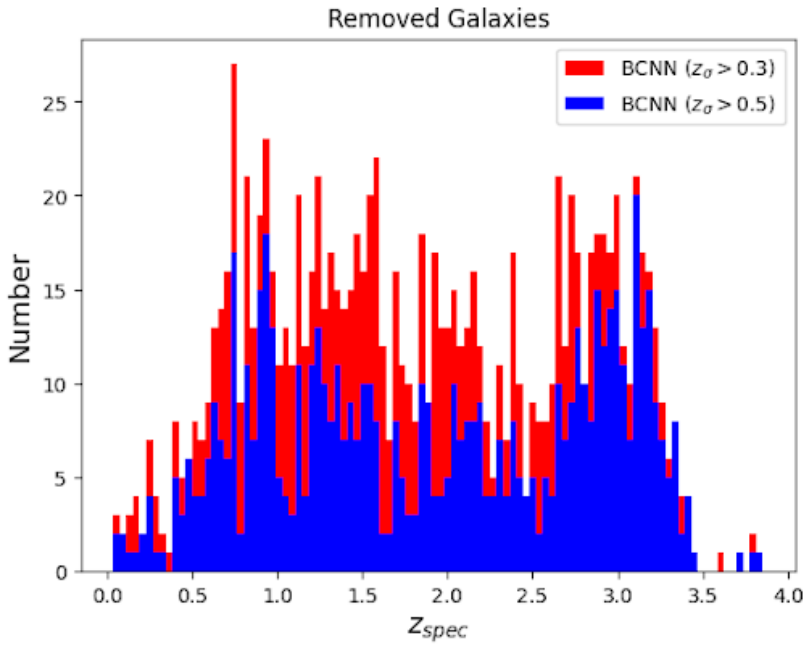

Figure 4.12 Histogram of photo-z uncertainties produced by the BCNN that exceed 0.3 and 0.5. By removing all galaxies in the evaluation sample with a photo-z uncertainty $\sigma_z < 0.3$, outliers were reduced by 70.1%, and catastrophic outliers were reduced by 80.43% – at the cost of removing 11% of the evaluation set. Using a photo-z uncertainty cutoff of 0.5 reduces the number of outliers by 70.1% and catastrophic outliers by 67.8% at the cost of removing 7.67% of the evaluation set.

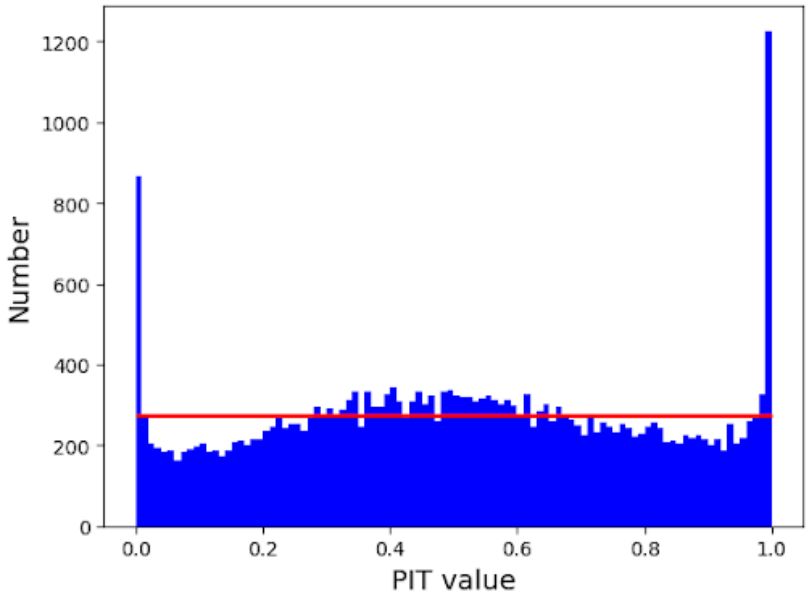

Figure 4.13 PIT histogram of the photo-z PDF produced by the BCNN. The red horizontal line indicates the ideal PIT histogram distribution: if the PIT histogram peaks at the center, the photo-z PDFs are generally too broad, and if the PIT histogram peaks at high and low PIT values, the PDF samples are too narrow or contain a large amount of catastrophic outliers.

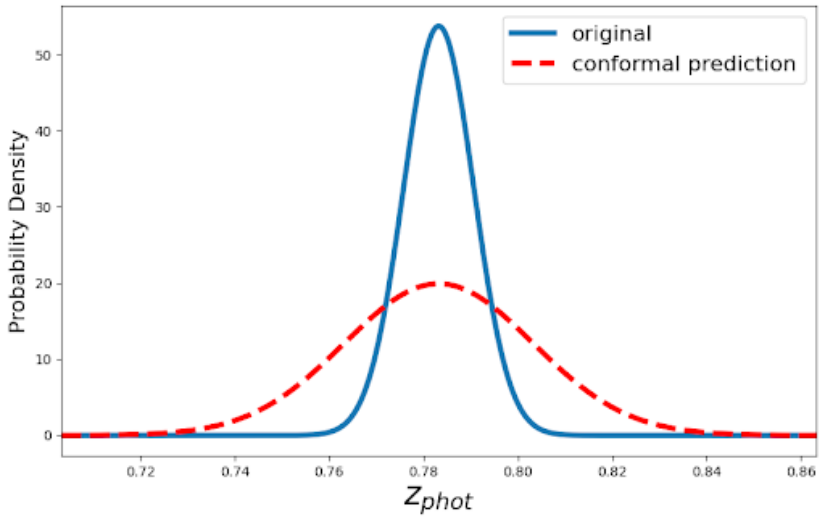

Figure 4.14 Visualisation of the photo-z probability distribution for an example galaxy in the evaluation set before and after conformal prediction transformation. The photo-z uncertainty in the original distribution was likely underestimated, which is why the conformal prediction transformation widened the width of the PDF.

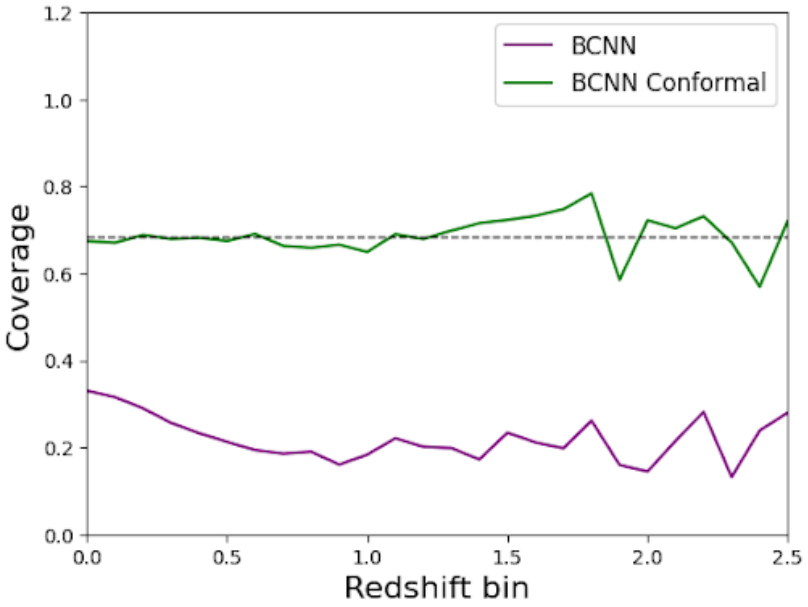

Figure 4.15 The fraction of galaxies that have a spectro-z within their 68% confidence interval for the original BCNN uncertainties and the adjusted uncertainties resulting from conformal prediction analysis. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies. The BCNN demonstrates accurate coverage throughout the redshift range.

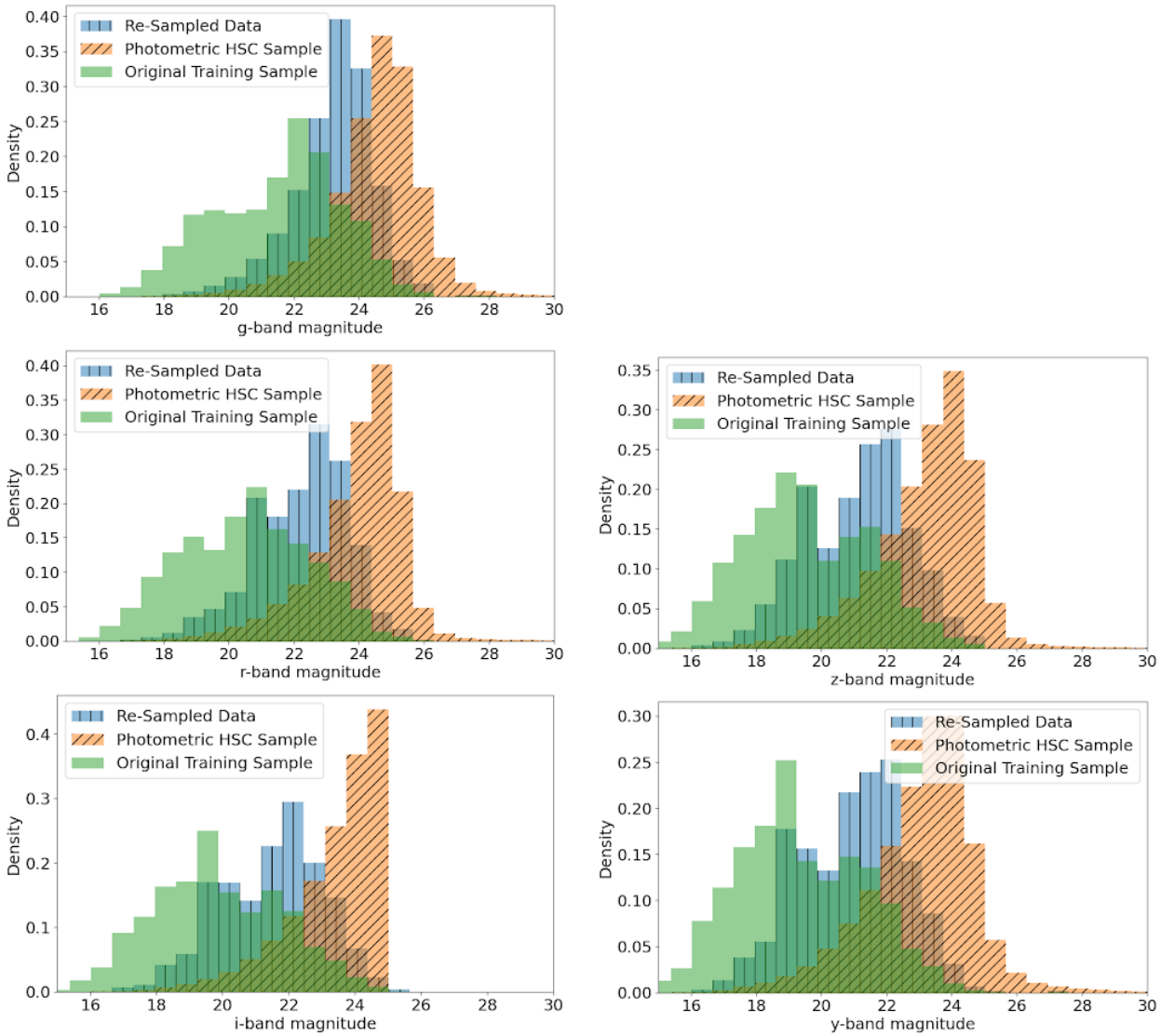

Figure 4.16 Visualisation of the *grizy* bands before and after the data is re-sampled to approximate the bulk HSC photometry sample. The green distribution is the original training sample for the dataset discussed in §2. The orange distribution indicates the HSC photometry sample with no spectroscopic bias. The blue distribution is a subset resulting from re-sampling the green distribution to more closely approximate the HSC photometric sample.

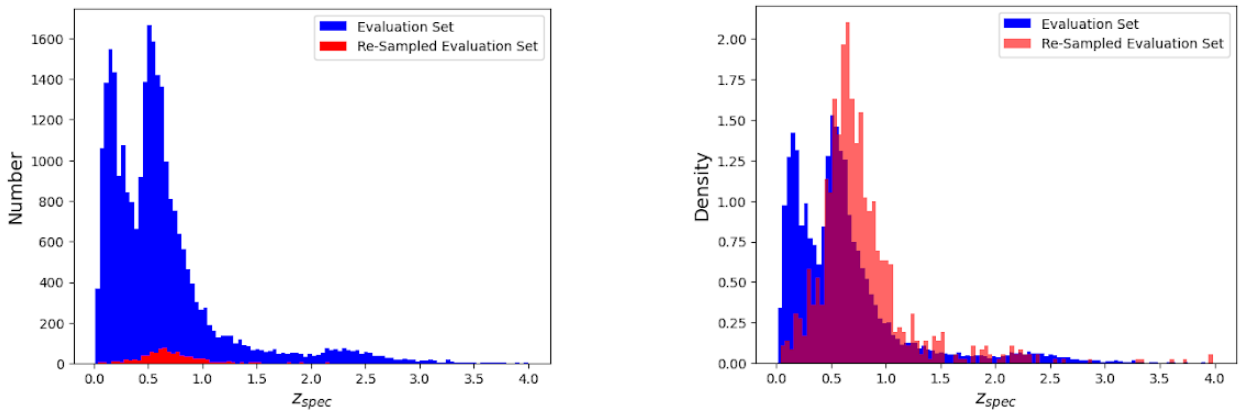

Figure 4.17 $N(z)$ distributions of the original evaluation set discussed in §4.3.1 and the re-sampled evaluation set.

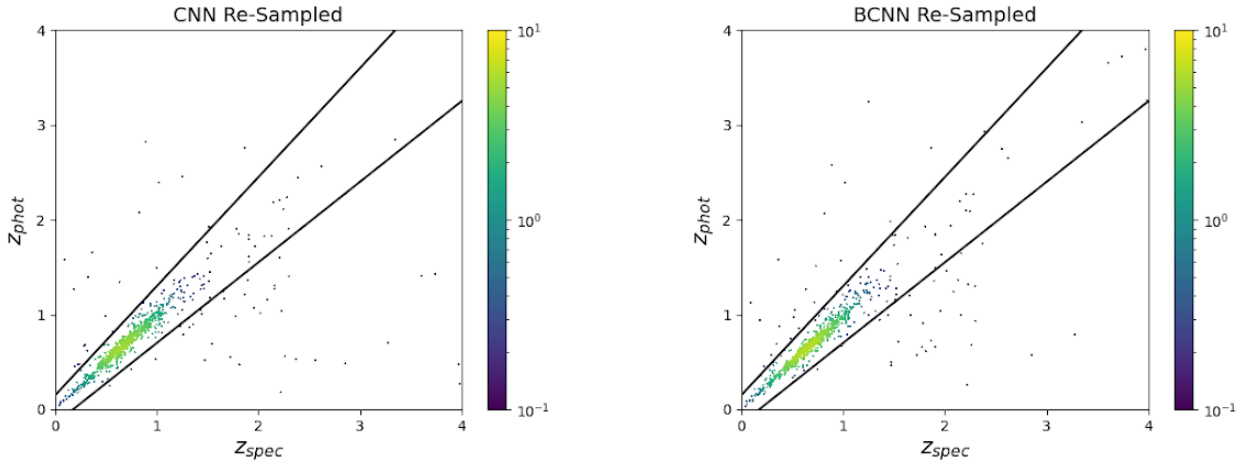

Figure 4.18 Visualization of the CNN and BCNN results using an evaluation set that is re-sampled to more closely approximate the bulk HSC photometry. The models are trained on the original data discussed in §4.3.1.

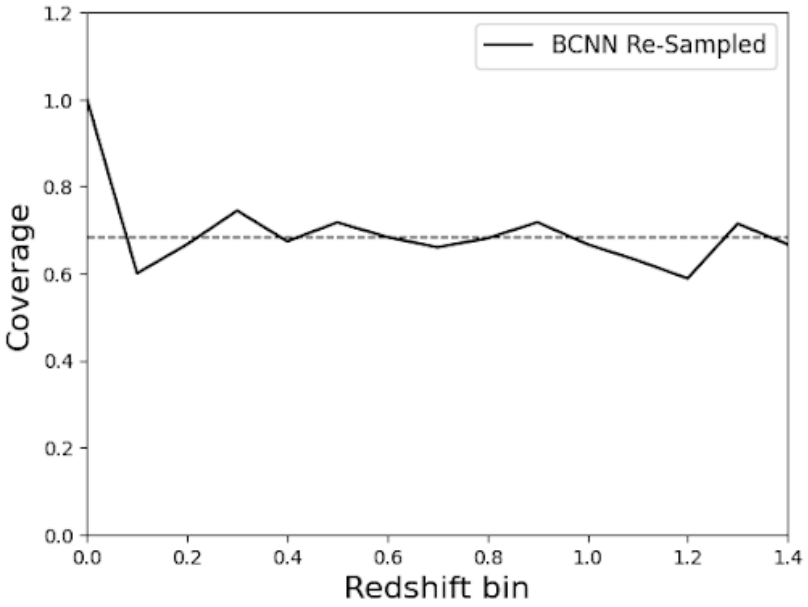

Figure 4.19 Coverage of the re-sampled evaluation sample. Coverage is defined as the fraction of galaxies that have a spectro-z within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true spectro-zs within their 68% confidence interval. If more than 68% of evaluated galaxies have spectro-zs within their 68% confidence interval, the galaxies are considered 'over-covered' because their photo-z uncertainties are too large. The same logic applies for 'under-covered' galaxies.

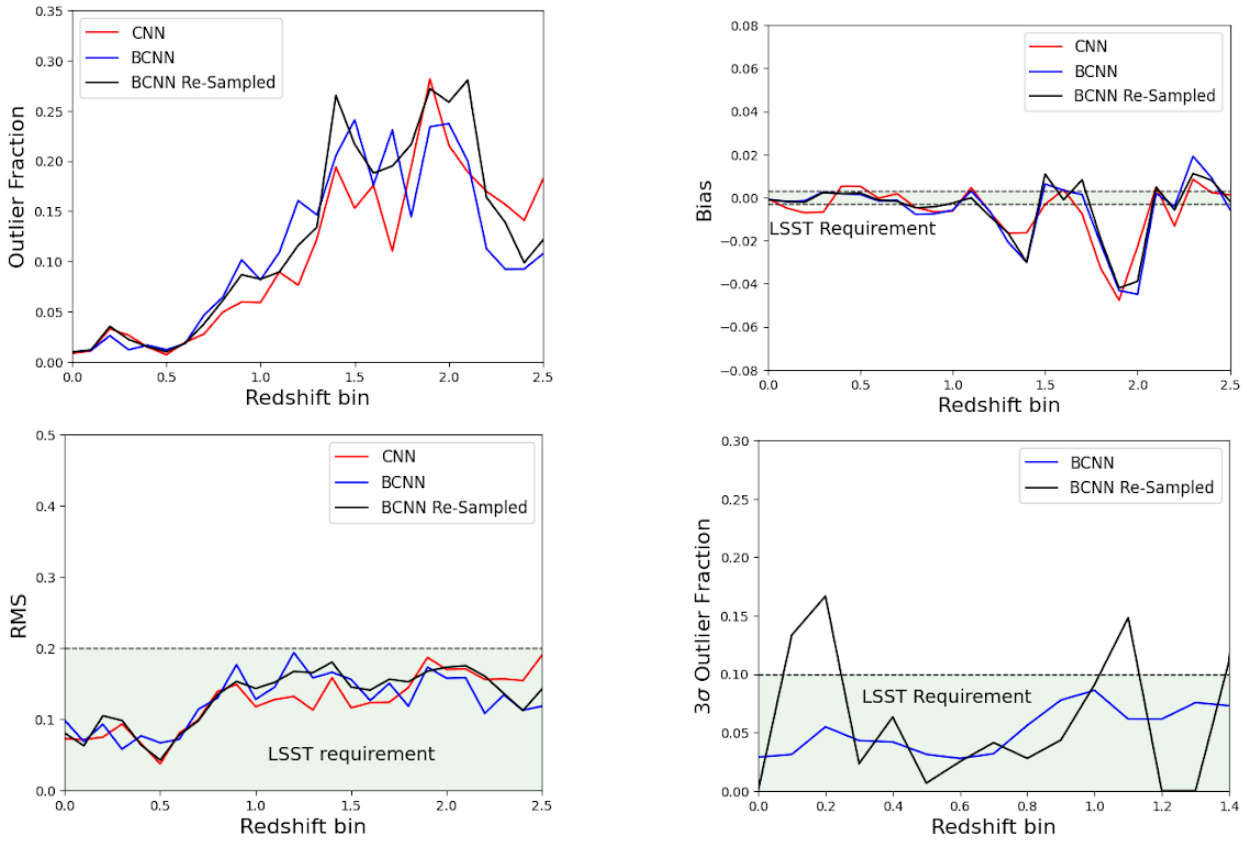

Figure 4.20 BCNN and CNN performance with respect to LSST photo-z requirements using an evaluation set with photometry that is re-sampled to approximate the bulk HSC photometry. We note that the 3σ outlier fraction can only be calculated with the BCNN because the metric requires photo-z uncertainties so we additionally include the standard outlier fraction for the CNN and BCNN for comparison. The plots reflect results with 70% of galaxies for training, 10% for validation, and 10% for evaluation. We include only those results in the redshift range $0 < z < 2.5$ because the $N(z)$ distribution of the data set degrades significantly at higher redshifts (see Fig. A.2) and would likely significantly improve given sufficient training data.

CHAPTER 5

Cosmic Shear Estimates for Cosmology with a Bayesian Convolutional Neural Network

5.1 Introduction

Dark matter and dark energy comprise $\sim 95\%$ of the energy density of the universe, but their natures are largely unknown. To investigate these entities, large-scale extragalactic surveys such as the Large Scale Survey of Space and Time (LSST – e.g. Ivezić et al., 2008) and Euclid (e.g. Collaboration et al., 2022) will soon provide observations of billions of galaxies. Cosmological probes of dark matter and dark energy aim to measure the structure and evolution of the universe, and thus rely in part on accurately and precisely measuring galaxy redshifts and galaxy shears of hundreds of millions of galaxies with well-constrained uncertainties. Therefore the task of obtaining sufficiently accurate shear estimates and understanding the error properties of these estimates is a major challenge.

Traditional techniques of shear estimation fall significantly short of the LSST shear science requirements for weak lensing. There is a need to investigate alternative shear estimation techniques that can leverage the recent advancements in image-based probabilistic machine learning with galaxy images. A key challenge in shear estimation is that the 'truth' shear values for real galaxies are unknown. For example, when utilizing HSC galaxies in our shear estimation network, the systematic uncertainties affecting the method with which

the HSC team estimated galaxy shears (discussed below) serve as the lower limit on shear estimation uncertainties that one can achieve using a machine learning network. Applying a machine learning method for shear estimation on real galaxy images while bypassing the systematic uncertainties present in traditional shear estimation methods requires using a training set of simulated galaxy images to evaluate real galaxy images. This is challenging because the accuracy of any machine learning model requires the training set be sufficiently representative of the evaluation set (i.e. overlap in parameter space). The impact of using simulated galaxy images to train a shear estimation model to predict real galaxy shears has not been well-studied.

Here we turn toward utilizing a real galaxy dataset obtained from the HSC shear catalog to investigate the performance of image based probabilistic machine learning for reproducing galaxy shape characteristics. Because we lack a 'known' shear values for each galaxy, we use galaxy ellipticity as a proxy for shear in order to gauge the efficacy of any given model. Ellipticities tend to be roughly 30-50 times the magnitude of galaxy shears and can be used to test the ability of a model to extract the shape information of a galaxy. We also lack 'true' ellipticity estimates and thus need to rely on HSC's provided ellipticity measurements in the HSC shear dataset. Utilizing HSC's ellipticity measurements as training and evaluation labels in our networks will necessarily increase the systematic bias floor in the model. In a future work, we will build a simulated shear dataset that is representative of future large scale survey galaxy images to expand this analysis.

Cosmic shear results from the bending of light from distant galaxies due to gravitational interactions with large scale structure. Weak lensing refers to gravitational lensing in the limit that deflected photons cause only small changes in the observed position, size, brightness, and shape of galaxies. The extent to which galaxy orientations deviate from a random distribution is thought to result from lensing. Galaxy shape distortions from lensing are typically on the order of 1% the size of the galaxy, which is a far smaller contribution than typical instrinsic galaxy ellipticities (~ 0.3) . This relative signal weakness is compounded

by coherent distortions produced from light propagating through the atmosphere, telescope optics, and incomplete knowledge of point spread functions (PSFs). Accurately measuring lensing shears for galaxies with low signal-to-noise ratios (SNRs) is an on-going challenge in weak lensing analyses. Systematic errors resulting from shape measurement must be reduced by factors of 5-10 in the next decade (Mandelbaum et al. (2014)).

Weak lensing cosmological analyses aim to quantify the small but spatially coherent distortions of galaxy shapes to probe the distribution of mass in the universe. Cosmic shear measurements are used to quantify the correlation of pairs of galaxy shapes as a function of angular separation and redshift to measure precision cosmology. A critical obstacle in performing weak lensing measurements is that weak lensing is not the only source of galaxy shape distortions; additional sources of distortion need to be removed in order to separate the weak lensing signal. Current survey plans rely on accurate shear measurements, however traditional moments-based approaches to shear measurements are limited by the systematic error contribtution from shape-noise. In order to leverage the increase of data from LSST and Euclid, shear needs to be measured with precision better than $\sim 2\%$. Sources of weak lensing systematic uncertainties include (Mandelbaum, 2015):

- approximations to the PSF in traditional methods (software used to extract morphological information)
- biased photometric redshifts
- accounting for the "shape noise" resulting from intrinsic, randomly oriented galaxy shapes
- instrument systematics
- incomplete PSF knowledge
- uncertainty in the impact of baryons on shears

The systematic error contribution from shape noise is the dominating error source in shear measurements (Springer et al., 2019). Shape noise is represented as σ_e in equation 12, and results from natural variation in galaxy intensity profiles and is considered the lower-limit of the statistical error that traditional shear estimators can achieve (Springer et al., 2019). A similar analysis was performed by Springer et al. (2019) who used a non-probabilistic CNN on galaxy images and found a reduction in shear RMS scatter of 26% compared to a traditional shear measurement technique. In this work we aim to improve on the shear estimate as well as provide uncertainties rather than only point-estimate shears. Applying machine learning to the task of shear estimation may also provide reliable shear estimates for galaxies with low SNR, increasing the number of viable galaxies and reducing statistical errors.

Two of the main sources of shear estimation uncertainties include incomplete PSF knowledge + method of PSF approximation and intrinsic galaxy shape noise. An important objective of this work is to quantify the extent to which probabilistic machine learning techniques can be used to reduce shear estimation biases resulting from treatment of galaxy PSFs and intrinsic galaxy shape noise.

The goal for this paper is to serve as a building block toward future shear analyses by providing a machine learning method for extracting shape information from galaxy images. We use real galaxy image data from HSC to estimate galaxy ellipticities. In a future analysis we will build a simulated dataset with known shear values and apply the model discussed in this work to real galaxy images. In this paper we explore a Bayesian convolutional neural network and non-probabilistic convolutional neural network approach using 5-band grizy galaxy images to produce ellipticity measurements with uncertainties from machine learning. Additionally, we perform ellipticity estimation with and without inclusion of PSFs as input along with galaxy images into the BCNN network to measure the effect of utilizing galaxy PSFs in image-based machine learning methods of shear estimation. We use $\sim$ 300k 5-band grizy images from the Hyper Suprime-Cam survey. In §5.2 we discuss shear formalisms; in

§5.3 we discuss past shear and ellipticity estimation techniques; in §5.4 we discuss the data and models used in this investigation; in §5.5 we state the results, and in §5.6 we provide a discussion.

5.2 Past shear and ellipticity estimation techniques

Many recent works measure shear with the Re-Gaussianization PSF correction method (Hirata & Seljak (2003), Mandelbaum (2018)) based on moments of the image and PSF to correct for effects of the PSF on galaxy shapes. This method has been studied on both real and simulated data (e.g., Mandelbaum et al. (2005, 2014, 2015), Hoekstra et al. (2017), Pujol et al. (2020), Tewes et al. (2019)). A potential flaw of this and similar methods is that it does not capture the detailed morphology (beyond second moments) of galaxies and the PSFs. Further, shear measurements of this kind suffer when applied to low SNR galaxies (Mandelbaum et al. (2018)). Shear estimation with machine learning techniques applied to galaxy images do not suffer from morphology simplification and may extend beyond the 'shape-noise limit' that prevents traditional shear estimation methods from obtaining shear estimates of low SNR galaxies (Mandelbaum et al., 2018; Springer et al., 2019).

See Table 5.1 for a summary of past works that have investigated shear estimation with machine learning. Previous works have applied neural network techniques to estimate galaxy shear and galaxy shear bias (Tewes et al., 2019; Pujol et al., 2020). Network inputs include ellipticity components, flux, size of the observed galaxy image, noise of the sky background, and the ellipticity and size of the PSF model at the location of the considered galaxy (see Tables 3 and 4). Tewes 2019 directly outputs a galaxy shear estimate. Gruen et al. 2010 uses a cost function that minimizes the difference between ensemble average of NN output and the true shear shared by the whole sample, noting specifically that the output of the model cannot be regarded as true ellipticity, but rather a quantity that, after averaging, gives good estimation for the shear. Pujol et al. 2020 outputs the multiplicative bias, which

is equivalent to the image response to shear, and additive bias of shear estimations, where the NN serves as a calibration method.

Author	**Model**	**Estimate**	**PSF treatment**
Gruen 2010	NN	Ellipticity	Pre-NN, PSF circularization or KSB
Tewes 2019	NN	Shear	Assume known PSF
Pujol 2020	NN	Shear bias	Constant PSF while training
Springer 2019	CNN	Shear	Constant PSF
Ribli 2019	CNN	Ellipticity	Varying PSF
Zhang 2023	CNN+NN	Shear	assume known PSF
Voigt 2024	CNN	Ellipticity	Constant PSF

Table 5.1 Summary of previous shear and ellipticity measurements and their treatment of PSFs.

There are few CNN models that directly output shear values from their models. Existing approaches involve: 1) output shape parameters from CNN and subsequently feeding the shape parameters to an NN model with MSB loss function (Zhang 2023); 2) training CNN models to predict ellipticity and averaging over all ellipticity in the set to obtain shear (Voigt 2024); 3) training CNN model to predict two ellipticity components which are later calibrated with MetaCalibration for shear (Ribli 2019). Only Springer 2019 uses CNN model that directly outputs shear.

While some CNN models consider PSF correction, no existing model directly addresses the problem of varying PSF through their models. They either treat the determination of PSF as a separate problem, assuming perfect knowledge about PSF during their training, or use a constant PSF for their training. In Ribili 2019, where PSF leakage is considered a potential problem, non-machine-learning method, i.e. MetaCalibration, is applied for correction. For NN models, the method by Gruen addresses PSF prior to training. PSF circularization is applied before shear is measured though the actual variation in PSF for each sample is very moderate. Tewes model assumes knowledge about PSF and mentions that to deal with varying PSF, two approaches can be taken: 1) for well-defined diffraction-limited PSFs, the field position of each galaxy can be added as input features; 2) for cases of stochastic at-

mospheric PSF where PSF would change for each exposure, features that describes the PSF can be added as input features, and the model will be trained with variations of potentially encountered PSF. In the paper Voigt 2024, some discussions regarding PSF dependence can be found, which is briefly summarized in the corresponding paper summary.

According to the LSST Science Requirements Document (SRD)[1], sufficiently accurate shear estimates for $\sim$ four billion galaxies are required to meet the LSST science goals for their main cosmological sample.

Currently, no published model satisfies the LSST shear science requirements needed for weak lensing analyses. Beyond the LSST metrics stated in the SRD, we consider additional probabilistic metrics for quantifying the quality of uncertainty estimates (see Table 1) (Malz & Hogg, 2020; Schmidt et al., 2020a; Jones et al., 2022a). The requirement thresholds for the probabilistic metrics are not as well quantified at this time as those for point metrics, but they allow us to compare the performance between different probabilistic models evaluated on the same data.

5.3 Data and Methods

For the analysis in this work we use data from the Hyper Suprime Cam survey, which is intended to approximate the data produced by future large-scale deep surveys for shear estimation (Collaboration et al., 2021). We use the Hyper-Suprime Cam (HSC) shear catalog which provides images from (insert) galaxies with measured shear values. We note that this dataset does not contain 'known' shear values and the method by which the shear estimates are made introduces significant systematic uncertainties to the shear values. We therefore turn to utilizing ellipticities as a proxy for shear, since shears are typically on the scale of 1-3 percent of galaxy ellipticities and therefore ellipticities are easier to predict with noisey data. In a future work we will apply the machine learning models used in this work to a

[1]https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17

simulated dataset with 'known' shear values with the intention of evaluating the model on real galaxy images.

Shear estimate accuracy needed for weak lensing analyses requires accurately modeling PSFs. PSFs are needed for mapping a point in sky coordinates to a surface brightness profile measured by a detector in pixel coordinates, but are prone to producing systematic biases. PSF anisotropies result in a multiplicative and additive bias for shear measurements (Mandelbaum et al 2018). Accurate PSF modeling allows for these biases to be mitigated. Here we discuss the approaches to both PSF modeling and shear measurement used by HSC and DC2 and contrast that with our own method. We also discuss the LSST science requirements for shear estimation that we use for examining the performance of our model.

5.3.1 Hyper Suprime-Cam (HSC) Survey Shape Catalog

There are three currently operating weak lensing surveys – KiDS (Heymans et al., 2020), DES (Collaboration et al., 2018), and HSC, (Hamana et al., 2020; Hikage et al., 2019). I choose to use the HSC shape catalogue data for this analysis because it provides high-resolution imaging extending to deep redshifts $z > 4$ (Mandelbaum et al., 2018). The shear catalog covers an area of 136.9 deg^2 over six fields, with a mean i-band seing of 0.58" and 5σ point-source depth of i $\sim$ 26. This catalogue contains over 9 million galaxies, which is only a small portion of the larger HSC dataset. To create a galaxy image dataset for ellipticity estimation, I use the HSC Shape Catalog[2] (Mandelbaum, 2018) and the HSC photometry database to obtain 5-band grizy galaxy images paired with grizy PSF images for 299,000 galaxies. Figure 5.3 depicts the full dataset creation process starting with the HSC shape dataset. The 6 shape fields are first queried to obtain roughly 12 million galaxies. After imposing magnitude cuts, we randomly sample 300,000 galaxies with which to obtain 5-band grizy image cutouts. We choose 300,000 for our dataset size because using a significantly larger dataset size provides

[2]https://hsc-release.mtk.nao.ac.jp/doc/index.php/s16a-shape-catalog-pdr2/

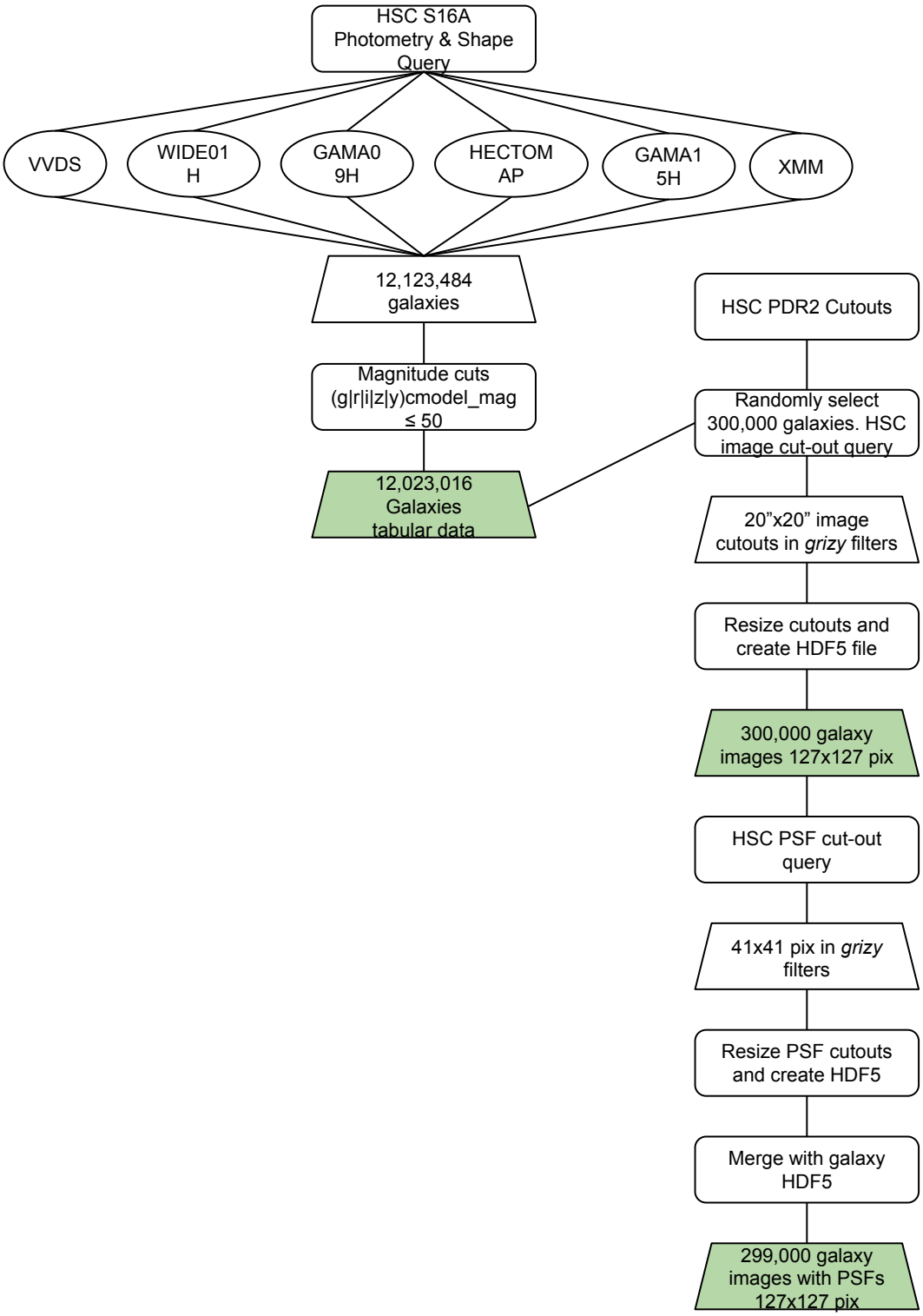

Figure 5.1 Visualisation of the process required to obtain the HSC grizy galaxy images and PSFs that are used for ellipticity prediction.

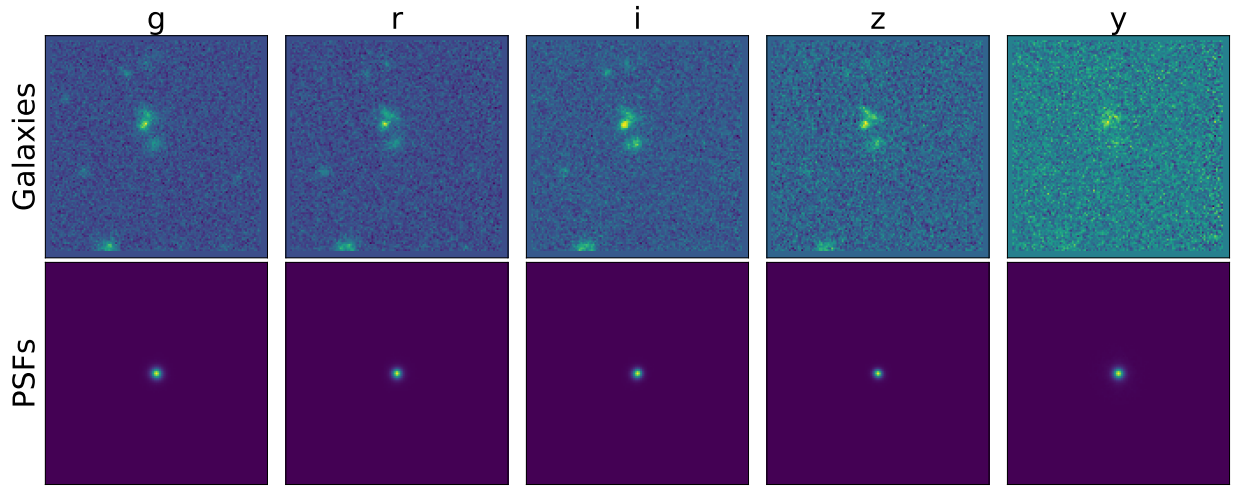

Figure 5.2 Example of a set of grizy galaxy images and grizy PSF images for a randomly selected galaxy from our dataset used for the ellipticity prediction analysis. The images are 127x127 pixels, where each pixel represents 0.186 arcseconds.

negligible benefit to the model performance in exchange for significant additional computing time for model testing.

Figure 5.3.4 contains example images for galaxies with small and large ellipticity values. Figure 5.5 depicts the e_1 and e_2 distributions and Figure 5.6 depicts the σ_e and e_{RMS} distributions. The photometry distribution is shown in figure 5.3.

5.3.1.1 Shear measurement with HSC

The HSC Shape Catalog provides the following information for each galaxy:

- RA, DEC
- object ID
- Galaxy distortions e_1, e_2
- distortion uncertainty σ_e
- e_{RMS}

- shape uncertainty weight
- multiplicative shear bias m, averaged over components
- additive bias for each component c_1, c_2

The reduced shear components g_1 and g_2 can be calculated from these quantities with equation 5.1 and used as training labels in a machine learning model. Shear measurements are calculated for every galaxy i with

$$\hat{\gamma_i} = \frac{1}{1+\bar{m}}\left[\frac{e_i}{2R} - c_i\right] \tag{5.1}$$

where the weighted-average multiplicative bias factor is

$$\bar{m} = \frac{\sum_i w_i m_i}{\sum_i w_i} \tag{5.2}$$

and R is the shear responsivity quantifying the response of distortion to a small shear r (Bernstein & Jarvis, 2002), defined as

$$R = 1 - \frac{\sum_i w_i e_{rms}^2}{\sum_i w_i} \tag{5.3}$$

The HSC pipeline obtains shear measurements with GalSim, which uses a moments-based shape measurement method where the shear is estimated by averaging galaxy shapes. Galaxy shapes themselves are produced by processing the coadded i-band images using a re-Gaussianization PSF correction method (Hirata & Seljak, 2003). The basic principle of galaxy shape estimation using this method is to fit a Gaussian profile with elliptical isophotes to the image, and to define the components of the distortion

$$(e1, e2) = \frac{1-(b/a)^2}{1+(b/a)^2}(cos2\phi, sin2\phi) \tag{5.4}$$

where b/a is the axis ratio and ϕ is the position angle of the major axis with respect to the equatorial coordinate system. See Figure 5.5 for a visual of the distributions of both ellipticity components.

The ensemble average distortion is obtained by

$$(\hat{g_1}, \hat{g_2}) = \frac{1}{2R}\big\langle (e_1, e_2) \big\rangle. \tag{5.5}$$

Shape uncertainty weights are defined as the inverse variance of the shape noise

$$w = (\sigma_e^2 + e_{rms}^2)^{-1}, \tag{5.6}$$

where σ_e is the shape measurement error for each galaxy and e_{rms} is defined per galaxy as the signal-to-noise ratio and resolution factor calibrated by using an ensemble of galaxies with SN and resolution values similar to the given galaxy. See Figure 5.6 for a visualisation of the distributions of σ_e and e_{RMS} from HSC.

HSC uses an empirical PSF modeling algorithm called Point Spread Function Extractor (PSFEx) (Bertin 2013) for PSF interpolation. PSFEx models the PSF as a linear combination of basis vectors (of pixel values) and interpolates the basis vectors across the CCD. To characterize the PSF for each CCD, HSC selects bright stars (SNR > 50) to feed into PSFex and model the position-dependent PSF Mandelbaum et al. (2018).

5.3.2 Our treatment of PSFs

The PSF used to produce a galaxy image plays a significant role in our ability to measure the shape information of that galaxy. Therefore we explore the use of integrating galaxy PSFs as inputs into the machine learning models in addition to galaxy images themselves in order to measure galaxy ellipticity.

HSC and DC2 do not provide their PSF information, so we use bright nearby star images

to directly measure PSFs for each galaxy. Stars serve as point sources, so the measurement of their surface brightness profiles provide a direct measurement of a PSF. This measurement is variable with distance between a given star-galaxy pair, so the PSF must be interpolated to the location of a galaxy if there is no nearby star for a given galaxy. We impose a distance requirement between a star galaxy pair of 10 arcseconds to reduce systematics from PSF variations with distance. See Figure 5.2 for example star images that are used as PSFs in our model.

5.3.3 LSST DESC Science Requirements for Shear Estimation

The shear science requirements are informed by the WL 3x2-point correlation function measurements (shear-shear, galaxy-shear, and galaxy-galaxy). In weak lensing analyses there are four main sources of systematic uncertainty: redshift, number density, multiplicative shear uncertainty, and additive shear uncertainty, which are allocated 0.7, 0.2, 0.7, and 0.2 of the total systematic error budget. Here we focus on the multiplicative and addictive shear biases. See the SRD for a more detailed discussion. We note that in order to achieve the required statistical precision in the LSST weak lensing analysis, one needs data from the full hemisphere sky coverage of LSST (producing 4 billion galaxy sources with i ¡ 25.3 in this sample).b

Statistical shear power is the total shear measurement signal resulting from gravitational lensing by large scale structure. Residual shear power is the leftover signal in shear measurements after accounting for known shear sources, which can be attributable to systematic errors or noise. The ultimate goal of residual shear error minimisation is to reach the scale of the statistical error floor established by the coadded images. Galaxy shear error on small angular scales (< a few arc minutes) is dominated by intrinsic galaxy noise, while on large angular scales the error is dominated by large scale structure cosmic variance. In the full LSST sample, these two errors sum to a 'shear cross correlation noise level' of $3x10^{-7}$, which sets the requirement that the systematic component be 30% of this noise level for it

to become negligible when added in quadrature. In other words,

- Residual shear power systematics (after corrections) resulting from the galaxy shear method and measurement hardware must be less than a third of the total statistical errors

In the LSST DESC SRD, there are different requirements for the WL shear sample for the Y10 and Y1 observations. For completeness we provide both. For Y10, the sample is divided into 10 tomographic photo-z bins of $z = 0.1$ between $0.2 < z < 1.2$ and 5 photo-z bins of $z = 0.2$ over the same range for Y1.

- Systematic uncertainty in the redshift-dependent shear measurement shall not exceed 0.003 in the Y10 DESC weak lensing analysis
- Systematic uncertainty in the PSF model size defined using the trace of the second moment matrix shall not exceed 0.1% in the Y10 DESC weak lensing analysis
- Systematic uncertainty in the stellar contamination of the source sample shall not exceed 0.1% in the Y10 DESC weak lensing analysis.
- Systematic uncertainty in the redshift-dependent shear measurement should not exceed 0.013 in the Y1 DESC WL analysis
- Systematic uncertainty in the PSF model size defined using the trace of the second moment matrix should not exceed 0.4% in the Y1 DESC WL analysis.
- Systematic uncertainty in the stellar contamination of the source sample should not exceed 0.4% in the Y1 DESC WL analysis.

A requirement on additive shear bias is deferred until a future version of the LSST DESC SRD, so we use the previous error allotment for multiplicative and additive shear biases (0.7 and 0.2, respectively) to constrain an additive shear bias requirement:

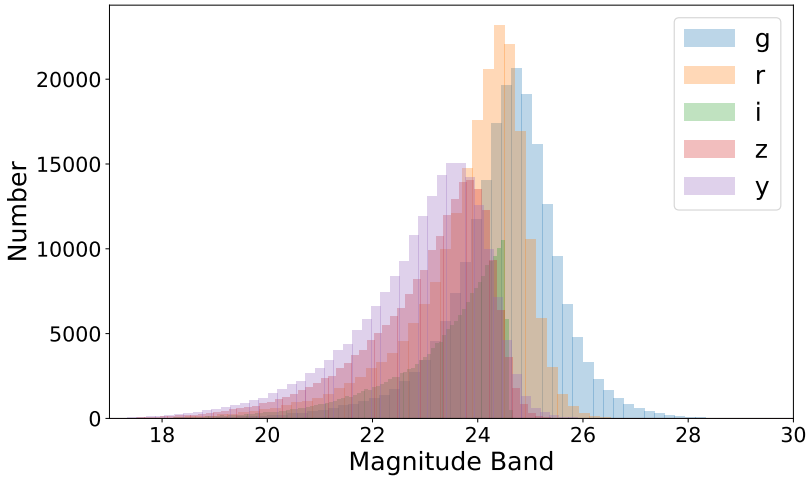

Figure 5.3 Distribution of photometry for the dataset used for the ellipticity prediction analysis.

- For Y10: Approximately 0.00082
- For Y1: Approximately 0.00165

and a multiplicative shear bias requirement:

- For Y10: Approximately 0.00287
- For Y1: Approximately 0.00578

5.3.4 Building the shape dataset

We use the HSC shape catalog and the HSC photometry database to obtain 5-band grizy galaxy images paired with grizy PSF images for 299,000 galaxies. Figure 5.3 depicts the full dataset creation process starting with the HSC shape dataset. Figure 5.3.4 contains example images for galaxies with small and large ellipticity values. Figure 5.5 depicts the e_1 and e_2 distributions and Figure 5.6 depicts the σ_e and e_{RMS} distributions. The photometry Distribution is shown in figure 5.3.

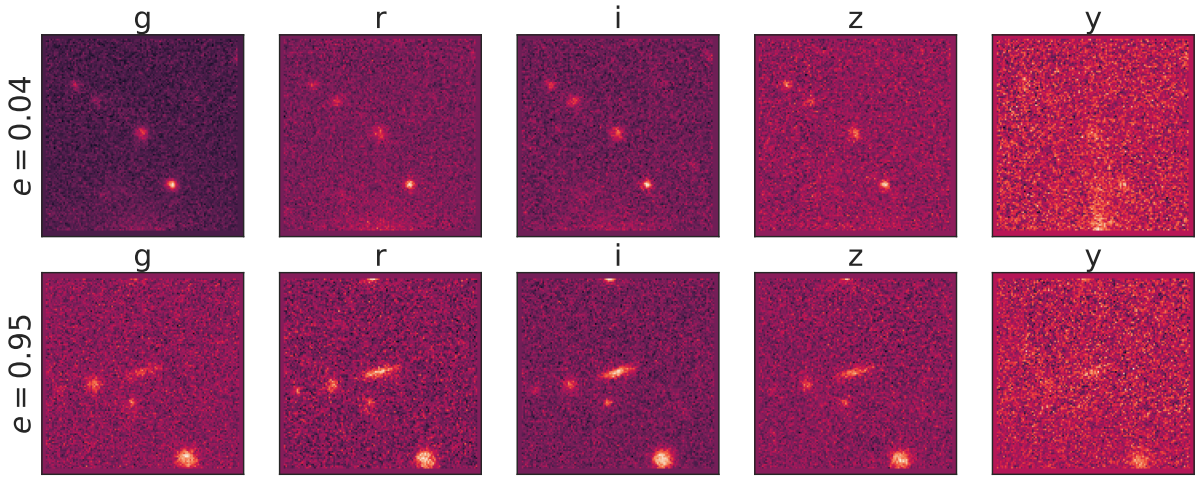

Figure 5.4 Example grizy galaxy images for high (BOTTOM) and low (TOP) ellipticities, as provided by the HSC shape catalog. The ellipticity measurement corresponds to the central galaxy located in the center of the image. The images are 127x127 pixels, where each pixel represents 0.186 arcseconds.

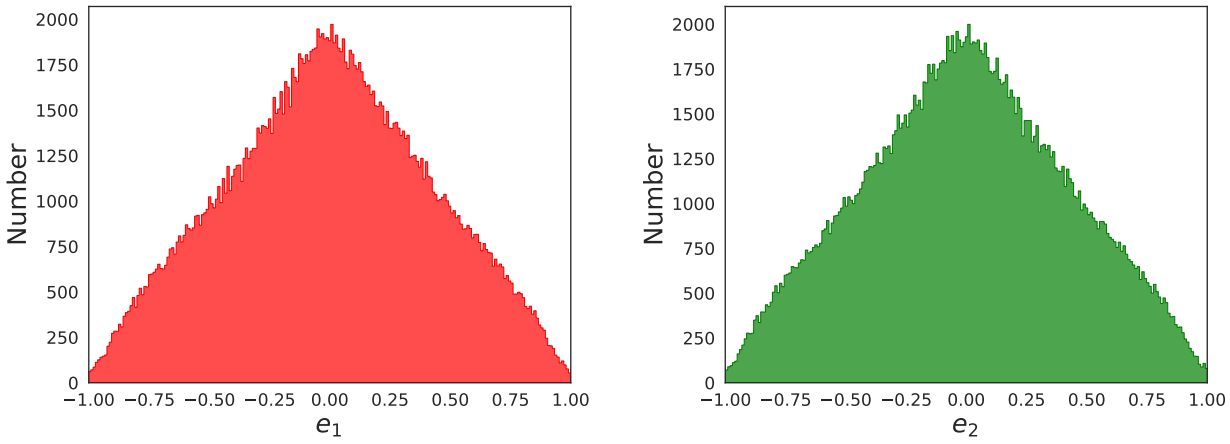

Figure 5.5 Visualisation of the two components of the ellipticities provided by the HSC Shape Catalogic that we use in the final ellipticity estimation datasets.

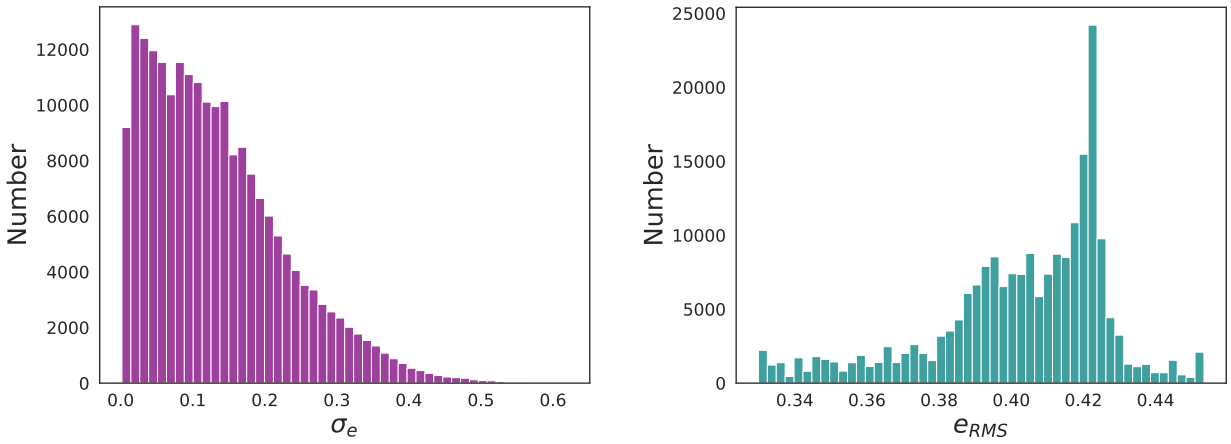

Figure 5.6 Distributions of σ_e and e_{RMS} provided in the HSC Shape Catalog for the galaxy sample used for the ellipticity estimation analysis in this work.

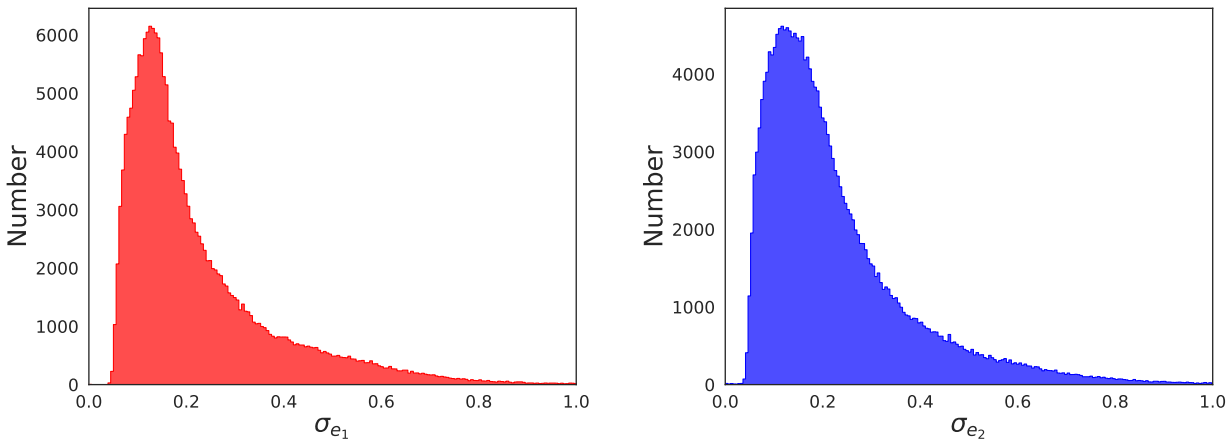

Figure 5.7 Ellipticity estimation uncertainties in each dimension in the ellipticity estimation analysis performed in this work using grizy galaxy images as inputs into the BCNN.

5.3.5 Network architectures

We use a BCNN and a CNN model for ellipticity estimation. Model inputs for each galaxy consist of an ellipticity label and 5-band galaxy images with or without 5-band PSF images. All images contain 127x127 pixels. Data outputs for the CNN and BCNN differ: the CNN outputs single ellipticity estimates while the BCNN outputs an ellipticity probability distribution consisting of a mean and standard deviation. We treat the mean of the distribution as the ellipticity estimate and use the standard deviation as the prediction uncertainty. We predict ellipticity using one dimensional inputs (e_1 or e_2) using the BCNN and CNN. With the CNN we also predict ellipticity using two-dimensional inputs (e_1 and e_2). We do not use ellipticity uncertainties provided by HSC as inputs. We attempted using combined ellipticity e as a training label and the performance was significantly worse than predicting the individual or combined components.

We investigated a number of different probabilistic deep neural network models for ellipticity estimation, but the best performing final model we adopted is based largely on the GoogLeNet deep convolutional neural network developed by Szegedy et al. 2014. The only difference between the CNN and BCNN googlenet model variations we use for this analysis is that the BCNN network contains a final output layer that is a probabilistic dense variational layer which outputs a normal distribution, whereas the CNN network contains a final dense layer that outputs a single ellipticity estimate. We converted this base model architecture into a probabilistic architecture (see Fig. 5.8) using TensorFlow (Abadi et al., 2016). We performed a parameter grid search to optimize for the number of epochs, number and type of layers, number of nodes per layer, learning rate, loss function, activation functions, and optimizers. We train using an AMD Ryzen Threadripper PRO 3955WX with 16-Cores and NVIDIA RTX A6000 GPU. Training runtimes are typically over 24 hours for 500 epochs for the final models and evaluation runtimes are on the scale of minutes.

We performed a hyperparameter grid search that iterated over a number of hyperparam-

eters:

- # layers
- types of layers
- # nodes per layer
- learning rate
- loss function
- activation functions
- # epochs
- image pixel scaling
- batch sizes
- kernel sizes
- filters
- strides

The final BCNN and CNN models used for ellipticity estimation can be viewed here[3]

5.3.6 Conformal Prediction

To ensure good statistical coverage of the BCNN, we use conformal prediction to rescale the predicted uncertainties. Conformal prediction is a promising method for uncertainty quantification that is agnostic to the method of shear or ellipticity prediction and does not need to assume a probability distribution (Papadopoulos et al., 2002; Vovk, 2012; Lei &

[3]https://colab.research.google.com/drive/1P15p-COOUCuzZq7B4CS20qwWBrr82GbP?usp=sharing

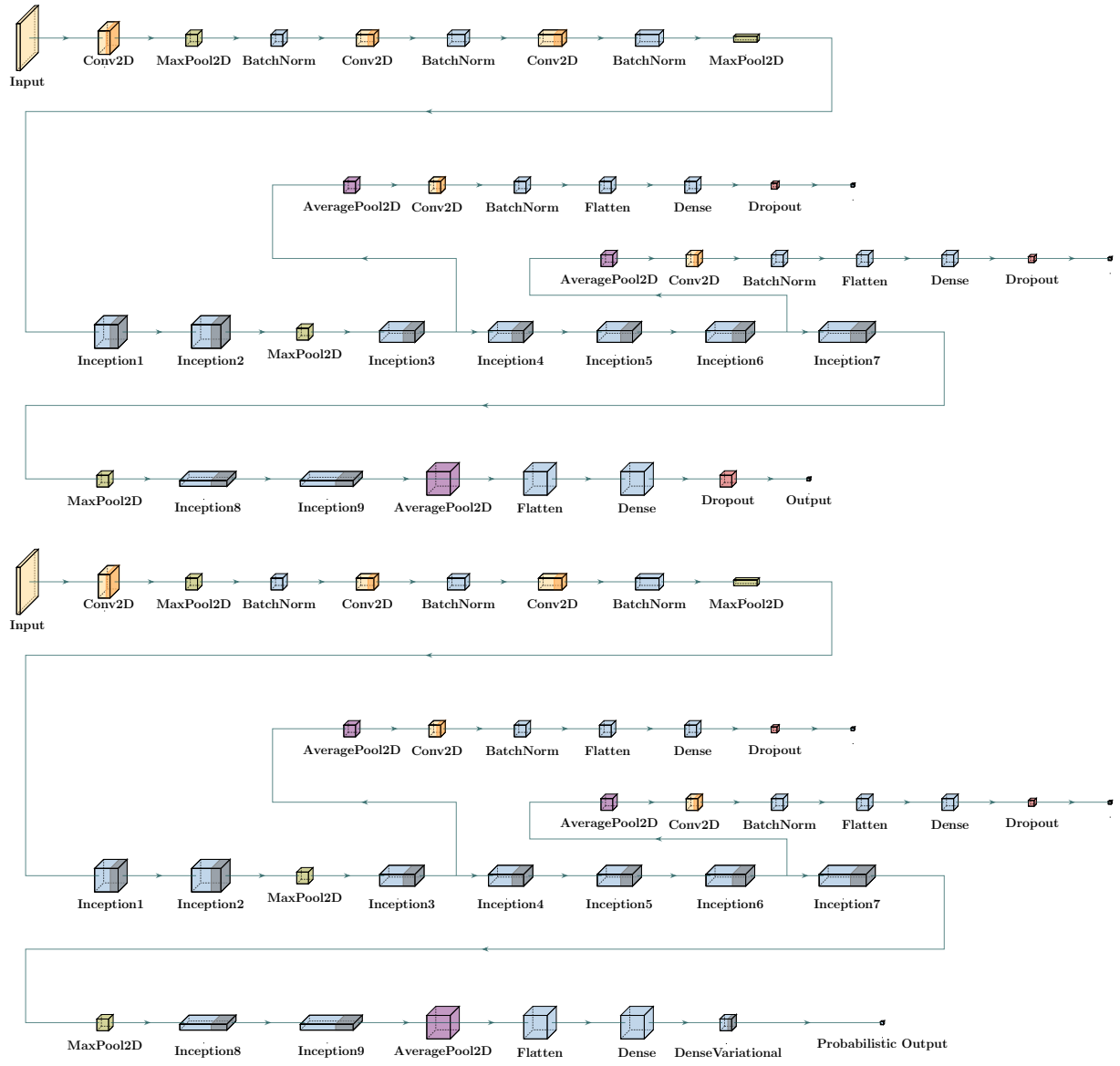

Figure 5.8 TOP: CNN architecture. BOTTOM: BCNN architecture. The inputs for both networks are five-band galaxy images in the g,r,i,z,y filters paired with ellipticity labels from HSC. The output for the CNN is a single point ellipticity estimate while the output for the BCNN is an ellipticity PDF. We assume Gaussianity in the creation of the PDF, so an ellipticity uncertainty is produced by the standard deviation of the PDF.

Wasserman, 2014). We have previously utilized conformal prediction for photo-z estimation using a BCNN (Jones et al. 2024). It works by including an extra calibration step after a model is trained to create credible intervals using a parameterisation dataset with which to re-scale an evaluation dataset to maintain desired statistical coverage. Given a required credible interval (ie. 90% coverage), this calibration step allows us to determine how a given prediction score maps to the range of predicted values that has that credible interval. For Bayesian models like the BCNN, we can calibrate how to scale the predicted variance to ensure exact coverage (Hoff, 2021). In addition, this method can add uncertainty quantification to networks (like CNNs using quantile loss) that previously only supported point predictions by mapping between prediction scores and statistical coverage (Angelopoulos & Bates, 2022).

We apply the conformal prediction analysis to individual ellipticity bins themselves rather than the entire dataset. We implement a binned approach to conformal prediction where we use a calibration dataset with known ellipticity values and separate galaxies into ellipticity bins of $e_i = 0.1$. In each bin we calculate a nonconformity score

$$S = \frac{|e_i - e_{i,HSC}|}{\sigma_e} \tag{5.7}$$

and calculate the quantile of nonconformity scores using the desired coverage of 0.683

$$q = Q(S, 0.683) \tag{5.8}$$

which are used to scale the uncertainties associated with evaluation set galaxies. We divide evaluation set galaxies into bins based on their ellipticity and we scale their uncertainties by the corresponding quantile scaling that was calculated from the calibration dataset ellipticities.

$$\sigma_{e,f} = Q_i(S, \alpha) * \sigma_{z,i}$$

where $Q_i(S, \alpha)$ is the quantile scaling parameter calculated from the calibration data set in bin i, $\sigma_{e,i}$ is the ellipticity uncertainty produced by the BCNN for an evaluation galaxy in bin i, and $\sigma_{e,f}$ is the final ellipticity uncertainty for a galaxy in bin i.

5.3.7 Metrics

Shear uncertainties are propagated to measurement uncertainties on dark matter and dark energy. Therefore, our choice of metrics to evaluate the ellipticity determinations in this work include the main two metrics for evaluating shear predictions in LSST science requirements document: multiplicative and additive bias averaged over the entire evaluation sample. We also include scatter and RMS error to compare model performance.

Table 5.3.7.1 contains the definition of the metrics we use for evaluating the quality of ellipticity estimates. Since we are treating ellipticity as a proxy for shear, we use the primary metrics for evaluating the quality of shear predictions as well: multiplicative and additive bias (Eqs. 1 and 2). We additionally include RMS error (Eq. 3) and scatter (Eq. 4.) The RMS error is given by a standard definition where n_{gals} is the number of galaxies in the evaluation testing set and Σ_{gals} represents a sum over those galaxies.

5.3.7.1 Probability metrics

In addition to the previously mentioned point metrics, we also perform a probabilistic analysis of the output produced by the BCNN. Here we focus on the results from using single component ellipticities as training and evaluation labels.

The Probability Integral Transform (PIT) is one probabilistic metric that can detect systematic error in the distribution width for galaxy samples with known evaluation labels Malz & Hogg (2020); Malz (2021). For the case of ellipticity (or shear) estimation, the PIT value for a single galaxy is defined in Eqn. 6 in Table 5.3.7.1, where $p(e)$ is the predicted ellipticity PDF. A histogram of PIT values for a galaxy sample should be uniform for an

accurate collection of $p(e)$ samples. Ideally, the PIT histogram is flat across all ellipticity bins. If the PIT histogram peaks at the center, the $p(e)$ collection is too broad. If the PIT histogram peaks at high and low PIT values, the p(e) samples are too narrow.

Coverage is another probabilistic metric used to assess whether confidence intervals are accurate (Eq. 5). In this case, we define coverage as the fraction of galaxies that have an ellipticity within their 68% confidence interval. Ideally, 68% of evaluated galaxies should have true ellipticities within their 68% confidence interval. If the coverage is over 68%, then the estimated uncertainties are on average too large. Similarly, if the coverage is below 68%, the estimated uncertainties are on average too small.

Table 5.2 Metrics used to assess model performance.

Point Metrics		Probabilistic Metrics	
Add. bias	$b = e_{pred} - e_{HSC}$ (1)	Coverage	$\sum_{i}^{n_{gals}} \frac{(\bar{e}_{pdf,i} - e_{HSC,i}) < e_{\sigma,i}}{n_{gals}}$ (5)
Mult. bias	m = 1 - $\frac{e_{pred}}{e_{HSC}}$ (2)		
RMS error	$\sqrt{\frac{1}{n_{gals}} \Sigma_{gals} \left(\frac{e_{pred} - e_{HSC}}{1 + e_{HSC}} \right)^2}$ (3)	PIT:	$\int_{-\infty}^{e_{HSC}} p(e) de$ (6)
Scatter	Median($\lvert \Delta z - \text{Median}(\Delta z_i) \rvert$) (4)		

5.4 Results

In this work we used HSC ellipticity measurements as 'ground truth' to build probabilistic and non-probabilistic convolution neural network models to learn shape characteristics from galaxy images. Specifically, we used these models to recover the ellipticity measurements made by HSC. We investigated the use of galaxy PSFs as inputs into the model along with 5-band photometric galaxy images. We also examined the performance differences between predicting individual ellipticity components or by jointly predicting components.

We found no significant benefit to ellipticity estimates by including galaxy PSFs with galaxy images as input into the CNN and BCNN, however we did find that jointly estimating

Network	Additive Bias	Multiplicative Bias	RMS	Scatter
e_1 (CNN, images)	-0.0172	-0.174	0.150	0.165
e_1 (CNN, images + PSFs)	-0.0211	-0.187	0.173	0.154
e_2 (CNN, images)	0.017	-0.222	-0.050	0.157
e_2 (CNN, images + PSFs)	-0.029	-0.31	0.189	0.159
e_1 (BCNN, images)	-0.0363	-0.142	0.175	0.164
e_1 (BCNN, images + PSFs)	-0.0382	-0.176	0.148	0.136
e_2 (BCNN, images)	0.0165	-0.0159	-0.0124	0.1637
e_2 (BCNN, images + PSFs)	0.0289	-0.179	0.154	0.154

Table 5.3 Comparison of the performance results with the CNN and BCNN using single component ellipticities for training and evaluation.

the ellipticity components provided significantly better performance than individually estimating ellipticity components. Figures 5.11 and 5.12 contain a visualisation of additive bias, multiplicative bias, scatter, and RMS error for the CNN and BCNN using both 1)galaxy images alone and 2) galaxy images with PSFs to estimate both ellipticity components. These results averaged over the entire evaluation set are provided in Table 5.3. Figure 5.13 visualizes the performance of jointly predicting both ellipticity components. These results averaged over the entire evaluation set are provided in Table 5.4. Jointly predicting e_1 and e_2 improved e_2 more than e_1. For the case of galaxy image inputs with the CNN, e_2 saw a reduction in additive bias by 73% and a reduction in multiplicative bias by 70% compared to the case where the CNN predicted e_1 individually. Comparatively, e_1 saw a reduction in multiplicative bias by 35% while additive bias increased by 138%.

We find that conformal prediction significantly improved the PIT histograms of ellipticity PDFs (see Figs. 5.14 and 5.15). A visualisation of the BCNN ellipticity uncertainty estimates on e_1 and e_2 after conformal prediction are shown in Fig. 5.7 and a comparison to uncertainties provided by the HSC team is shown in Fig. 5.16. The original uncertainties were tremendously underestimated (narrow PDFs) and performing conformal prediction re-scaled the uncertainties to an acceptable distribution.

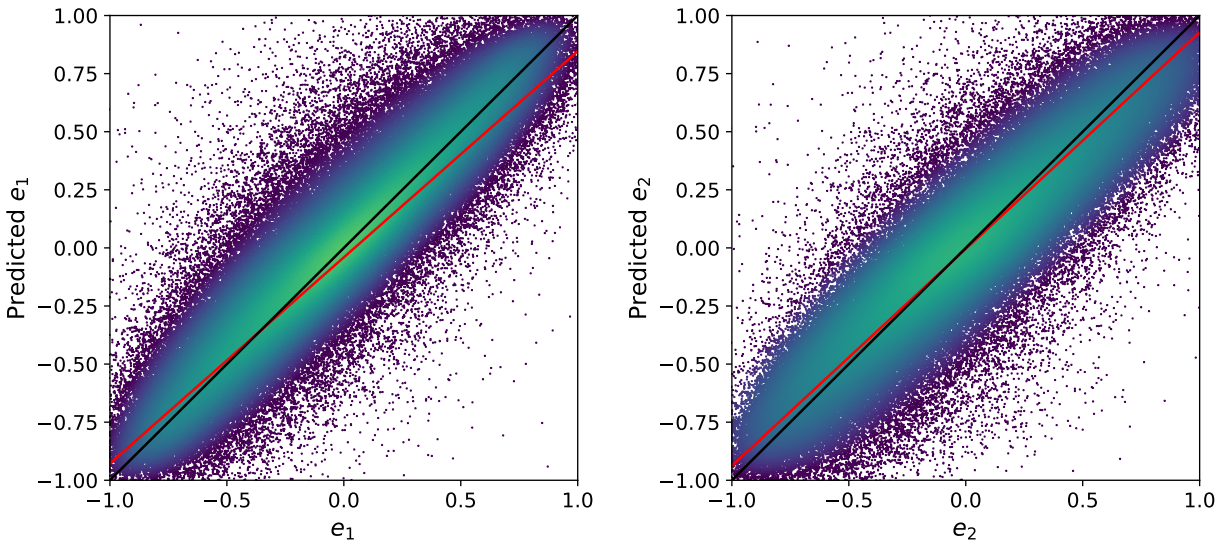

Figure 5.9 CNN ellipticity predictions with inputs consisting of 5-band grizy images. e_1 and e_2 were fed to the model as training labels together and are predicted together. The multiplicative and additive bias for e_1 are $m_{e_1} = -0.113$, $b_{e_1} = -0.041$. The multiplicative and additive bias for e_2 are $m_{e_2} = -0.067$, $b_{e_2} = -0.0046$.

Network	Additive Bias	Multiplicative Bias	RMS	Scatter
e_1 (CNN, images)	-0.041	-0.113	0.164	0.175
e_1 (CNN, images + PSFs)	-0.0035	-0.130	0.196	0.174
e_2 (CNN, images)	-0.0046	-0.067	0.166	0.173
e_2 (CNN, images + PSFs)	-0.070	-0.042	0.196	0.211

Table 5.4 Comparison of the performance results with the CNN when jointly predicting e_1 and e_2.

5.5 Discussion

Shear estimation using machine learning with galaxy images as input faces many challenges. Galaxy shears are typically only 1-3 percent of the intrinsic ellipticity, which makes the task of distinguishing shear from ellipticity a challenging one. Utilizing machine learning techniques to optimize galaxy shear estimates at the level required for weak lensing analyses will likely require the construction of a training set with 'known' shear values. In a future work we will apply the machine learning models used in this work to a simulated dataset

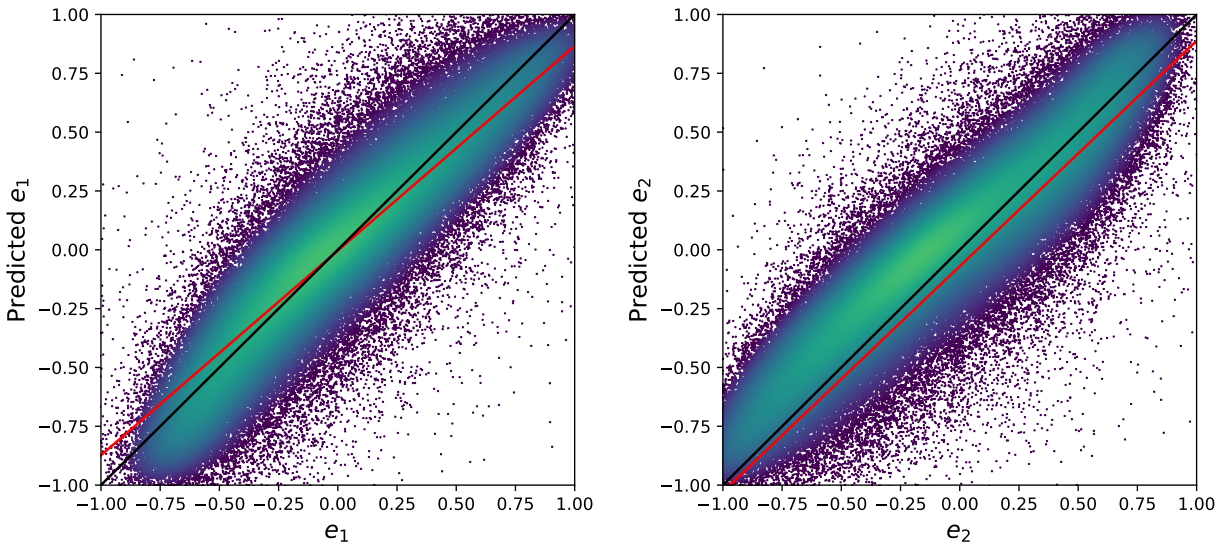

Figure 5.10 CNN ellipticity predictions with inputs consisting of 5-band grizy images with 5-band PSF images. e_1 and e_2 were fed to the model as training labels together and are predicted together. The multiplicative and additive bias for e_1 are $m_{e_1} = -0.130$, $b_{e_1} = -0.0035$. The multiplicative and additive bias for e_2 are $m_{e_2} = -0.042$, $b_{e_2} = -0.070$.

with 'known' shear values with the intention of evaluating the model on real galaxy images.

Since we used the HSC shape data catalog as training and evaluation labels in the ellipticity analysis in this work, we would expect the uncertainty of our predictions to exceed the uncertainty in the HSC measurements. Figure 5.16 visualizes the fractional uncertainty of our predictions (after conformal prediction) compared to HSC, which clearly indicates this is the case. in Fig. 5.16, the fractional uncertainties provided by the BCNN after conformal prediction are generally double the size of the fractional uncertainties provided by the HSC team, which is consitent with the hypothetsis that the noise present in the HSC ellipticity measurements are the dominating source of uncertainty in the BCNN estimates.

The lack of correlation between ellipticity estimate accuracy and the inclusion of galaxy PSFs with galaxy images as input into the BCNN and CNN models may be attributable to the inherent noise present in the ellipticity measurements themselves. We hypothesized that the ellipticities from HSC, which are subject to the same sources of systematic uncertainty

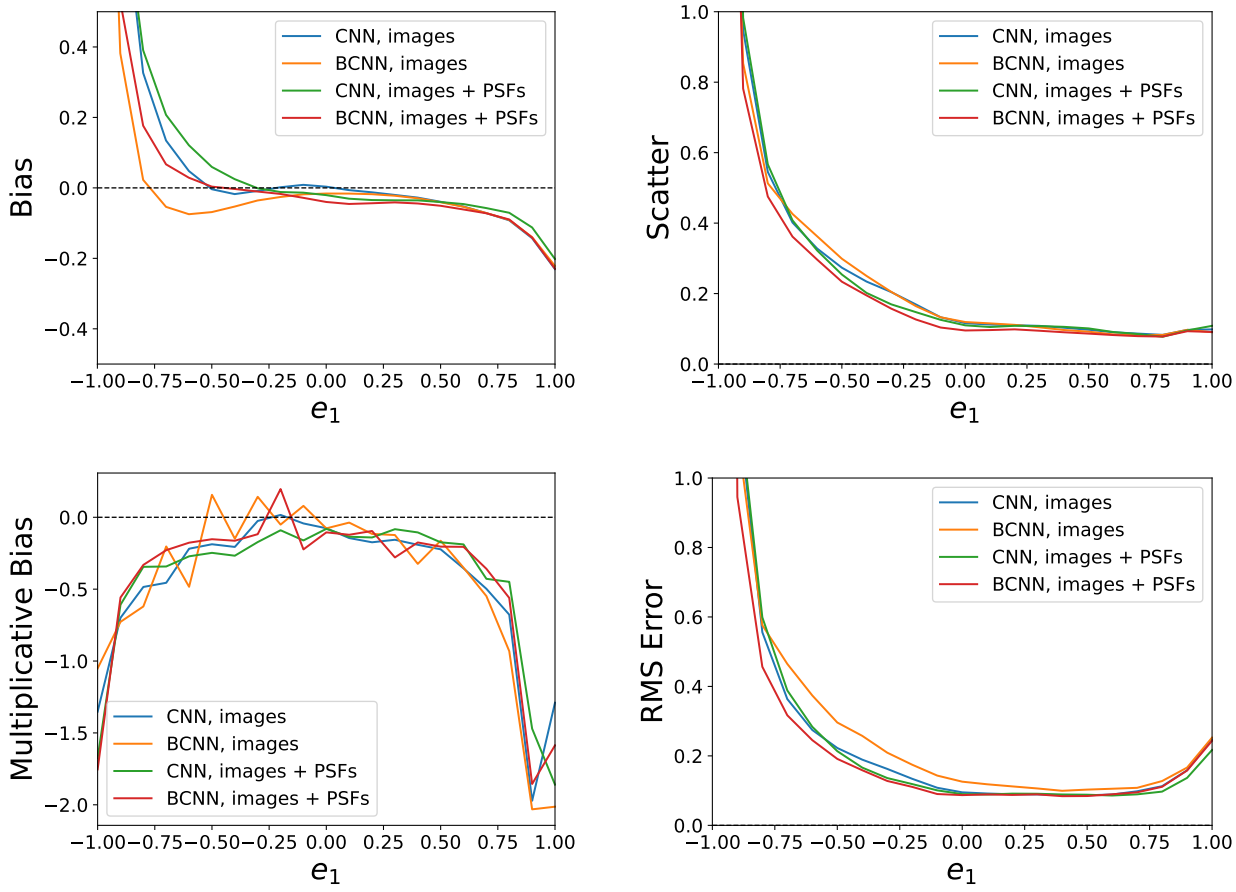

Figure 5.11 BCNN and CNN performance when trained on e_1 with respect to additive bias, multiplicative bias, scatter, and RMS error. The plots reflect results with 70% of galaxies for training, 10% for validation, and 10% for evaluation.

present in HSC shear estimates, may be aided by PSF inclusion in the model despite the noisey nature of the ellipticity measurements due to the fact that ellipticity is significantly easier to predict than shear. It's also possible that the lack of correlation between PSF inclusion and ellipticity estimate quality is due to deficiencies in the model architecture that may be overcome with further model hyperparameter tuning. We will perform additional model tuning in a future work using this same data and also apply a BCNN to simulated data with 'known' shear values that we can evaluate on real HSC galaxies. Lastly,

Training and evaluating a CNN model on both ellipticity components at the same time

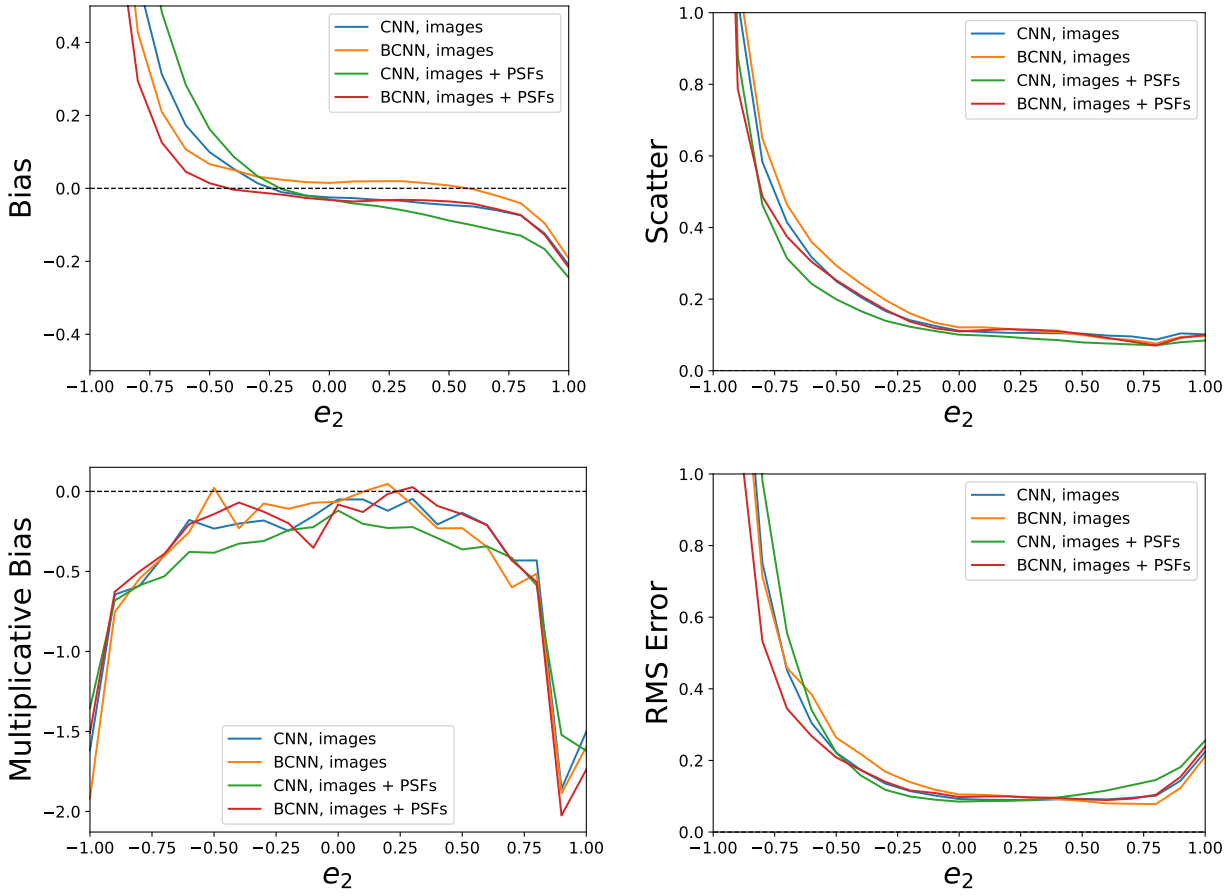

Figure 5.12 BCNN and CNN performance when trained on e_2 with respect to additive bias, multiplicative bias, scatter, and RMS error. The plots reflect results with 70% of galaxies for training, 10% for validation, and 10% for evaluation.

provided superior results to the singular approach. Jointly predicting e_1 and e_2 improved e_2 more than e_1. For the case of galaxy image inputs with the CNN, e_2 saw a reduction in additive bias by 73% and a reduction in multiplicative bias by 70% compared to the case where the CNN predicted e_1 individually. Comparatively, e_1 saw a reduction in multiplicative bias by 35% while additive bias increased by 138%. Based on these results, we argue that future machine learning based approaches for shear estimation should be performed on both components at the same time.

The current status quo of traditional shear measurements do not achieve the necessary

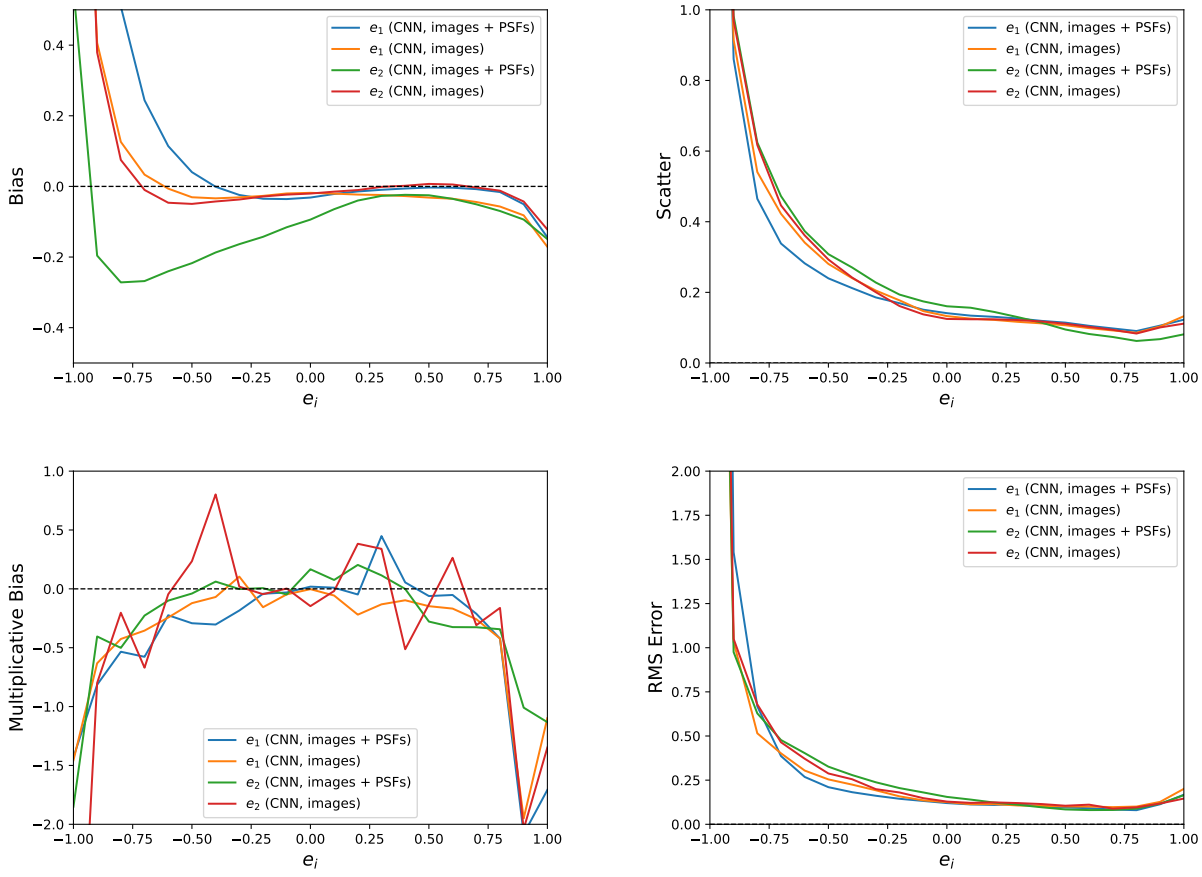

Figure 5.13 BCNN and CNN performance when trained on both e_1 and e_2 at the same time, with respect to additive bias, multiplicative bias, scatter, and RMS error. The plots reflect results with 70% of galaxies for training, 10% for validation, and 10% for evaluation.

precision required for weak lensing surveys, as defined by the LSST science requirements for multiplicative and additive bias. One path forward toward achieving these shear estimation requirements may lie in probabilistic machine learning approaches using simulated data with 'known' shear values. Based on the results here, we believe the CNN and BCNN models provide promising results recovering ellipticity measurements, indicating their potential capability to measure galaxy shears from galaxy images. Our hope is that by applying these models to simulated galaxy images with known shear values, we may be able to finally achieve the shear estimation requirements needed for future shear estimation initiatives as

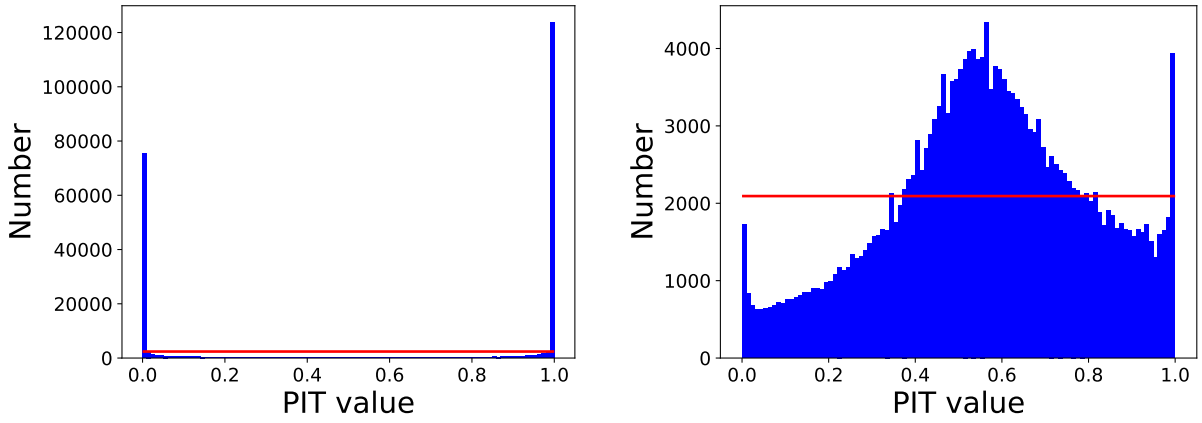

Figure 5.14 PIT histograms of the ellipticity PDF produced by the BCNN before (LEFT) and after (RIGHT) conformal prediction was applied to the BCNN uncertainties on e_1. The red horizontal line indicates the ideal PIT histogram distribution: if the PIT histogram peaks at the center, the ellipticity PDFs are generally too broad, and if the PIT histogram peaks at high and low PIT values, the PDF samples are too narrow.

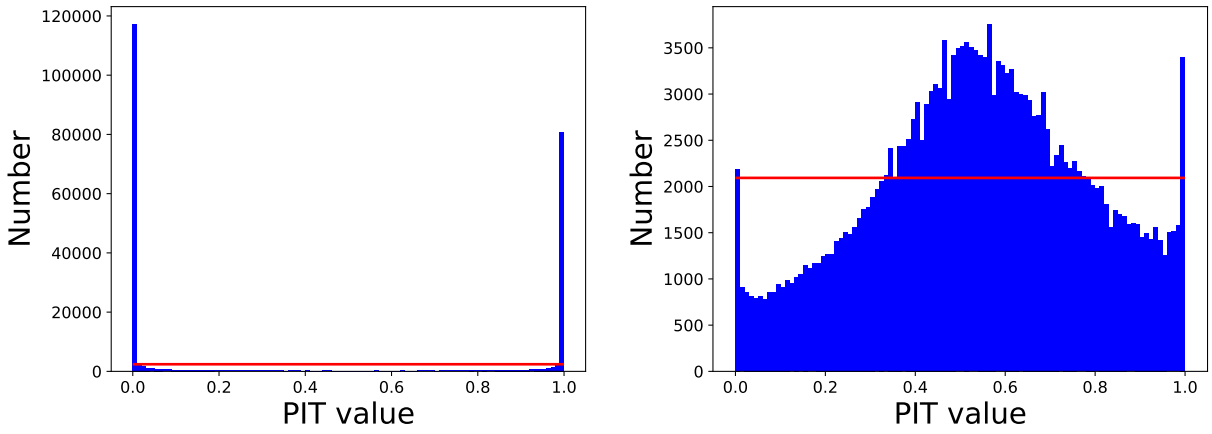

Figure 5.15 PIT histograms of the ellipticity PDF produced by the BCNN before (LEFT) and after (RIGHT) conformal prediction was applied to the BCNN uncertainties on e_2. The red horizontal line indicates the ideal PIT histogram distribution: if the PIT histogram peaks at the center, the ellipticity PDFs are generally too broad, and if the PIT histogram peaks at high and low PIT values, the PDF samples are too narrow.

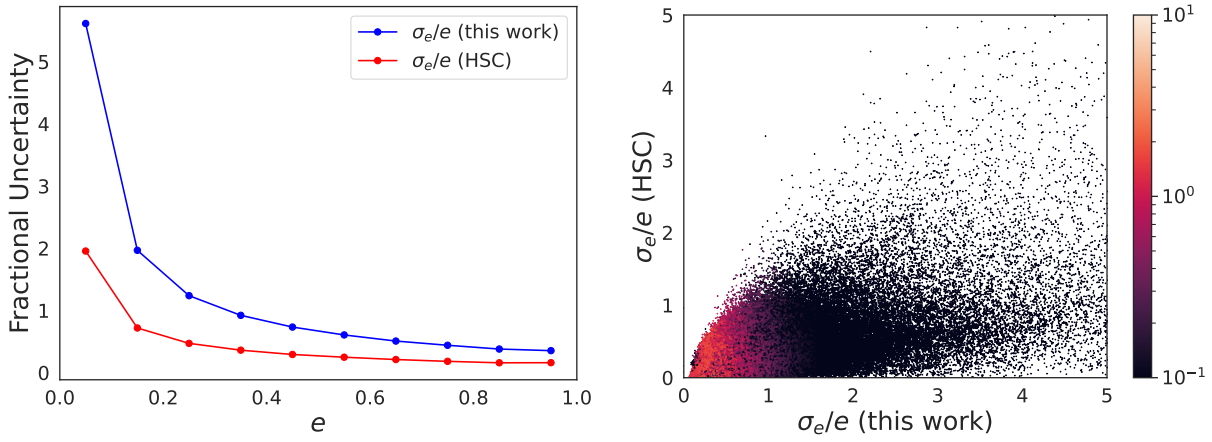

Figure 5.16 LEFT: Fractional uncertainty versus ellipticity in bins of 0.1 for this work compared to the ellipticity measurement and uncertainty provided in the HSC Shape Catalog. Since the HSC data was used as training labels in the dataset, we expect the uncertainty for this work to be higher than HSC, which is reflected here. RIGHT: Fractional uncertainty in the ellipticity provided by HSC versus the ellipticity estimates provided in this work.

part of weak lensing cosmological probes.

There are a number of future steps required to advance the approach in this work toward a final shear analysis on data from large scale surveys. The principal task is to simulate galaxy images that are representative of HSC. For this objective we will need to likely modify the real galaxy images and impose known shear values and random noise/orientation adjustments. The quality of the dataset can be assessed by training on the simulated sample and recovering the shape information contained in the real sample. Another avenue to consider toward developing a machine learning model for shear estimation is to investigate the inclusion of galaxy position data as inputs into the model. Because galaxy shears are correlated on positional scales on the order of the dark matter clump distributions that shear galaxy images, having position data may enable the model to utilize information from multiple images on individual shear measurements. Alternatively, one can modify a machine learning model to process large galaxy fields as input rather than individual galaxy images.

APPENDIX A

Appendix

The purpose of this Appendix is to detail the intended method by which one can utilize the photo-z and potential shear estimation models in this work for performing a weak lensing cosmological measurement. This work will be carried out by the Galaxies ML research group led by Tuan Do.

A.1 Testing ΛCDM with Weak Lensing

Lensing measurements obtained by an observer result from 1) the true mass distribution present in the observations, 2) the intrinsic alignment of galaxies in the observations (summed contribution of physical and gravitational interactions of galaxies that influence their orientation and shape), and 3) photo-z estimation uncertainties and other measurement uncertainties. All systematic uncertainties affecting lensing measurements are propagated as uncertainties in cosmological parameter constraints.

Overdense regions in the matter density field are quantified with respect to the the average density in time and space:

$$\delta = \frac{(\rho(x,t) - \bar{\rho}(x,t))}{\bar{\rho}(x,t)} \tag{A.1}$$

Gravitational lensing produces cosmic shear, which is quantified statistically with a 2-point

correlation function:

$$\xi(\vec{r})_{\phi\psi} = \int \phi(\vec{r}')\psi(\vec{r}' - \vec{r})d^3r' \tag{A.2}$$

where ϕ and ψ are two homogeneous and isotropic fields.

Let's define the likelihood of encountering an overdense region in a volume V as P_V. If one observes an overdense region (as defined by δ), the likelihood of encountering a neighboring overdense region with separation $\vec{r}$ is

$$P_V^2[1 + \xi_{\delta\delta}]. \tag{A.3}$$

where $\xi_{\delta\delta}$ is the 2-point correlation function between overdense region peaks. If the 2-point correlation function between regions separated by $\vec{r}$ is 0, the regions are statistically independent.

For a weak lensing analysis using the photo-z and shear estimation models in this work, we assume a spatially flat universe under ΛCDM Cosmology. The expansion rate of the universe is connected to the total energy density

$$H^2(t) = \left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \sum_i \Omega_i(t). \tag{A.4}$$

where $a = \frac{1}{1+z}$ is the scale factor, H_0 is the Hubble parameter today, and the energy density constituents of the universe are given by $\Omega_i = \frac{\rho_i}{\rho_c}$, where ρ_c is the critical density.

H_0 can be defined in terms of the critical density today

$$H_0^2 = \frac{8\pi G\rho_{crit}(t_0)}{3}. \tag{A.5}$$

The equation of state parameter w is related to the density and expansion rate of the

universe:

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \tag{A.6}$$

and the pressure is related to density by

$$P = w\rho. \tag{A.7}$$

The equation of state parameter values for radiation, matter, and dark energy are 1/3, 0, and -1, respectively. Equation A.4 can be stated in terms of the Hubble parameter, matter densities, and scale factor:

$$H(a) = H_0\sqrt{\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_\Lambda}, \tag{A.8}$$

Under the ΛCDM paradigm, the different density parameters for our universe are Ω_C for cold dark matter matter, Ω_b for baryonic matter, Ω_Λ for dark energy, and Ω_R for radiation (photos and relativistic neutrinos). Ω_R is negligible ($\sim 10^{-4}$) compared to the contributions of matter $\Omega_M = \Omega_C + \Omega_b$ and dark energy.

Table A.1 contains example cosmological parameters that are probed in a weak lensing analysis by the HSC team using traditional photo-z and shear estimation techniques (Hikage et al., 2019). Among the energy density constituents are σ_8, which is the root mean square of mass fluctuations, the Hubble constant h, and the scalar spectral index n_s. Another useful parameter to model is S_8, which combines Ω_m and σ_8:

$$S_8 = \sigma_8\sqrt{\frac{\Omega_m}{0.3}} \tag{A.9}$$

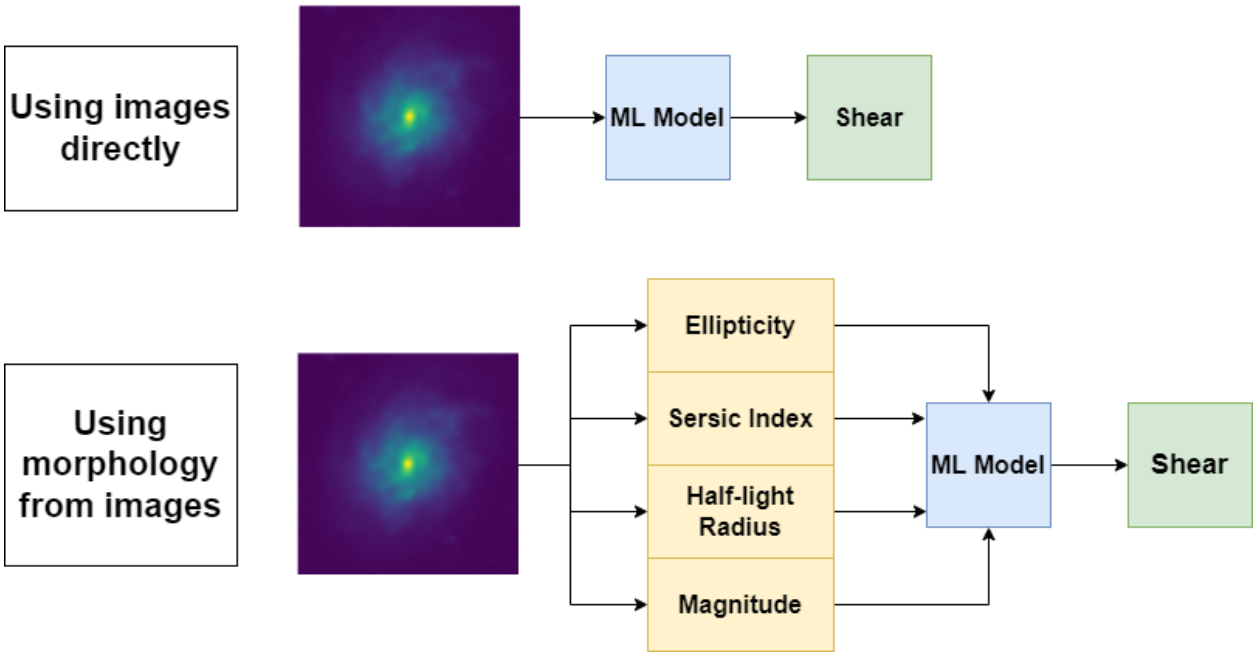

Figure A.1 Visualization of two distinct approaches to measure galaxy shear that can be used as inputs into a weak lensing probe. The top flowchart utilizes galaxy images as input into a machine learning model, such as a BCNN, and the bottom flowchart visualizes galaxy morphological features as input into a machine learning model, such as a BNN. In principle, the approach using images directly should contain more information, but both models can be explored to assess their strengths and weaknesses.

A.2 From photometric redshifts and shears to cosmological parameters

One can use the Pseudo-Cl method applied in Hikage et al. (2019) to infer cosmological measurements from photometric redshifts and shear estimates. The Pseudo-Cl method characterizes cosmic shear using the power spectrum – the mean square of fluctuation amplitudes as a function of multipole ℓ – in Fourier space. Fourier space is ideal for working with statistics of translation-invariant fields (Hamimeche & Lewis, 2008) and covariance matrix estimation is easier in Fourier space (Alonso et al., 2019). This method is faster than other methods, which is advantageous due to the high dimensionality of the theoretical model in the likeli-

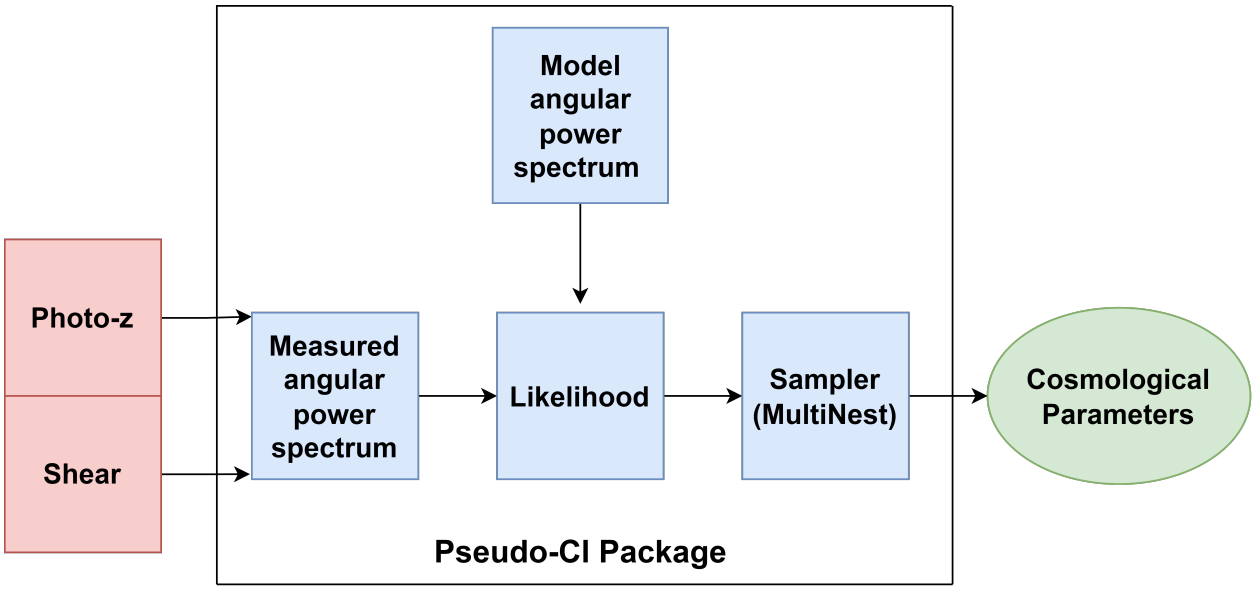

Figure A.2 Flow chart of steps used involved in using the Pseudo-Cl method for weak lensing cosmology. Photo-z and shear estimates are used to produce dimensionless binned angular power spectra. Model power spectra are calculated over a range of cosmological parameter values, which are jointly constrained using nested sampling. The blue boxes refer to steps within the publicly available Psuedo-Cl code package.

hood sampling of cosmological parameters (Hikage et al., 2019). Further, this method avoids error contributions due to incompleteness in: 1) the sky coverage due to complicated survey geometry resulting from bright star masks, 2) survey boundaries and depths, and 3) galaxy shape weights. The authors make their code[1] and shape data publicly available[2], so their results are very suitable for comparison.

The observed shear field is given by

$$\gamma^{obs}(\theta) = W(\theta)\gamma^{true}(\theta) \tag{A.10}$$

where $W(\theta)$ is the survey windows resulting from the sum of shear weights in each pixel.

[1] $https://github.com/chiaki-hikage/Likelihood_pseudoCl_HSCY1$

[2] https://hsc-release.mtk.nao.ac.jp/doc/index.php/s16a-shape-catalog-pdr2/

The shear field for each of the six HSC shear regions are transformed into Fourier space with the Pseudo-Cl method to obtain unbiased shear power spectrum estimates by correcting for survey window variabilities. In this method, the likelihood in Equation 26 is stated in terms of the observed pseudo-spectrum C_ℓ^{obs} obtained from the Fourier transform of $\gamma^{obs}(\theta)$ and the model pseudo-spectrum C_b^{model} (Hikage et al., 2011; Kitching et al., 2012; Hikage & Oguri, 2016; Asgari et al., 2018). The measured power spectrum for a given multipole bin is given by

$$C_b^{true} = M_{bb'}^{-1} \sum_{\ell}^{\ell \in \ell_b} P_{b'\ell}(C_\ell^{obs} - \langle N_\ell \rangle_{MC}). \tag{A.11}$$

$M_{bb'}$ is the mode coupling matrix of binned power spectra obtained from the survey window (see equation A7 in Hikage et al. (2019)). C_ℓ^{obs} is the observed pseudo-spectrum obtained from the Fourier transform of $\gamma^{obs}(\theta)$. $P_{b'\ell} = \frac{\ell^2}{2\pi}$ is a conversion factor to produce a dimensionless power spectrum. Finally, $\langle N_\ell \rangle$ is a convolved noise spectrum produced from taking the average of shot noise power spectra N_ℓ from 10000 Monte Carlo simulations with random galaxy orientations. The model power spectrum is computed in each multipole bin as:

$$C_b^{model} = \frac{\sum_\ell^{\ell \in \ell_b} P_{b\ell} C_\ell^{model}}{\sum_\ell^{\ell \in \ell_b} P_{b\ell}}. \tag{A.12}$$

The likelihood is stated in terms of $\Delta C_b = C_b^{true} - C_b^{model}$:

$$-2\log(\mathcal{L}) = \sum_{iji'j'}^{\ell_{min} \leq \ell_b, \ell_{b'} \leq \ell_{max}} \sum_{b,b'} \Delta C_b^{(ij)} \left[Cov\right]^{-1} \Delta C_{b'}^{(i'j')} + \ln|Cov| + const \tag{A.13}$$

Because the authors provide the data and code used in Hikage et al. (2019), one can reproduce this method using improved photo-z and shear estimation techniques. Any difference in results is attributable to the photo-z and shear estimates.

Table A.1 Summary of parameters and priors used in Hikage et al. (2019) with the Psuedo-cl method to sample tomographic cosmic shear power spectra.

Parameter	symbols	prior
physical dark matter density	$\Omega_c h^2$	flat [0.03,0.7]
physical baryon density	$\Omega_b h^2$	flat [0.019,0.026]
Hubble parameter	h	flat [0.6,0.9]
scalar amplitude on $k = 0.05\mathrm{Mpc}^{-1}$	$\ln(10^{10} A_s)$	flat [1.5,6]
scalar spectral index	n_s	flat [0.87,1.07]
optical depth	τ	flat [0.01,0.2]
neutrino mass	$\sum m_\nu$ [eV]	$(0)^\dagger$, (0.06) or flat [0,1]
dark energy EoS parameter	w	fixed $(-1)^\dagger$ or flat $[-2, -0.333]$
amplitude of the intrinsic alignment	A_{IA}	flat $[-5, 5]$
redshift dependence of the intrinsic alignment	η_{eff}	flat $[-5, 5]$
baryonic feedback amplitude	A_B	fixed $(0)^\dagger$ or flat $[-5, 5]$
PSF leakage	$\tilde{\alpha}$	Gauss $(0.057, 0.018)$
residual PSF model error	$\tilde{\beta}$	Gauss $(-1.22, 0.74)$
uncertainty of multiplicative bias m	$100\Delta m$	Gauss $(0, 1)$
photo-z shift in bin 1	$100\Delta z_1$	Gauss $(0, 2.85)$
photo-z shift in bin 2	$100\Delta z_2$	Gauss $(0, 1.35)$
photo-z shift in bin 3	$100\Delta z_3$	Gauss $(0, 3.83)$
photo-z shift in bin 4	$100\Delta z_4$	Gauss $(0, 3.76)$

Bibliography

Abadi, M., Agarwal, A., Barham, P., et al. 2016, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv, doi: 10.48550/arXiv.1603.04467

Aihara, H., Armstrong, R., Bickerton, S., et al. 2018, Publications of the Astronomical Society of Japan, 70, doi: 10.1093/pasj/psx081

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, Publications of the Astronomical Society of Japan, 71, 114, doi: 10.1093/pasj/psz103

Ait-Ouahmed, R., Arnouts, S., Pasquet, J., Treyer, M., & Bertin, E. 2023, Multimodality for improved CNN photometric redshifts, arXiv. http://arxiv.org/abs/2310.02185

Alonso, D., Sanchez, J., & Slosar, A. 2019, Monthly Notices of the Royal Astronomical Society, 484, 4127, doi: 10.1093/mnras/stz093

Angelopoulos, A. N., & Bates, S. 2022, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, arXiv, doi: 10.48550/arXiv.2107.07511

Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, Monthly Notices of the Royal Astronomical Society, 310, 540, doi: 10.1046/j.1365-8711.1999.02978.x

Asgari, M., Taylor, A., Joachimi, B., & Kitching, T. D. 2018, Monthly Notices of the Royal Astronomical Society, doi: 10.1093/mnras/sty1412

Benítez, N. 2000, The Astrophysical Journal, 536, 571, doi: 10.1086/308947

Bernstein, G. M., & Jarvis, M. 2002, The Astronomical Journal, 123, 583, doi: 10.1086/338085

Bertin, E., & Arnouts, S. 1996, 117, 393, doi: 10.1051/aas:1996164

Bradshaw, E. J., Almaini, O., Hartley, W. G., et al. 2013, Monthly Notices of the Royal Astronomical Society, 433, 194, doi: 10.1093/mnras/stt715

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Carrasco Kind, M., & Brunner, R. J. 2013, Monthly Notices of the Royal Astronomical Society, 432, 1483, doi: 10.1093/mnras/stt574

Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco California USA: ACM), 785–794, doi: 10.1145/2939672.2939785

Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, The Astrophysical Journal, 741, 8, doi: 10.1088/0004-637X/741/1/8

Collaboration, D., Abbott, T. M. C., Abdalla, F. B., et al. 2018, Physical Review D, 98, 043526, doi: 10.1103/PhysRevD.98.043526

Collaboration, E., Ilić, S., Aghanim, N., et al. 2022, Astronomy & Astrophysics, 657, A91, doi: 10.1051/0004-6361/202141556

Collaboration, T. L. D. E. S., Mandelbaum, R., Eifler, T., et al. 2021, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document, arXiv. http://arxiv.org/abs/1809.01669

Collister, A. A., & Lahav, O. 2004, Publications of the Astronomical Society of the Pacific, 116, 345, doi: 10.1086/383254

Cool, R. J., Moustakas, J., Blanton, M. R., et al. 2013, The Astrophysical Journal, 767, 118, doi: 10.1088/0004-637X/767/2/118

Cortes, C., & Vapnik, V. 1995, Machine Learning, 20, 273, doi: 10.1007/BF00994018

Davis, M., Faber, S. M., Newman, J., et al. 2003, 4834, 161, doi: 10.1117/12.457897

Dusenberry, M. W., Tran, D., Choi, E., et al. 2020, in ACM Conference on Health, Inference, and Learning (ACM CHIL). https://arxiv.org/abs/1906.03842

Filos, A., Farquhar, S., Gomez, A. N., et al. 2019, A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks, arXiv, doi: 10.48550/arXiv.1912.10481

Firth, A. E., Lahav, O., & Somerville, R. S. 2003, Monthly Notices of the Royal Astronomical Society, 339, 1195, doi: 10.1046/j.1365-8711.2003.06271.x

Garilli, B., Guzzo, L., Scodeggio, M., et al. 2014, Astronomy & Astrophysics, 562, A23, doi: 10.1051/0004-6361/201322790

Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, The Astrophysical Journal, 715, 823, doi: 10.1088/0004-637X/715/2/823

Hamana, T., Shirasaki, M., Miyazaki, S., et al. 2020, Publications of the Astronomical Society of Japan, 72, 16, doi: 10.1093/pasj/psz138

Hamimeche, S., & Lewis, A. 2008, Physical Review D, 77, 103013, doi: 10.1103/PhysRevD.77.103013

Hearin, A. P., Zentner, A. R., Ma, Z., & Huterer, D. 2010, The Astrophysical Journal, 720, 1351, doi: 10.1088/0004-637X/720/2/1351

Heymans, C., Tröster, T., Asgari, M., et al. 2020, doi: 10.1051/0004-6361/202039063

Hikage, C., & Oguri, M. 2016, Monthly Notices of the Royal Astronomical Society, 462, 1359, doi: 10.1093/mnras/stw1721

Hikage, C., Takada, M., Hamana, T., & Spergel, D. 2011, Monthly Notices of the Royal Astronomical Society, 412, 65, doi: 10.1111/j.1365-2966.2010.17886.x

Hikage, C., Oguri, M., Hamana, T., et al. 2019, Publications of the Astronomical Society of Japan, 71, 43, doi: 10.1093/pasj/psz010

Hirata, C., & Seljak, U. 2003, Monthly Notices of the Royal Astronomical Society, 343, 459, doi: 10.1046/j.1365-8711.2003.06683.x

Hoekstra, H., Viola, M., & Herbonnet, R. 2017, Monthly Notices of the Royal Astronomical Society, 468, 3295, doi: 10.1093/mnras/stx724

Hoff, P. 2021, arXiv:2105.14045 [math, stat]

Hsieh, B. C., & Yee, H. K. C. 2014, The Astrophysical Journal, 792, 102, doi: 10.1088/0004-637X/792/2/102

Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, Monthly Notices of the Royal Astronomical Society, 366, 101, doi: 10.1111/j.1365-2966.2005.09782.x

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, Astronomy & Astrophysics, 457, 841, doi: 10.1051/0004-6361:20065138

Ivezic, 2018, 58

Ivezić, , Kahn, S. M., Tyson, J. A., et al. 2008, doi: 10.3847/1538-4357/ab042c

Jee, M. J., Tyson, J. A., Schneider, M. D., et al. 2013, The Astrophysical Journal, 765, 74, doi: 10.1088/0004-637X/765/1/74

Jones, E., Do, T., Boscoe, B., Wan, Y., & Nguyen, Z. 2021a, Photometric Redshifts for Cosmology: Improving accuracy and uncertainty estimates using Bayesian Neural Networks, v6, Zenodo, doi: 10.5281/zenodo.5528827

Jones, E., Do, T., Boscoe, B., et al. 2022a, Photometric Redshifts for Cosmology: Improving Accuracy and Uncertainty Estimates Using Bayesian Neural Networks, arXiv, doi: 10.48550/arXiv.2202.07121

—. 2022b, doi: 10.48550/ARXIV.2202.07121

Jones, E., Do, T., Boscoe, B., Wan, Y., & Singal, J. 2023, doi: 10.48550/ARXIV.2202.07121

Jones, E., Do, T., Wan, Y., et al. 2021b, in Debating the Potential of Machine Learning in Astronomical Surveys, Paris

Jones, E., & Singal, J. 2017, Astronomy & Astrophysics, 600, A113, doi: 10.1051/0004-6361/201629558

—. 2020, Publications of the Astronomical Society of the Pacific, 132, 024501, doi: 10.1088/1538-3873/ab54ed

Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., & Bennamoun, M. 2020, arXiv:2007.06823 [cs, stat]. http://arxiv.org/abs/2007.06823

Kitching, T. D., Balan, S. T., Bridle, S., et al. 2012, Monthly Notices of the Royal Astronomical Society, 423, 3163, doi: 10.1111/j.1365-2966.2012.21095.x

Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2016, doi: 10.48550/arXiv.1612.01474

Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, Astronomy and Astrophysics, 559, A14, doi: 10.1051/0004-6361/201322179

Lei, J., & Wasserman, L. 2014, Journal of the Royal Statistical Society Series B: Statistical Methodology, 76, 71, doi: 10.1111/rssb.12021

Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, The Astrophysical Journal Supplement Series, 184, 218, doi: 10.1088/0067-0049/184/2/218

Lin, Q., Fouchez, D., Pasquet, J., et al. 2022, Astronomy & Astrophysics, 662, A36, doi: 10.1051/0004-6361/202142751

Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, Monthly Notices of the Royal Astronomical Society, 452, 2087, doi: 10.1093/mnras/stv1436

Malz, A. I. 2021, Physical Review D, 103, 083502, doi: 10.1103/PhysRevD.103.083502

Malz, A. I., & Hogg, D. W. 2020, arXiv:2007.12178 [astro-ph]. http://arxiv.org/abs/2007.12178

Mandelbaum, R. 2015, Journal of Instrumentation, 10, C05017, doi: 10.1088/1748-0221/10/05/C05017

—. 2018, Annual Review of Astronomy and Astrophysics, 56, 393, doi: 10.1146/annurev-astro-081817-051928

Mandelbaum, R., Hirata, C. M., Seljak, U., et al. 2005, Monthly Notices of the Royal Astronomical Society, 361, 1287, doi: 10.1111/j.1365-2966.2005.09282.x

Mandelbaum, R., Rowe, B., Bosch, J., et al. 2014, The Astrophysical Journal Supplement Series, 212, 5, doi: 10.1088/0067-0049/212/1/5

Mandelbaum, R., Rowe, B., Armstrong, R., et al. 2015, Monthly Notices of the Royal Astronomical Society, 450, 2963, doi: 10.1093/mnras/stv781

Mandelbaum, R., Miyatake, H., Hamana, T., et al. 2018, Publications of the Astronomical Society of Japan, 70, doi: 10.1093/pasj/psx130

McLure, R. J., Pearce, H. J., Dunlop, J. S., et al. 2012, doi: 10.1093/mnras/sts092

Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, The Astrophysical Journal Supplement Series, 225, 27, doi: 10.3847/0067-0049/225/2/27

Newman, J. A., & Gruen, D. 2022, Annual Review of Astronomy and Astrophysics, 60, annurev, doi: 10.1146/annurev-astro-032122-014611

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, The Astrophysical Journal Supplement Series, 208, 5, doi: 10.1088/0067-0049/208/1/5

Nishizawa, A. J., Hsieh, B.-C., Tanaka, M., & Takata, T. 2020, Photometric Redshifts for the Hyper Suprime-Cam Subaru Strategic Program Data Release 2, arXiv, doi: 10.48550/arXiv.2003.01511

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. 2002, in Machine Learning: ECML 2002, ed. T. Elomaa, H. Mannila, & H. Toivonen, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer), 345–356

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, Astronomy & Astrophysics, Volume 621, id.A26, <NUMPAGES>15</NUMPAGES> pp., 621, A26, doi: 10.1051/0004-6361/201833617

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825, doi: 10.48550/arXiv.1201.0490

Pujol, A., Bobin, J., Sureau, F., Guinot, A., & Kilbinger, M. 2020, Astronomy & Astrophysics, 643, A158, doi: 10.1051/0004-6361/202038658

Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, Publications of the Astronomical Society of the Pacific, 128, 104502, doi: 10.1088/1538-3873/128/968/104502

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020a, Monthly Notices of the Royal Astronomical Society, staa2799, doi: 10.1093/mnras/staa2799

—. 2020b, Monthly Notices of the Royal Astronomical Society, staa2799, doi: 10.1093/mnras/staa2799

Schuldt, S., Suyu, S. H., Cañameras, R., et al. 2020a, arXiv:2011.12312 [astro-ph]. http://arxiv.org/abs/2011.12312

—. 2020b, arXiv:2011.12312 [astro-ph]. http://arxiv.org/abs/2011.12312

—. 2021, Astronomy & Astrophysics, 651, A55, doi: 10.1051/0004-6361/202039945

Simonyan, K., & Zisserman, A. 2015, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv, doi: 10.48550/arXiv.1409.1556

Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., & Lotz, J. 2011, Publications of the Astronomical Society of the Pacific, 123, 615, doi: 10.1086/660155

Singal, J., Silverman, G., Jones, E., et al. 2022, The Astrophysical Journal, 928, 6, doi: 10.3847/1538-4357/ac53b5

Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, The Astrophysical Journal Supplement Series, 214, 24, doi: 10.1088/0067-0049/214/2/24

Specht, D. F. 1990, Neural Networks, 3, 109, doi: 10.1016/0893-6080(90)90049-Q

Springer, O. M., Ofek, E. O., Weiss, Y., & Merten, J. 2019, Monthly Notices of the Royal Astronomical Society, stz2991, doi: 10.1093/mnras/stz2991

Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018a, Publications of the Astronomical Society of Japan, 70, doi: 10.1093/pasj/psx077

—. 2018b, Publications of the Astronomical Society of Japan, 70, doi: 10.1093/pasj/psx077

Terrizzano, I. G., Schwarz, P., Roth, M., & Colino, J. E. 2015, Data Wrangling: The Challenging Journey from the Wild to the Lake. https://www.semanticscholar.org/paper/Data-Wrangling%3A-The-Challenging-Journey-from-the-to-Terrizzano-Schwarz/2a24f587b68a1ef6539b4ed8725dfe76f0ed40e2

Tewes, M., Kuntzer, T., Nakajima, R., et al. 2019, Astronomy & Astrophysics, 621, A36, doi: 10.1051/0004-6361/201833775

Treyer, M., Ait-Ouahmed, R., Pasquet, J., et al. 2023, CNN photometric redshifts in the SDSS at $r\leq 20$, arXiv. http://arxiv.org/abs/2310.02173

Vovk, V. 2012, in Proceedings of the Asian Conference on Machine Learning (PMLR), 475–490

Wadadekar, Y. 2005, Publications of the Astronomical Society of the Pacific, 117, 79, doi: 10.1086/427710

Wyatt, M., & Singal, J. 2020, arXiv:1911.04572 [astro-ph]. http://arxiv.org/abs/1911.04572

Zhou, X., Gong, Y., Meng, X.-M., et al. 2022, Photometric redshift estimates using Bayesian neural networks in the CSST survey, arXiv. http://arxiv.org/abs/2206.13696